Bayesian Methods for Mixtures of Normal Distributions

Matthew Stephens

Magdalen College, Oxford

Michaelmas Term, 1997

A thesis submitted to the Faculty of Mathematical Sciences
for the degree of Doctor of Philosophy of the University of Oxford.

Abstract

Bayesian Methods for Mixtures of Normal Distributions

Matthew Stephens
Magdalen College, Oxford

D. Phil. Thesis
Michaelmas Term, 1997

Mixture distributions are typically used to model data in which each observation is assumed to have arisen from one of a number of different groups. They also provide a convenient and flexible class of models for density estimation.

While a Bayesian analysis of mixture models has certain advantages over a classical approach, it is not without its problems. In theory quantities of interest may be written down as integrals, but in practice these integrals cannot be done analytically. When the number of groups in the data is assumed known, the Gibbs sampler can be used to perform this integration numerically, but the *non-identifiability* of the mixture model parameters causes *label-switching* in the Gibbs sampler output and makes inference for the individual components of the mixture meaningless. We show that the usual method of dealing with this problem (imposing simple identifiability constraints on the mixture model parameters) is sometimes inadequate, and present a more flexible approach to solving this problem, which allows sensible clustering to be performed in a Bayesian context and allows interpretations for groups to be discovered rather than imposed. We illustrate the success of our approach on several examples.

When the number of groups in the data is considered unknown more sophisticated methods are required to perform the integration necessary for a Bayesian analysis. One method is described by Richardson and Green (1997), which they apply successfully to univariate data. We describe an alternative 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. We apply this method successfully to both univariate and bivariate data.

Finally we examine "on-line" methods for mixture models, in which the posterior distribution of the parameters is updated as observations arrive sequentially, and are then discarded. We show that the computationally trivial *Quasi-Bayes* method of Makov and Smith (1977) can be improved upon at the expense of small additional computational complexity.

Acknowledgements

I would like to thank my supervisor, Professor Brian Ripley, for his skillful supervision, allowing me freedom to make my own mistakes and discoveries, investigate alleys and cul-de-sacs, whilst ensuring I remained on the right track. Thanks also to Dave Flitney and Susan Hutchinson for all their computer advice, and Mark Mathieson for some stimulating discussions. The financial support of the Engineering and Physical Sciences Research Council is gratefully acknowledged.

I would also like to thank all my friends for providing me with good food and companionship along the way, particularly Dru, Jenny, Ian, Charlie, Marcus, Andy, Ed and (last but not least) Laura. A great eight.

Finally, the list of thanks would not be complete without a mention of my family, to whom I owe a great deal and more. This is for them.

Contents

Chapter 1

Introduction

Since the first attempt to analyse a mixture model by Pearson (1894), mixture models have been used in an incredible range of applications, from fish lengths (Hosmer, 1973) to philately (Izenman and Sommer, 1988). For a more comprehensive list of applications see the monograph by Titterington *et al.* (1985) and the more recent article by Titterington (1997). Pearson (1894) used a method of moments to estimate the parameters of a mixture of 2 univariate normal distributions, but mixture models are now more commonly analysed using maximum likelihood or Bayesian methods. Key texts for the non-Bayesian analysis of mixture models, including the maximum likelihood method, are Titterington *et al.* (1985) and McLachlan and Basford (1988). Bayesian methods have recently become popular due to advances in both methodology and computer power. A comprehensive general introduction to Bayesian theory can be found in Bernardo and Smith (1994), and chapters on the application of Bayesian methods to mixture models are included in the books by Robert (1994), and Gelman *et al.* (1995); see also Robert (1996). Some key papers on the Bayesian analysis of mixtures are Diebolt and Robert (1994), Escobar and West (1995) and Richardson and Green (1997).

In this thesis we consider some of the problems posed by application of Bayesian methods to finite mixture models. We consider models 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:

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \pi_1 f(x; \phi_1, \eta) + \cdots + \pi_k f(x; \phi_k, \eta) \tag{1.1}$$

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 thesis $p(\cdot \mid \cdot)$ will be used interchangeably to denote conditional densities and distributions.

Mixture models are typically used to model data where each observation is assumed to have arisen from one of k (with k possibly unknown) groups, each group

being suitably modelled by a density from the parametric family f. The mixture proportions then represent the relative frequency of occurance of each group in the population, and the model provides a framework by which observations may be clustered together into groups for discrimination or classification (see for example McLachlan and Basford, 1988).

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).

In this thesis we examine Bayesian methods of analysing mixture models in the context of both clustering and density estimation, including the case where the number of components k is unknown. We concentrate on mixtures of univariate and bivariate normal distributions, but most of our work could easily be extended to mixtures of other distributions. The structure of the thesis is as follows:

Chapter 1 provides a brief overview of methods of analysing mixture models, and some of the problems they present.

Chapter 2 provides an introduction to Markov Chain Monte Carlo (MCMC) methods, and their use in a Bayesian analysis of mixture models in the particular case where the number of components k in the mixture is assumed known. We will also see how symmetries in the posterior distribution of the mixture model parameters can lead to the problem of *label-switching* which makes inference for the individual components of the mixture difficult.

Chapter 3 is devoted to the development and illustration of a solution to the problem of label-switching.

Chapter 4 describes a method of performing a Bayesian analysis in the case where the number of components k in the mixture is considered unknown, based on ideas from the simulation of point processes. The method is demonstrated on several examples for mixtures of both univariate and bivariate normal distributions. These examples provide further demonstration of the solution to the label-switching problem described in Chapter 3.

Chapter 5 compares some methods of performing an approximate Bayesian analysis when the observations $x_1, x_2, \ldots$ are considered sequentially; each observation is processed when it arrives, and then discarded before the next observation arrives. This situation is often referred to as *online learning* in the engineering literature, and contrasts with the *batch* learning methods studied in other chapters, which require all the data to be available at the same time. Most work in this area was done at a

time when computers were less powerful than they are today, and we will see that methods developed then (and still in current use) can be improved upon at the expense of small additional computational complexity.

1.1 Introduction to the analysis of mixture models

We now take a look at some of the questions which may be of interest when analysing data which is assumed to have arisen from a mixture model, and at some of the problems which arise when attempting to answer these questions.

1.1.1 Missing data formulation

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 (that is, *missing*) component z_j of the mixture. We will refer to the missing data $z^n = z_1, \ldots, z_n$ as the *allocation variables*, and to (x^n, z^n) as the *completed data*. The model (1.1) can be written in terms of the missing data, with $z_1, \ldots, z_n$ assumed to be realisations of independent and identically distributed discrete random variables $Z_1, \ldots, Z_n$ with probability mass function

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

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

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

Integrating out the missing data $Z_1, \ldots, Z_n$ then yields the model (1.1):

$$\begin{aligned} p(x_j \,|\, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) &= \sum_{i=1}^{k} \Pr(Z_j = i \,|\, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(x_j \,|\, Z_j = i, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) \\ &= \sum_{i=1}^{k} \pi_i f(x_j; \phi_i, \eta) \end{aligned} \tag{1.2}$$

If the components of the mixture model have a physical interpretation, then inference for the Zs may be of interest in itself, and we may be interested in quantities such as the *classification probabilities*:

$$\begin{aligned} \Pr(Z_j = i \,|\, x_j, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) &\propto \Pr(Z_j = i \,|\, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(x_j \,|\, Z_j = i, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) \\ &\propto \pi_i f(x_j; \phi_i, \eta) \end{aligned} \tag{1.3}$$

which gives

$$\Pr(Z_j = i \,|\, x_j, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \frac{\pi_i f(x_j; \phi_i, \eta)}{\sum_{l=1}^{k} \pi_l f(x_j; \phi_l, \eta)}. \tag{1.4}$$

Even if the components of the mixture have no physical interpretation, as may be the case if the mixture model is being used purely as a method of approximating or estimating a density, introduction of the allocation variables is still a convenient notational and computational device.

1.1.2 Questions of interest

As an example of some of the questions which we may wish to answer in a mixture model analysis we consider briefly a dataset 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 analysed 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. The 82 observations used are shown in the appendix to this thesis (Section A.1), and a histogram of the data is shown in Figure 1.1. The view of the data given by a histogram may be crucially dependent on the bin-widths and starting points chosen (see for example Roeder, 1990). More reliable non-parametric density estimation devices are described in the appendix to this chapter (Section 1.2) and illustrated for this data in Figure 1.6.

Roeder (1990) gives a brief scientific background to the galaxy data:

> "After the Big Bang, it is believed that matter expanded at a tremendous rate. Because of the local attraction of matter, the galaxies formed. Astronomers predicted that gravitational pull would lead to some clustering of galaxies; however, there are data to suggest the presence of superclusters of galaxies, surrounded by large voids (de Lapparent *et al.*, 1986). . . . Given the expansion scenario of the universe, points furthest from our galaxy must be moving at greater velocities. Distance, then, is proportional to and can be estimated from velocity.
>
> If the galaxies are clumped, the distribution of velocities would be multimodal, each mode representing a cluster as it moves away at its own speed. Conversely, if there is no cluster effect, the distribution would be determined by the sampling scheme. From our galaxy we sample a conic section of space, but we have a declining ability to detect galaxies at greater distances; thus the velocity density should increase initially, and gradually tail off."

In this case then the components of the mixture may have a physical interpretation (the number of superclusters of galaxies for example), and inference for k

may be a major aim of the analysis. Inference for k is perhaps the most problematic aim of a mixture model analysis, and is addressed in Chapter 4.

Another common aim of a mixture model analysis, which may apply whether or not the components of the mixture have a physical interpretation, is to estimate the density (1.1). Figure 1.2 shows an estimate of the density for the galaxy data, obtained by fitting a mixture of $k = 3$ normal distributions to the data using Bayesian methods described in Chapter 2. When performing density estimation we may wish to allow k to vary even where the components of the mixture have no physical interpretation, as this will affect the smoothness of the density estimate obtained (see for example Figure 2.2). In fact, in a Bayesian setting density estimation is most naturally performed by taking a weighted average of the density estimates obtained for different values of k, rather than conditional on a fixed value of k, and we will see how this can be achieved in Chapter 4.

When the components of a mixture may have a physical interpretation, the following may also be aims of a mixture model analysis:

1. Estimating the density of the individual scaled components of the mixture
$$\pi_i f(x; \phi_i, \eta) \quad (i = 1, \ldots, k)$$
(see for example Figure 1.3).

2. Estimating the classification probabilities (1.4) either for the observed data points, or for a future data point not yet observed (see for example Figure 1.4).

3. Clustering the points into clusters of "similar" points. This may be done by choosing z_j to maximise $\Pr(Z_j = z_j)$ for $j = 1, \ldots, n$ (see for example Figure 1.5).

In contrast with density estimation, these aims only appear to make sense conditional on a fixed value of k, and this will be the approach we take. The aims are clearly highly related, and we will refer to them collectively as the application of mixture models to *clustering*, in which we include both classification and discrimination. The problems of performing clustering in a Bayesian setting are outlined in Section 1.1.5, and solutions to these problems are proposed in Chapter 3.

Mixture models are now usually analysed by either maximum likelihood or Bayesian methods, each of which present problems. We now examine briefly both these methods of analysis, and their associated problems.

1.1.3 The maximum likelihood approach

The maximum likelihood approach to parameter estimation in mixture models obtains point estimates $(\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\phi}}, \widehat{\eta})$ of the parameters $(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ by attempting to maximise the likelihood

$$L(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \prod_{j=1}^{n} \left[\pi_1 f(x_j; \phi_1, \eta) + \cdots + \pi_k f(x_j; \phi_k, \eta)\right]. \tag{1.5}$$

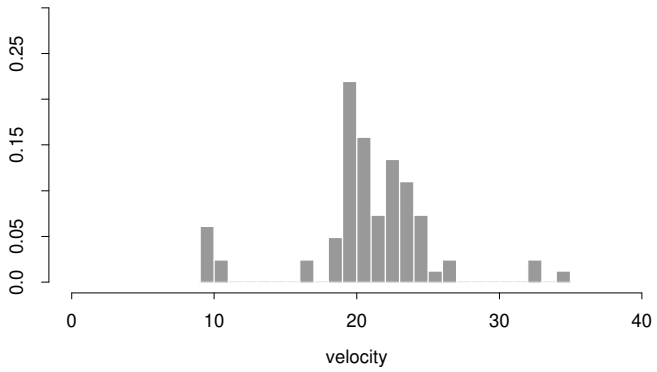


Figure 1.1: Histogram of the galaxy data, with bin-widths chosen by eye. Histograms are notoriously poor density estimation devices; for example Roeder (1990) illustrates how different bin-widths and starting points give histograms which give quite different pictures of this data. More reliable non-parametric density estimation devices are described in Section 1.2 and illustrated in Figure 1.6.

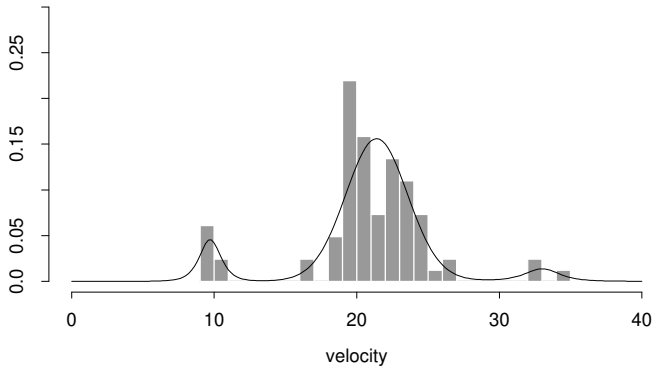


Figure 1.2: Histogram of the galaxy data overlaid with density estimate obtained by fitting a mixture of $k = 3$ normal components to the data using Bayesian methods, as described in Chapter 2 (Section 2.2), which includes some alternative density estimates based on fitting different mixture models to this data (Figures 2.2 and 2.11). Alternative non-parametric methods of density estimation are described in Section 1.2 and illustrated in Figure 1.6.

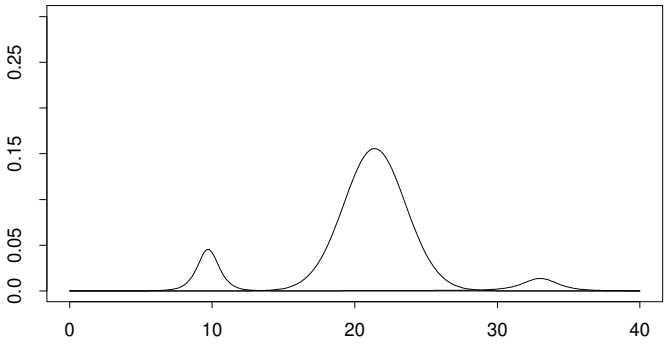


Figure 1.3: Estimates of $k = 3$ normal components, scaled by their weights, fitted to the galaxy data using Bayesian methods described in Chapter 2. The problems of obtaining sensible estimates of the scaled component densities in a Bayesian setting are outlined in Section 1.1.5, and solutions to these problems are proposed in Chapter 3.

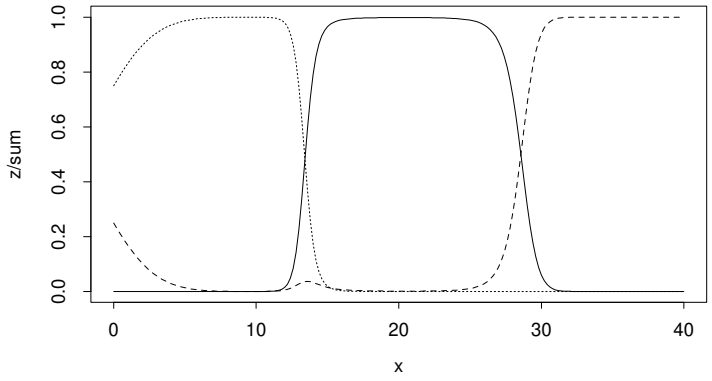


Figure 1.4: Estimates of classification probabilities for the three components fitted to the galaxy data in Figure 1.3.

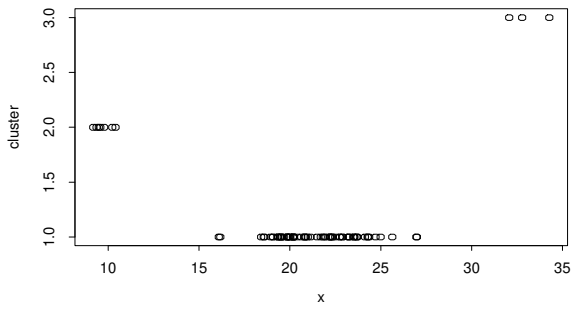


Figure 1.5: Clustering of the galaxy data into 3 clusters, based on maximising the classification probabilities shown in Figure 1.4.

Quantities of interest may then be estimated by "plugging in" the point estimates. For example, the density $p(x \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ may be estimated by

$$p(x \mid \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\phi}}, \widehat{\eta}) \tag{1.6}$$

which is given by (1.1), and the classification probabilities may be estimated by

$$\Pr(Z_j = i \mid x_j, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\phi}}, \widehat{\eta}) \tag{1.7}$$

which is given by (1.4).

Although popular, the maximum likelihood approach to mixture models is beset with difficulties, mostly caused by the fact that for many choices of parametric family f the likelihood (1.5) is unbounded. In particular, consider a mixture of univariate normal distributions, with likelihood

$$L(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \prod_{j=1}^{n} \left[\pi_1 \mathcal{N}(x_j; \mu_1, \sigma_1^2) + \cdots + \pi_k \mathcal{N}(x_j; \mu_k, \sigma_k^2) \right] \tag{1.8}$$

where $\mathcal{N}(x; \mu, \sigma^2)$ denotes the univariate normal density function, with mean μ and variance σ^2. This likelihood tends to infinity if we set $\mu_1 = x_1$ and allow σ_1^2 to tend to zero, and although this problem can be avoided by constraining the variances of all components to be equal, application of this constraint is not always appropriate.

Point estimates of the parameters corresponding to such singularities in the likelihood surface are of no real interest, and it is usual to seek a parameter estimate which corresponds to a large *local* maximum of the likelihood surface. Aside from the computational problems associated with finding such maxima, there may be several reasonable local maxima between which to choose, each of which may give quite different plug-in estimates for quantities of interest such as the density or classification probabilities. In many cases it will be difficult to justify choosing one of these point estimates of the parameters above the others. Furthermore, the

absence of a single dominant local maximum in the likelihood means that neither the standard asymptotic theory for maximum likelihood estimation, nor the theory underlying standard methods of choosing a suitable value for k, such as Akaike's AIC (Akaike, 1973), apply in the mixture context. Such problems have encouraged the development of a Bayesian approach to the problem.

1.1.4 The Bayesian approach

We assume the reader is familiar with the basics of the Bayesian paradigm, in which parameters are treated as random quantities, and point estimates for parameters are replaced by distributions on the parameter space which represent our knowledge or belief about the value of the parameters. A comprehensive description is given by Bernardo and Smith (1994).

The Bayesian approach avoids the problems associated with the maximum likelihood approach described above. It is not necessary to choose between many plausible point parameter estimates, as quantities of interest are found by averaging over the parameter space, weighting by the posterior distribution of the parameters. For example, if we assume that the number of components k is known, and specify a suitable prior distribution $p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ for the parameters of the mixture model (choice of suitable prior is a controversial topic which we will return to later) then in theory the posterior distribution $p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \,|\, x^n)$ is given by Bayes theorem, and the *predictive density* (the density of a future observation) given x^n is given by

$$\begin{aligned} p(x_{n+1} \,|\, x^n) &= \int p(x_{n+1} \,|\, x^n, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \,|\, x^n)\, d\boldsymbol{\pi}\, d\boldsymbol{\phi}\, d\eta \\ &= \int p(x_{n+1} \,|\, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \,|\, x^n)\, d\boldsymbol{\pi}\, d\boldsymbol{\phi}\, d\eta. \end{aligned} \tag{1.9}$$

Similarly, the classification probabilities for the observations $x_1, \ldots, x_n$ are given by

$$\begin{aligned} \Pr(Z_j = i \,|\, x^n) &= \int \Pr(Z_j = i \,|\, x^n, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \,|\, x^n)\, d\boldsymbol{\pi}\, d\boldsymbol{\phi}\, d\eta \\ &= \int \frac{\pi_i f(x_j; \phi_i, \eta)}{\sum_l \pi_l f(x_j; \phi_l, \eta)} p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \,|\, x^n)\, d\boldsymbol{\pi}\, d\boldsymbol{\phi}\, d\eta \end{aligned} \tag{1.10}$$

and the classification probabilities for a future observation x_{n+1} are given by

$$\begin{aligned}
\Pr(Z_{n+1} = i \mid x^{n+1}) &= \int \frac{\pi_i f(x_{n+1}; \phi_i, \eta)}{\sum_l \pi_l f(x_{n+1}; \phi_l, \eta)} p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \mid x^{n+1})\, d\boldsymbol{\pi}\, d\boldsymbol{\phi}\, d\eta \\
&\propto \int \frac{\pi_i f(x_{n+1}; \phi_i, \eta)}{p(x_{n+1} \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)} p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta, x^{n+1})\, d\boldsymbol{\pi}\, d\boldsymbol{\phi}\, d\eta \\
&= \int \frac{\pi_i f(x_{n+1}; \phi_i, \eta)}{p(x_{n+1} \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)} p(x_{n+1} \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) \cdot \\
&\qquad \cdot p(x^n \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta)\, d\boldsymbol{\pi}\, d\boldsymbol{\phi}\, d\eta \\
&\propto \int \pi_i f(x_{n+1}; \phi_i, \eta) p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \mid x^n)\, d\boldsymbol{\pi}\, d\boldsymbol{\phi}\, d\eta.
\end{aligned} \tag{1.11}$$

In general all aspects of the posterior distribution of the parameters are valid quantities for inference, and we will be interested in evaluating integrals of the form

$$E\big(F(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) \mid x^n\big) = \int F(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \mid x^n)\, d\boldsymbol{\pi}\, d\boldsymbol{\phi}\, d\eta. \tag{1.12}$$

The Bayesian approach also provides a natural framework for considering the case where the number of components k is unknown. By allowing k to vary with the other parameters and specifying their joint prior distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$, inference may be based on the posterior distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta \mid x^n)$. In particular, inference for k may be based on the marginal posterior distribution of k:

$$\Pr(k = i \mid x^n) \quad i = 1, 2, 3, \ldots.$$

In theory the Bayesian approach appears to have many advantages over the maximum likelihood approach. In practice it presents new problems of its own, some of which we will address in this thesis.

1.1.5 Problems with the Bayesian approach

Computational problems

Bayesian methods provide a computational challenge, as integrals of the form (1.12) are not generally analytically tractable, and standard numerical methods for integration cannot be applied to give accurate results due to the high dimensionality of the parameter space.

In order to examine the problem in more detail we recall the missing data formulation of the model (Section 1.1.1) and note that if the allocation variables z^n were known then it might be possible to find the posterior distribution of the mixture model parameters given the completed data (x^n, z^n) analytically. For example, it is often possible to find a *conjugate* family of distributions $g(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta; a)$

(where a is a parameter of the parametric family of distributions $g(\cdot; a)$) such that if the prior distribution of the mixture parameters is of the form

$$p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = g(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta; a) \quad \text{for some } a \tag{1.13}$$

then the posterior distribution given the completed data (x^n, z^n) is of the same form

$$p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \mid x^n, z^n) = g(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta; a') \quad \text{for some } a'. \tag{1.14}$$

However, the posterior distribution of the parameters given only the data x^n may be written as

$$\begin{aligned} p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \mid x^n) &= \sum_{z^n} p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta, z^n \mid x^n) \\ &= \sum_{z^n} p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta \mid x^n, z^n) p(z^n \mid x^n) \end{aligned} \tag{1.15}$$

where the sum is over the set of all possible values for z^n, and so has k^n terms. Thus, even if the prior distribution of the model parameters is from the conjugate family, the posterior distribution is a mixture distribution with k^n components, and so is computationally intractable for moderate values of n.

Diebolt and Robert (1994) showed how a Markov Chain Monte Carlo (MCMC) method (a version of the Gibbs sampler implied by the Data Augmentation algorithm described by Tanner and Wong, 1987) can be used to approximate integrals of the form (1.12) in the case where k is considered known. We describe and illustrate this method in Chapter 2. More recently, Richardson and Green (1997) have applied the "reversible jump" methodology introduced by Green (1995), to perform inference for mixture models in which k is considered unknown. We present an alternative computational scheme for this case, based on ideas from the simulation of point processes (see Preston, 1976 and Ripley, 1977) in Chapter 4.

Choice of prior

Choice of suitable prior is generally a contentious issue in any situation where Bayesian methods are applied, and mixture models provide particular problems in this respect. In some situations we may have relatively strong prior information about the mixture model parameters, including the number of components present in the mixture — perhaps if we had a *training set* of *classified observations* (observations for which we know the component of the mixture from which they arose). The challenge is then how to represent this knowledge mathematically as a prior distribution for the parameters. However, we concentrate in this thesis on the case where we have very little prior information about the mixture model parameters, and we wish to use priors which are correspondingly "vague" about the value of the parameters.

Much effort has been expended by Bayesians in the search for so-called *non-informative* prior distributions which represent "ignorance" or lack of information about the parameters of a model. For example, Jeffreys (1967) suggests a method of specifying such a prior (the *Jeffreys prior*) when there is only a single unknown scalar parameter. More recently it seems to have been accepted that *any* prior distribution will contain *some* information about the parameters, and the emphasis has shifted towards the calculation of *reference priors* (introduced by Bernardo, 1979, and further developed by Berger and Bernardo, 1989, 1992) which result in posteriors which depend most heavily on the data (in a well-defined way). It is suggested that reference priors might provide a suitable starting point for a Bayesian analysis of the data, guaranteeing scientific objectivity without claiming to represent a definitively "correct" prior. More details, references and illuminating discussion may be found in Bernardo and Smith (1994) and Bernardo (1997).

Unfortunately the reference prior for most mixture models (and in particular for a mixture of normal distributions) gives an independent improper prior on the mixture model parameters, which cannot be used in a mixture context as it leads to improper posteriors for the component-specific parameters $\phi_1, \ldots, \phi_k$. We can see this by considering the posterior distribution of ϕ_1 given completed data (x^n, z^n), where z^n assigns no observations to the first component. If $\phi_1, \ldots, \phi_k$ are considered *a priori* independent then (x^n, z^n) contains no information about the parameters ϕ_1, and so the posterior distribution $p(\phi_1 \mid x^n, z^n)$, will be the same as the prior distribution $p(\phi_1)$. If this prior distribution is improper then $p(\phi_1 \mid x^n, z^n)$ will also be improper, and hence (from expression (1.15)) $p(\phi_1 \mid x^n)$ will also be improper.

In this thesis we use proper priors which attempt to be only "weakly informative" about the parameters, basing our priors for the univariate data on the hierarchical prior specification used by Richardson and Green (1997), and extending this specification to the case of bivariate data. However, we emphasise that inference can be very heavily influenced by the priors used, even when the priors appear to be relatively flat. In general we found that the prior used for the parameters of a mixture model is less critical when the model is to be used for density estimation, than when it is to be used for clustering. In particular, if k is considered unknown then independent and relatively flat priors on the mixture model parameters can be highly informative for k, and become more informative the flatter they become (see Section 4.10). The priors we use should therefore be regarded merely as convenient for the purposes of illustration, and we feel that much further work is required on the appropriate specification of priors, particularly in the cases where k is unknown, or the data is multivariate.

Symmetric posterior distributions

It is a well-known problem with mixture models that the parameters are not *identifiable* in that the likelihood

$$L(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \prod_{j=1}^{n} \left[\pi_1 f(x_j; \phi_1, \eta) + \cdots + \pi_k f(x_j; \phi_k, \eta) \right] \tag{1.16}$$

is symmetric in the components $1, \ldots, k$. That is, the likelihood is the same for all *permutations of the parameters*,

$$\nu(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \big((\pi_{\nu(1)}, \ldots, \pi_{\nu(k)}), (\phi_{\nu(1)}, \ldots, \phi_{\nu(k)}), \eta \big)$$

where ν is any permutation of $1, \ldots, k$. If the prior distribution of the parameters is also invariant under permutations of the parameters (as is usually the case if we have no real prior information about the components) then the posterior distribution will be similarly invariant, with up to $k!$ copies of each "genuine" mode. We will refer informally to the distribution possessing $k!$ symmetric sets of modes.

This symmetry in the posterior distribution of the parameters causes severe problems when attempting to perform inference regarding the individual components of the mixture, as the information we have for each component is exactly the same — there is no information which allows us to distinguish between the components in either the prior distribution or the likelihood. For example, by symmetry the classification probabilities

$$\Pr(Z_j = i \,|\, x^n) = \int \Pr(Z_j = i \,|\, x^n, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(\boldsymbol{\pi}, \boldsymbol{\phi}, \eta) \, d\boldsymbol{\pi} \, d\boldsymbol{\phi} \, d\eta \tag{1.17}$$

do not depend on i, and so are all equal to $1/k$. This is clearly a problem if we wish to use the mixture model for clustering, and we address this problem in Chapter 3.

1.2 Appendix: Non-parametric methods of density estimation

We examine briefly two non-parametric methods of performing density estimation: kernel density estimation and spline fitting to log-densities. These are illustrated on the galaxy data, and may be compared with the density estimates obtained using mixture models (see for example Figures 2.2, 2.3, 2.11, 2.12, and 4.6).

1.2.1 Kernel density estimation

Kernel density estimation is a non-parametric method of estimating the density which gave rise to observations $x_1, \ldots, x_n$. The estimate is of the form

$$\widehat{f}(x; h) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h} \right) \tag{1.18}$$

where K is a density known as the *kernel*, and h is a positive number, usually called the *bandwidth*. If K is chosen to be the density of the normal distribution with mean 0 and variance 1 then (1.18) becomes a mixture of n equally weighted normal distributions with means x_j ($j = 1, \ldots, n$) and variance h^2, and is referred to as a Gaussian kernel density estimator. The value of h used has a critical effect on the smoothness of the density estimator, larger values producing smoother density estimates.

Wand and Jones (1995) compare various methods of choosing h automatically in order to minimise estimates of the mean integrated squared error (MISE):

$$\text{MISE} = E\left[\int |\widehat{f}(x;h) - f(x)|\, dx\right]$$

and some of these methods are implemented by Venables and Ripley (1997a). One of the methods Wand and Jones (1995) recommend was introduced by Sheather and Jones (1991); Figure 1.6a shows the density estimate produced using this method on the galaxy data (computed using the S function `width.SJ` from Venables and Ripley, 1997a) and Figure 1.7a shows the corresponding estimate of the cumulative distribution function, obtained by numerically integrating the density estimate.

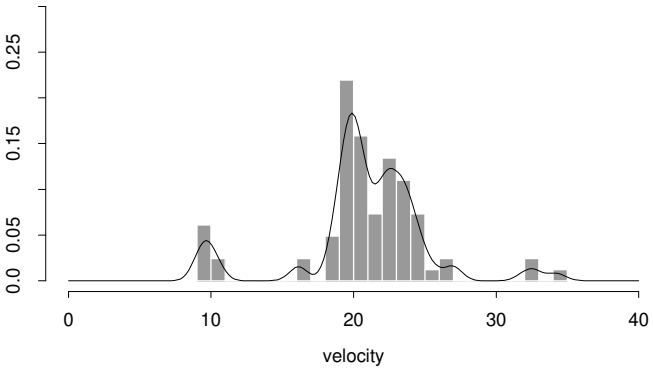
One problem with kernel density estimators is that the smoothness parameter is constant across the range of the data, making them less flexible than general mixture models. The double bump at the extreme right of Figure 1.6a might be viewed as a product of having to choose a smoothing parameter small enough to accurately model the data in other areas of the parameter space.

1.2.2 Fitting splines to log-densities

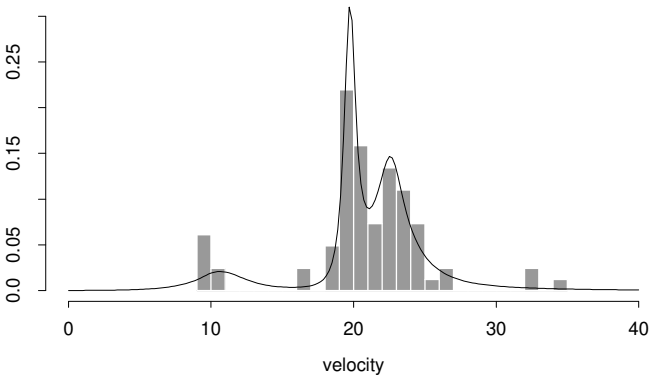
This method of density estimation fits a univariate density of the form

$$\widehat{f}(x) \propto \exp g(x;\theta) \tag{1.19}$$

to the data, where $g(\cdot;\theta)$ is a family of splines. We use a method of this type by Kooperberg and Stone (1992), in which g is a cubic spline, and the fit criterion is maximum likelihood, with a penalty on the number of knots used. This method is implemented in the S-PLUS library `logspline` by Charles Kooperberg, and further details are given in Venables and Ripley (1997b). The precise form of the penalty on the number of knots can be altered in the function `logspline.fit`, and we used the default penalty to produce estimates of the density (Figure 1.6b) and the cumulative distribution function (Figure 1.7b) for the galaxy data. The estimates can be seen to differ quite markedly from those obtained using the kernel density estimator described above and illustrated in the same Figures.

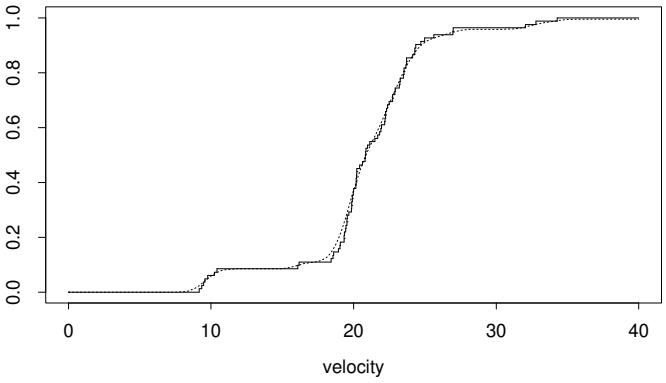


(a) Kernel density estimate

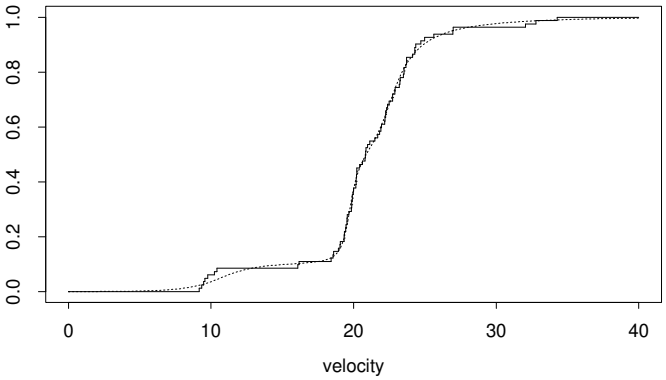


(b) Fitting splines to the log density

Figure 1.6: Histogram of the galaxy data overlaid with two different non-parametric density estimates. **Top:** Gaussian kernel density estimate using bandwidth chosen automatically according to a rule given by Sheather and Jones (1991); **Bottom:** Fitting splines to the log-density using the S-PLUS functions `logspline.fit` and `dlogspline` from the `logspline` library by Charles Kooperberg (see Venables and Ripley, 1997b).



(a) Kernel density estimate



(b) Fitting splines to the log density

Figure 1.7: The empirical cumulative distribution function for the galaxy data (solid line) overlaid with fitted cumulative density function (dashed line) obtained from **Top:** The Gaussian kernel density estimate using bandwidth chosen automatically according to a rule given by Sheather and Jones (1991); **Bottom:** Fitting splines to the log-density using the S-PLUS functions `logspline.fit` and `plogspline` from the `logspline` library by Charles Kooperberg (see Venables and Ripley, 1997b).

Chapter 2

Basic MCMC techniques for mixtures

In this chapter we consider the computational methods applicable to the analysis of a mixture with k components, where k is assumed known. Computational methods for the case where k is considered to be unknown are discussed in Chapter 4.

We recall that in a Bayesian analysis the parameters $\theta = (\boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ are treated as realisations of a random variable (Θ say), and that inference requires the evaluation of integrals of the form

$$E(F(\Theta) \mid x^n) = \int F(\theta) p(\theta \mid x^n)\, d\theta. \tag{2.1}$$

In most cases analytic evaluation of this integral is impossible, and if the integral is over a very high dimensional space (as is often the case) then traditional numerical methods of integration are impossible to apply accurately.

An alternative method of approximating such integrals is provided by Markov Chain Monte Carlo (MCMC) methods (see for example Gilks *et al.*, 1996), which rely on the construction of a Markov chain $\{\Theta^{(t)}\}$ with the property that the *sample path average*

$$\bar{F}_N = \frac{1}{N} \sum_{t=1}^{N} F(\Theta^{(t)}) \tag{2.2}$$

is a consistent estimator for $E(F(\Theta) \mid x^n)$, in that it converges almost surely to $E(F(\Theta) \mid x^n)$ as $N \to \infty$. Such a Markov chain can often be constructed in situations where it is not possible to sample from $p(\theta \mid x^n)$ directly, as is usually the case in the mixture model context.

We now describe briefly the theory underlying MCMC methods, and introduce the Gibbs sampler, an MCMC method which is particularly suited to the mixture model context. We then illustrate the use of the Gibbs sampler in fitting mixture distributions with a fixed number of components to both univariate and bivariate data.

2.1 Introduction to MCMC

A Markov chain in discrete time and general state space E is a sequence of random variables $(\Theta^{(0)}, \Theta^{(1)}, \dots)$, with $\Theta^{(t)} \in E$, which obeys the Markov Property in time. That is, given the current state $\Theta^{(t)}$ $(t \geq 0)$ the distribution of the next state $\Theta^{(t+1)}$ is independent of the past history of the chain, $(\Theta^{(0)}, \dots, \Theta^{(t-1)})$. The state space E may be quite general, and in particular it may include some discrete and some continuous components.

The distribution of a time-homogeneous Markov chain $\{\Theta^{(t)}\}$ on state space E is specified by the distribution of $\Theta^{(0)}$ and by its *transition kernel*

$$P(\theta, A) = \Pr(\Theta^{(t+1)} \in A \,|\, \Theta^{(t)} = \theta) \quad \text{for } A \subset E. \tag{2.3}$$

If $\theta^{(t)}$ has distribution ν on E, then $\theta^{(t+1)}$ has distribution νP given by

$$\nu P\,(A) = \int P(\theta, A)\, \nu(d\theta).$$

A Markov chain is said to have *invariant* (or *stationary*) distribution π if $\pi = \pi P$. That is if

$$\Theta^{(t)} \sim \pi \Rightarrow \Theta^{(t+1)} \sim \pi. \tag{2.4}$$

Intuitively we might expect $\Theta^{(t)}$ to be approximately distributed according to π for large t, provided it is able to reach all points in the state space in the support of π. We might then expect the sample path average $\bar{F}_N$ given by (2.2) to be an appropriate estimator for $\int F(\theta)\, \pi(d\theta)$. In order to make this precise we need to introduce the notion of a Markov chain being *irreducible*. The following is a generalisation of the definition of irreducibility for Markov chains in a discrete state space, where a chain is said to be irreducible if it is possible to move from any state to any other state in a finite number of steps.

Definition 1. A Markov chain is irreducible if there exists a probability distribution ϕ on E such that, for all $A \subset E$ with $\phi(A) > 0$

$$\Pr(\Theta^{(t)} \in A \text{ for some } t > 0 \,|\, \Theta^{(0)} = \theta) > 0$$

for all $\theta \in E$.

We can now state the following theorem (taken from Tierney, 1996, p65, with appropriate notational changes):

Theorem 2.1. *Suppose $(\Theta^{(0)}, \Theta^{(1)}, \dots)$ is an irreducible Markov chain on state-space E with transition kernel P and invariant distribution π, and let F be a real-valued function on E such that $\int |F(\theta)|\, \pi(d\theta) < \infty$. Then*

$$\Pr\left(\bar{F}_N \to \int F(\theta)\, \pi(d\theta) \;\middle|\; \Theta^{(0)} = \theta^{(0)}\right) = 1$$

for π-almost all $\theta^{(0)}$.

The theorem tells us that with suitable regularity conditions on F, the sample path average (2.2) of an irreducible Markov chain with invariant distribution $p(\theta \mid x^n)$ will converge almost surely to $E(F(\Theta) \mid x^n)$. Stronger distributional results are possible with further conditions, including asymptotic results on the speed of convergence. A full discussion can be found in Tierney (1996) with proofs given in Tierney (1994) and references therein. We discuss the practical implications of the speed of convergence in Section 2.1.2.

MCMC methods rely on the construction of a Markov chain fulfilling the conditions of Theorem 2.1 with invariant distribution $p(\theta \mid x^n)$, and construction of such a chain is often surprisingly straightforward. An algorithm which has found wide application and is particularly suited to the mixture context is the Gibbs sampler, which was given its name by Geman and Geman (1984) who used it to sample from Gibbs distributions.

2.1.1 The Gibbs sampler

The Gibbs sampler is a method of constructing a Markov chain with stationary distribution $p(\theta \mid x^n)$ when $\Theta \in E$ can be partitioned into components $(\Theta_1, \ldots, \Theta_r) \in E_1 \times \cdots \times E_r$, of possibly differing dimensions, where we cannot sample directly from $p(\theta \mid x^n) = p(\theta_1, \ldots, \theta_r \mid x^n)$ but can sample directly from the full conditional distributions

$$p(\theta_1 \mid x^n, \theta_2, \ldots, \theta_r), \ldots, p(\theta_r \mid x^n, \theta_1, \ldots, \theta_{r-1}).$$

Algorithm 2.1. Given the state $\Theta^{(t)} = \theta^{(t)}$ at time t, simulate a value for $\Theta^{(t+1)}$ in r steps as follows:

Step 1: sample $\Theta_1^{(t+1)}$ from $p(\theta_1 \mid x^n, \theta_2^{(t)}, \ldots, \theta_r^{(t)})$.

Step 2: sample $\Theta_2^{(t+1)}$ from $p(\theta_2 \mid x^n, \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_r^{(t)})$.

...

Step r: sample $\Theta_r^{(t+1)}$ from $p(\theta_r \mid x^n, \theta_1^{(t+1)}, \ldots, \theta_{r-1}^{(t+1)})$.

Proposition 2.2. *Algorithm 2.1 defines a Markov chain with stationary distribution $p(\theta_1, \ldots, \theta_r \mid x^n)$ which will be irreducible if the full conditional distribution of θ_j gives positive probability to any subset of E_j for $j = 1, \ldots, r$.*

Proof. $\Theta^{(1)}, \Theta^{(2)}, \ldots$ is clearly a Markov chain, as the distribution of $\Theta^{(t+1)}$ is completely determined by the value of $\Theta^{(t)}$. We will show that $p(\theta_1, \ldots, \theta_r \mid x^n)$ is the stationary distribution by showing that (2.4) holds. Suppose $(\Theta_1^{(t)}, \ldots, \Theta_r^{(t)})$ is distributed according to $p(\theta_1, \ldots, \theta_r \mid x^n)$. Then $(\Theta_1^{(t+1)}, \Theta_2^{(t)}, \ldots, \Theta_r^{(t)})$ is distributed according to

$$p(\theta_2, \ldots, \theta_r \mid x^n) p(\theta_1 \mid x^n, \theta_2, \ldots, \theta_r) = p(\theta_1, \theta_2, \ldots, \theta_r \mid x^n).$$

Similarly $(\Theta_1^{(t+1)}, \Theta_2^{(t+1)}, \Theta_3^{(t)}, \ldots, \Theta_r^{(t)})$ is distributed according to

$$p(\theta_1, \theta_2, \theta_3, \ldots, \theta_r \mid x^n)$$

and continuing in this manner we find that $(\Theta_1^{(t+1)}, \ldots, \Theta_r^{(t+1)})$ is distributed according to $p(\theta_1, \ldots, \theta_r \mid x^n)$.

Thus we have shown that

$$\Theta^{(t)} \sim p(\theta \mid x^n) \Rightarrow \Theta^{(t+1)} \sim p(\theta \mid x^n)$$

and so $p(\theta \mid x^n) = p(\theta_1, \ldots, \theta_r \mid x^n)$ is the stationary distribution of this Markov chain. □

2.1.2 Practical considerations: Starting points, burn-in and convergence

Algorithm 2.1, in common with all MCMC algorithms, requires us to choose (at random or otherwise) a starting value $\Theta^{(0)}$. Ideally we would sample $\Theta^{(0)}$ from the invariant distribution $p(\theta \mid x^n)$ of the Markov chain, but in most cases this is not possible, and so $\Theta^{(0)}$ is typically chosen either at random from the prior distribution for θ, or from near a mode of the posterior distribution. In order to reduce the dependence of the estimator $\bar{F}_N$ (given by (2.2)) on the choice of starting point, it is standard practice to discard the results of the first m iterations of the MCMC sampler, for suitably chosen m. These initial m iterations are sometimes referred to as the *burn-in* period. If the Markov chain is able to move quickly to and between all areas of the sample space with reasonable support in the posterior distribution (sometimes referred to as *good mixing*) then $\Theta^{(m)}$ will be virtually independent of $\Theta^{(0)}$ for relatively small m, and only a short burn-in period will be required before the choice of starting point becomes unimportant. In contrast, if the chain tends to get stuck in small areas of the state space for long periods of time then a long burn-in period may be required.

The mixing behaviour of the chain also determines the rate of convergence of the estimator $\bar{F}_N$ to $E\big(F(\theta) \mid x^n\big)$, and so determines the length of chain N required for acceptably accurate inference. Convergence will be more rapid if the chain exhibits good mixing behaviour. Many different formal methods have been proposed for choosing suitable values of m and N; see for example the review article by Cowles and Carlin (1996). We have taken a less formal approach, assessing the mixing behaviour of our chains on the basis of graphical output of the results, and comparing results of chains run from more than one starting point where we felt further investigation seemed necessary. We fixed on running the chain for $N = 20\,000$ iterations, discarding the first $m = 10\,000$ values. In most of our examples these values give a reasonable balance between computational cost and accuracy of inference (although $m = 10\,000$ is probably an unnecessarily large value for most of the cases we consider). Starting points were generally chosen by sampling at random from a relatively flat prior distribution over the sample space.

2.2 Gibbs sampling for mixtures of univariate normal distributions

We now illustrate the use of the Gibbs sampler in fitting a mixture of univariate normal distributions to the galaxy data which was introduced in Chapter 1. In particular we use the output of the Gibbs sampler to estimate the predictive density given by (1.9) (page 9).

We assume that the data $x^n = (x_1, \ldots, x_n)$ (with $n = 82$ for the galaxy data) are independent observations from a mixture of k (k assumed known) univariate normal distributions with density

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \pi_1 \mathcal{N}(x; \mu_1, \sigma_1^2) + \cdots + \pi_k \mathcal{N}(x; \mu_k, \sigma_k^2) \tag{2.5}$$

where $\mathcal{N}(x; \mu_i, \sigma_i^2)$ denotes the density function of the univariate normal distribution with mean μ_i and variance σ_i^2.

The assumption that k is known is rather restrictive, and we examine the case where k is considered unknown in Chapter 4. Previous work on this dataset with mixtures of normal distributions has suggested a value of k between 3 and 7, and for the purposes of illustration we will consider the cases $k = 3$ and $k = 6$ here.

2.2.1 Parameter priors

As explained in Chapter 1 (Section 1.1.5) we take a hierarchical approach to specifying the prior distribution of the parameters. We follow the suggestions of Richardson and Green (1997) who base their recommendation on a careful analysis of the sensitivity of the results to the choice of prior for three real datasets, one of which is the galaxy data considered here. Although they consider k to be unknown, they partition their prior as $p(k, \theta) = p(k)p(\theta \mid k)$, and so we use the prior they suggest for $p(\theta \mid k)$:

$$\mu_i \sim \mathcal{N}(\xi, \kappa^{-1}) \quad (i = 1, \ldots, k) \tag{2.6}$$

$$\sigma_i^{-2} \mid \beta \sim \Gamma(\alpha, \beta) \quad (i = 1, \ldots, k) \tag{2.7}$$

$$\beta \sim \Gamma(g, h) \tag{2.8}$$

$$\boldsymbol{\pi} \sim \mathcal{D}(\delta, \ldots, \delta) \tag{2.9}$$

where $\Gamma(n, \lambda)$ denotes the gamma distribution with mean n/λ and variance n/λ^2, $\mathcal{D}(\delta_1, \ldots, \delta_k)$ denotes the Dirichlet distribution on the simplex

$$\{(\pi_1, \ldots, \pi_{k-1}, 1 - \pi_1 - \cdots - \pi_{k-1}) : \pi_1 + \cdots + \pi_{k-1} \leq 1\}$$

with density proportional to

$$\pi_1^{\delta_1 - 1} \ldots \pi_{k-1}^{\delta_{k-1} - 1} (1 - \pi_1 - \cdots - \pi_{k-1})^{\delta_k - 1},$$

β is a hyperparameter, and $\xi, \kappa, \alpha, g, h$ and δ are constants which are to be found from the following formulae, which depend on the observed interval of variation

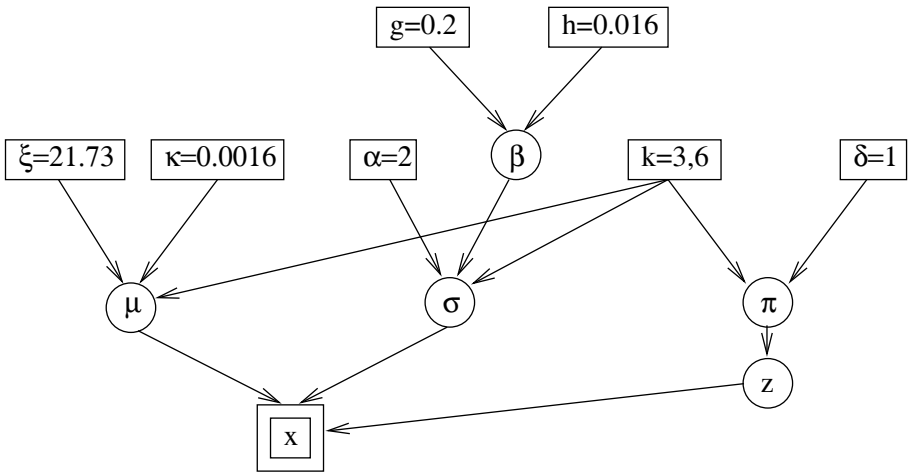


Figure 2.1: Directed Acyclic Graph (DAG) showing the hierarchical structure of prior, and values of constants used in our analysis of the galaxy data. We have used a single box around a quantity to indicates that it is considered to be a known constant, a circle to indicate an unknown quantity, and a double box to indicate observed data. The arrows indicate the conditional independence structure of the model; see Spiegelhalter *et al.* (1996) and references therein for more details of the uses and interpretation of DAGs. The structure of the prior and the values of the constants used follow Richardson and Green (1997).

of the data, and the length of this interval R:

$$\xi = \text{midpoint of the observed range of the data} = 21.73 \tag{2.10}$$

$$\kappa = \frac{1}{R^2} = 0.0016 \tag{2.11}$$

$$\alpha = 2 \tag{2.12}$$

$$g = 0.2 \tag{2.13}$$

$$h = \frac{100g}{\alpha R^2} = 0.016 \tag{2.14}$$

$$\delta = 1. \tag{2.15}$$

The hierarchical structure of the prior, and values of the constants used are shown in Figure 2.1. The prior chosen for the means and variances is not the natural conjugate prior, but gives some of the advantages of conjugacy when implementing the Gibbs sampler. The values chosen for ξ and κ give a prior for μ which is fairly flat over the observed range of data. The choice of $\alpha = 2$ with a relatively flat hyperprior on the hyperparameter β expresses the belief that the σ_i^2 are similar, without being very informative about their absolute size. This is less restrictive than constraining all the variances to be equal. The Dirichlet distribution is the natural conjugate distribution for the mixture proportions, and choosing $\delta = 1$ gives a uniform prior over the space $\pi_1 + \cdots + \pi_k = 1$.

2.2.2 The Gibbs sampler

In order to implement the Gibbs sampler, we return to the missing data formulation of the model described in the previous chapter (Section 1.1.1, page 3) and consider the *augmented* parameter vector $\theta = (z^n, \beta, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$, which is the vector of parameters and hyperparameters, augmented by the (unknown) allocation variables z^n. If we partition θ into its 5 natural component parts, then due to the carefully chosen form of the hierarchical prior the full conditional posterior distributions can be found analytically, and are as follows (using $\mid \cdots$ to denote conditioning on the values of all other parameters and on the data x^n):

$$p(z_j = i \mid \cdots) \propto \pi_i \mathcal{N}(x_j; \mu_i, \sigma_i^2) \tag{2.16}$$

$$\beta \mid \cdots \sim \Gamma\Big(g + k\alpha, h + \sum_i \sigma_i^{-2}\Big) \tag{2.17}$$

$$\boldsymbol{\pi} \mid \cdots \sim \mathcal{D}(\delta + n_1, \ldots, \delta + n_k) \tag{2.18}$$

$$\mu_i \mid \cdots \sim \mathcal{N}\left(\frac{\sigma_i^{-2} \sum_{j:z_j=i} x_j + \kappa\xi}{\sigma_i^{-2} n_i + \kappa}, (\sigma_i^{-2} n_i + \kappa)^{-1}\right) \tag{2.19}$$

$$\sigma_i^{-2} \mid \cdots \sim \Gamma\Big(\alpha + \tfrac{1}{2} n_i, \beta + \tfrac{1}{2} \sum_{j:z_j=i} (x_j - \mu_i)^2\Big) \tag{2.20}$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, n = 82$, where n_i is the number of observations allocated to class i

$$n_i = \#\{j : z_j = i\}.$$

A single iteration of the Gibbs sampler (Algorithm 2.1) consists of simulating from each of these conditional posterior distributions in turn (we updated the parameters in the order given), and by construction the stationary distribution of the Markov chain defined by this algorithm will be $p(z^n, \beta, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2} \mid x^n)$.

2.2.3 Results

The Gibbs sampler was run for $N = 20\,000$ iterations for both $k = 3$ and $k = 6$, starting at parameter values drawn from the prior (and thus starting in a region of parameter space with low likelihood), and the sampled values of $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$ are shown in Figures 2.5 ($k = 3$) and 2.6 ($k = 6$). The runs took 71 seconds ($k = 3$) and 117 seconds[1] ($k = 6$). The first $m = 10\,000$ sample points were discarded as burn-in, and estimates of the predictive density:

$$\begin{aligned} p(x_{n+1} \mid x^n) &= \int p(x_{n+1} \mid x^n, \theta) p(\theta \mid x^n)\, d\theta \\ &\approx \frac{1}{N-m} \sum_{t=m+1}^{N} p(x_{n+1} \mid \theta^{(t)}) \end{aligned} \tag{2.21}$$

[1] CPU time on a Sun UltraSparc 200 workstation

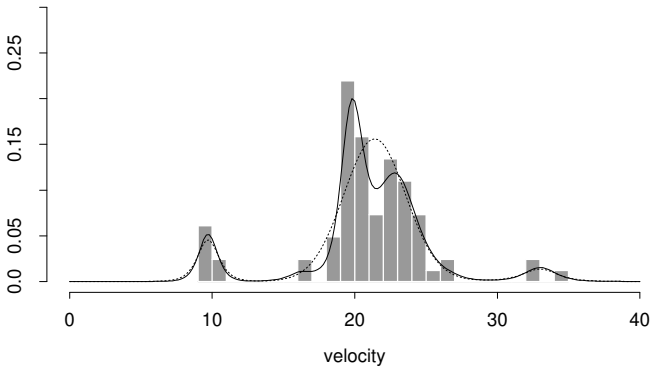


Figure 2.2: Histogram of the galaxy data overlaid with estimates of predictive density (2.21) based on: **Dotted line:** the output of the Gibbs sampler when fitting $k = 3$ normal distributions to the data (see Figure 2.5); **Solid line:** the output of the Gibbs sampler when fitting $k = 6$ normal distributions to the data (see Figure 2.6).

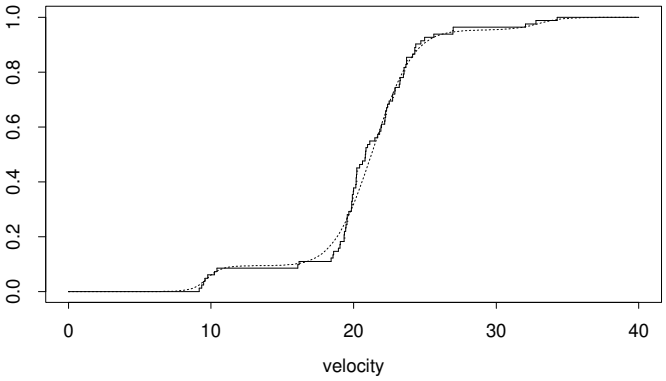
were formed using the remaining $10\,000$ sample points. These estimates are shown overlaid on a histogram of the data in Figure 2.2, and the corresponding estimates of the cumulative distribution function (obtained by integrating the density estimates numerically) are shown in Figure 2.3. As might be expected, the density estimate for $k = 6$ is somewhat less smooth than for $k = 3$ which may not have enough flexibility to adequately model the observed data. Both are smoother than the Gaussian kernel density estimate shown in the previous chapter (Figure 1.6a), and differ significantly from the `logspline` density estimate (Figure 1.6b) in that they each possess a mode near 34. Discussion of methods of choosing a suitable value for k is deferred to Chapter 4.

2.2.4 Mixing properties and label-switching

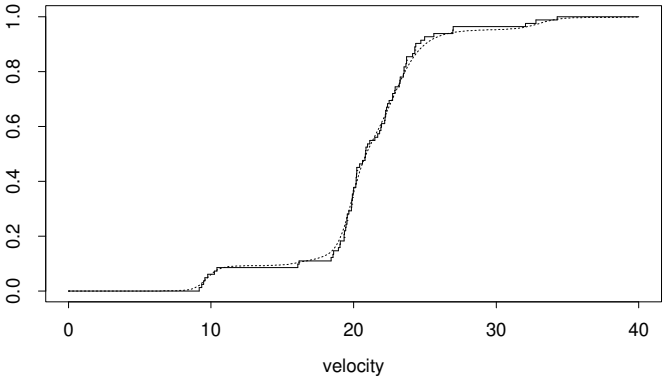
Some previous analyses of normal mixture distributions using Gibbs sampling have reported problems with "trapping states". Robert (1996, page 448) reports

> "Despite the theoretical irreducibility of the chain, the [Gibbs sampler] becomes effectively trapped when one of the components of the mixture is allocated very few observations."

These trapping states appear to be due to allowing components which contain few observations to have very small variance, and thus allowing the parameter values to approach singularities in the likelihood corresponding to setting the mean of one mixture component to be at an observed data point, and allowing the variance of this component to tend to zero. This can be avoided by constraining the component variances to be equal, or (less restrictively) by putting a hierarchical prior structure on the variances which favours "similar" variances for the components,



(a) Fitting $k = 3$ normal distributions



(b) Fitting $k = 6$ normal distributions

Figure 2.3: Empirical cumulative distribution function for the galaxy data (solid line) overlaid with fitted cumulative density function (dashed line) obtained by numerically integrating the estimated predictive densities shown in Figure 2.2.

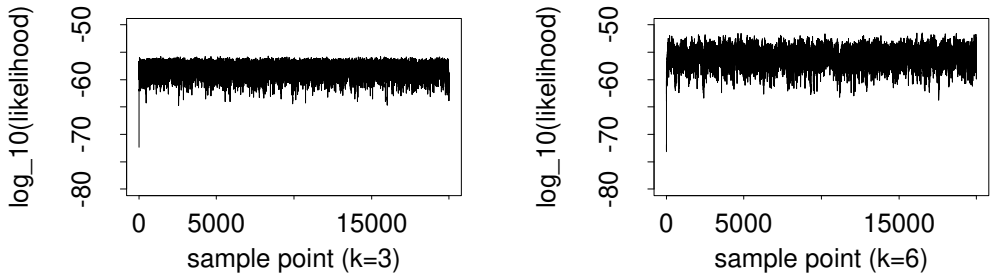


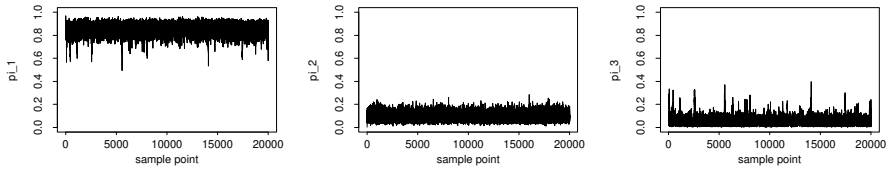
Figure 2.4: $\log_{10}$(likelihood) of the sampled parameter values when applying the Gibbs sampler to fit a mixture of normal distributions (2.5) to the galaxy data. **Left:** $k = 3$; **Right:** $k = 6$. In both cases the sampler moves quickly to an area of higher likelihood and does not appear to get stuck in areas of low or high likelihood.

as we do here. This has a regularising effect on the posterior density, reducing or eliminating the local maxima near singularities in the likelihood. Figure 2.4 shows the $\log_{10}$ likelihood for the $20\,000$ sampled parameter values for each value of k. In both cases the sampler moves quickly from a region of low likelihood to a region of higher likelihood, and there is no evidence of trouble with trapping states due to singularities in the likelihood.

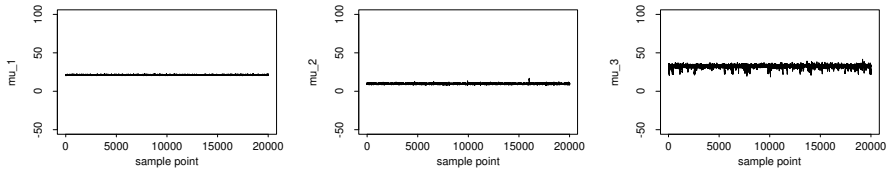
If we compare the graphs of the sampled parameter values for $k = 3$ (Figure 2.5) and $k = 6$ (Figure 2.6), we notice that they exhibit quite contrasting behaviour. For $k = 3$ the sampled values of the parameters lie in a relatively restricted part of the parameter space, while for $k = 6$ they explore the parameter space more freely, with periods of stability alternating with more "noisy" behaviour (particularly evident in the graphs of the means).

The restricted movement about the sample space for $k = 3$ may be because the sampler is stuck in a local mode of the posterior distribution (and thus mixing poorly) or may be because the posterior distribution of the parameters for $k = 3$ is concentrated in this restricted region of the parameter space. We investigated this by running the simulation for $k = 3$ a further ten times, with each simulation starting from a different point randomly drawn from the prior. All ten runs gave similar results, with the three fitted components settling down to have a heavily weighted component with mean near 20, and two smaller components with means near 10 and 34; the results from one of these runs are shown in Figure 2.7. Comparing Figures 2.5 and 2.7, the only obvious difference is in the labelling of the components: in the first run component 1 has mean near 20, component 2 has mean near 10 and component 3 has mean near 34; in the second run components 1 and 3 have switched.

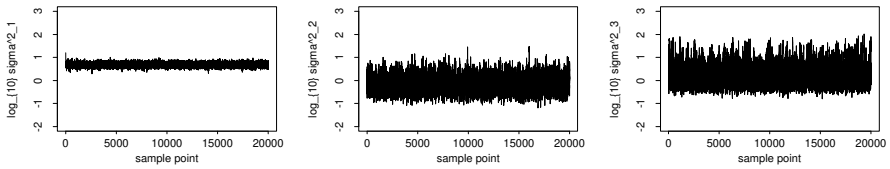
This *label-switching* is due to the symmetries in the posterior distribution of the parameters described in the previous chapter (Section 1.1.5, page 13). For $k = 3$ the posterior distribution of the model parameters has $3!$ symmetric sets



(a) Mixture proportions

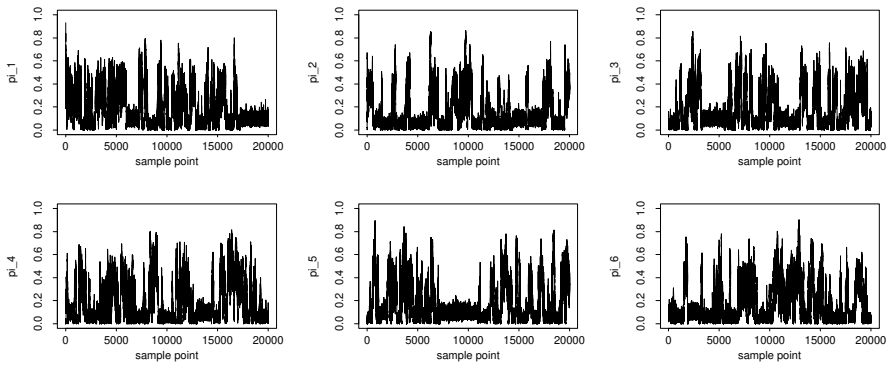


(b) Means

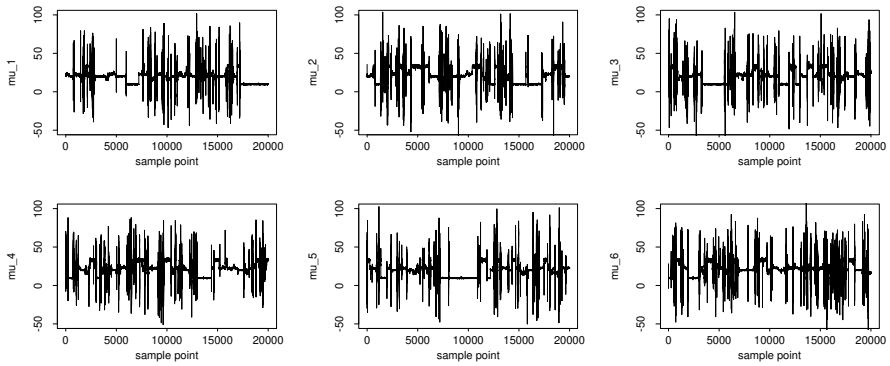


(c) $\log_{10}$ (variances)

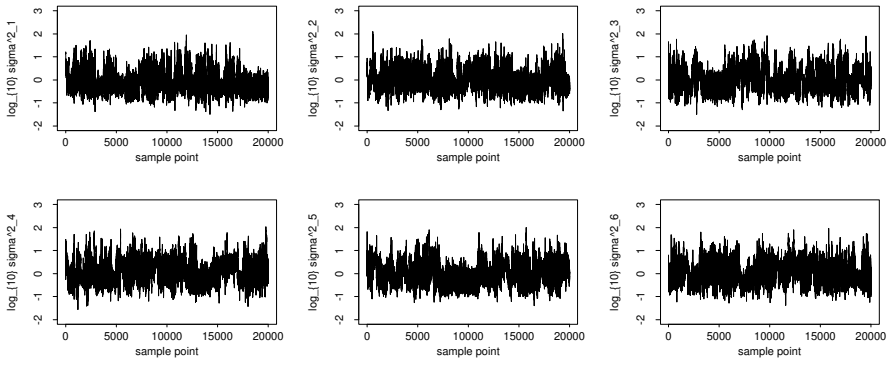
Figure 2.5: Sampled values of mixture proportions, means and $\log_{10}$(variances) when fitting $k = 3$ normal distributions to the galaxy data, using the parameter priors given in Section 2.2.1 and the Gibbs sampler steps described in Section 2.2.2. The scale of the graph of the means has been chosen to be comparable with later figures. The results of a second run of the Gibbs sampler for $k = 3$ starting from a different random starting point are shown in Figure 2.7.



(a) Mixture proportions

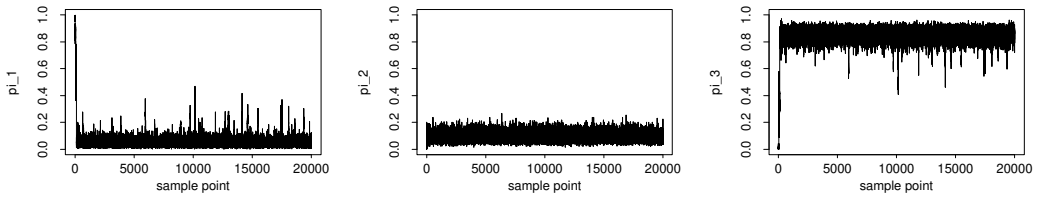


(b) Means

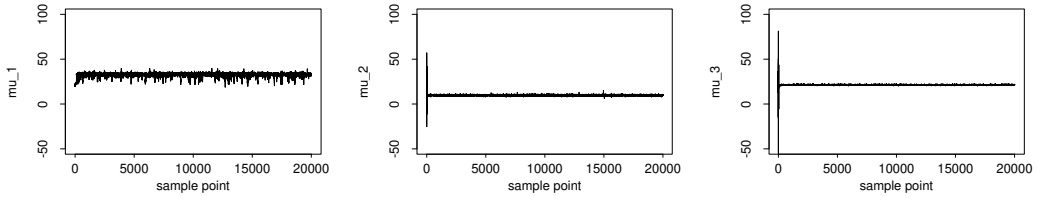


(c) $\log_{10}$ (variances)

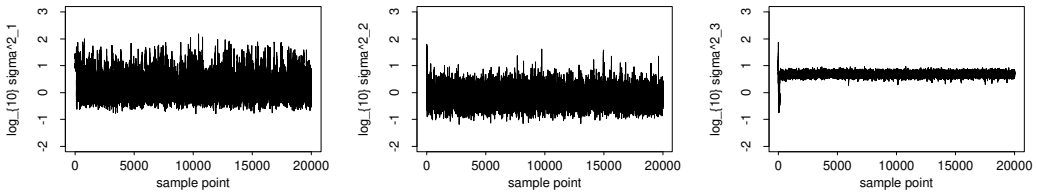
Figure 2.6: Sampled values of mixture proportions, means and $\log_{10}$(variances) when fitting $k = 6$ normal distributions to the galaxy data, using the parameter priors given in Section 2.2.1 and the Gibbs sampler steps described in Section 2.2.2.



(a) Mixture proportions



(b) Means



(c) $\log_{10}$ (variances)

Figure 2.7: Sampled values of mixture proportions, means and $\log_{10}$(variances) for a second run of the Gibbs sampler used to fit $k = 3$ normal distributions to the galaxy data. The second run used the same priors and Gibbs sampler as the first run (see Figure 2.5) but used a different random starting point. The graphs are similar to those for the first run, except that the labels of components 1 and 3 have been "switched".

of modes, corresponding to the $3!$ different ways of labelling the components. Figures 2.5 and 2.7 show samplers which are exploring 2 different such sets of modes. The fact that for each of our ten starting points our sampler explores only one of the $3!$ sets of modes suggests that in this case the $3!$ sets of modes lie in $3!$ well-separated regions of parameter space, and that it will typically take a long time for the sampler to move from one of these regions to another.

In contrast, the results for $k = 6$ (Figure 2.6) show that within this one run label-switching occurs frequently, as the sampler moves between some of the $6!$ sets of modes in the posterior distribution. The switching behaviour is particularly noticeable in the graphs of the means as they switch between several distinctive levels and "noise".

To some extent the presence of label-switching is evidence of good mixing, and lack of label-switching is evidence of poor mixing. However, since we *know* that the posterior distribution of the parameters possesses this symmetry, it is only necessary for the sampler to explore *one* of the symmetric sets of modes adequately, and if necessary we can then permute each sample point to give a sample which explores all $k!$ modes.

The similarities shown between the ten different runs for $k = 3$, exemplified by Figures 2.5 and 2.7, suggests that in each case the sampler is indeed exploring one set of modes adequately, and not getting stuck for long periods in local modes of the posterior distribution. The lack of label-switching within the runs for $k = 3$ need not then concern us greatly, and in some ways it provides a natural way of identifying the components of the mixture, allowing us to use the results sensibly for clustering and giving sensible estimates of the scaled predictive component densities:

$$\int \pi_i \mathcal{N}(x; \mu_i, \sigma_i^2) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2} \mid x^n)\, d\boldsymbol{\pi}\, d\boldsymbol{\mu}\, d\boldsymbol{\sigma^2} \approx \frac{1}{N-m} \sum_{t=m+1}^{N} \pi_i^{(t)} \mathcal{N}\left(x; \mu_i^{(t)}, \sigma_i^{2(t)}\right) \quad (2.22)$$

which are shown in Figure 2.8.

In contrast, estimates of the scaled predictive component densities based on the results of the Gibbs sampler for $k = 6$ (Figure 2.10a) demonstrate the problems caused by the symmetry on the posterior distribution of the parameters when attempting to perform inference relating to the individual components of the mixture. The label-switching mixes up any potentially interpretable components, to give 6 very similar density estimates. It would seem to be helpful if we could "undo" the label-switching in some way, so that the sampler only explores one set of modes of the posterior distribution; Chapter 3 is devoted to investigation of this problem. We applied methods developed there (Algorithm 3.2, Section 3.2) to the results for $k = 6$. Figure 2.9 shows the resulting sampled means with the label-switching "undone", and Figure 2.10 contrasts estimates of the six scaled

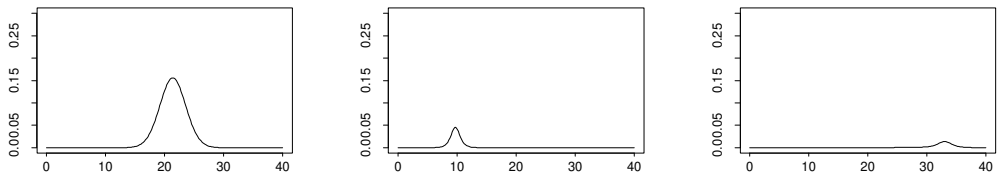


Figure 2.8: Estimates of the three scaled predictive component densities (2.22) when fitting a mixture of $k = 3$ normal distributions to the galaxy data. The estimates are based on the output of the Gibbs sampler shown in Figure 2.5, with the first $m = 10\,000$ sample points discarded as burn-in.

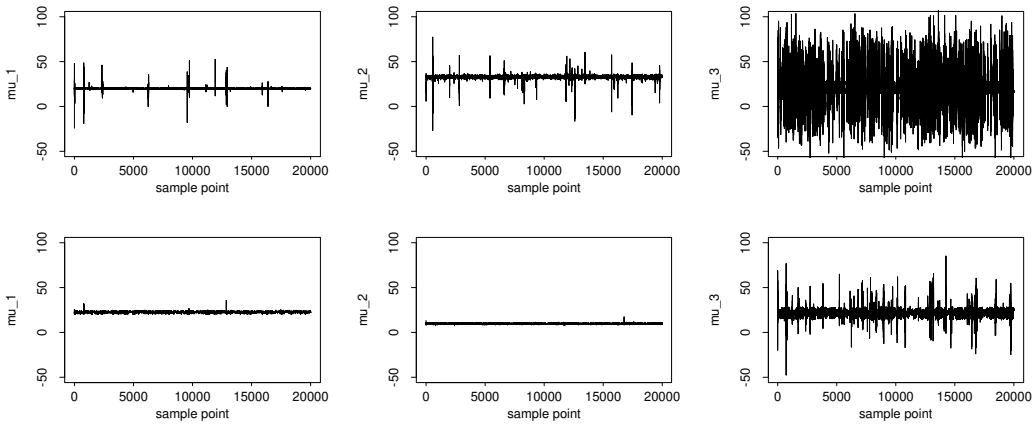
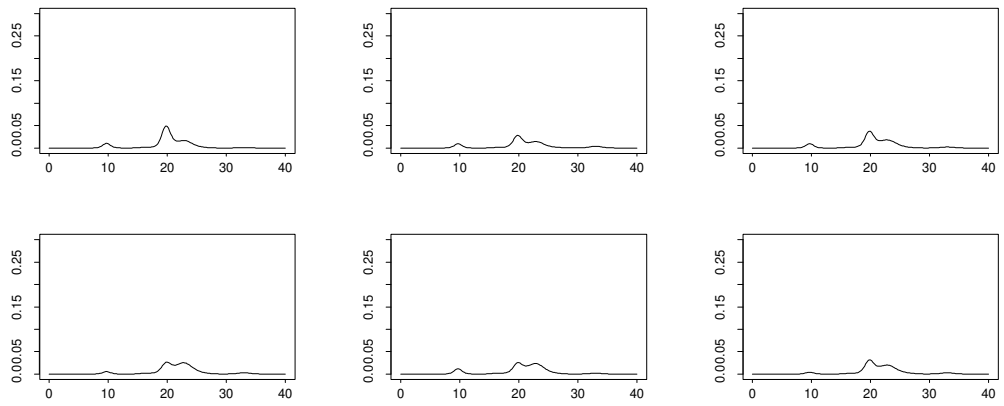
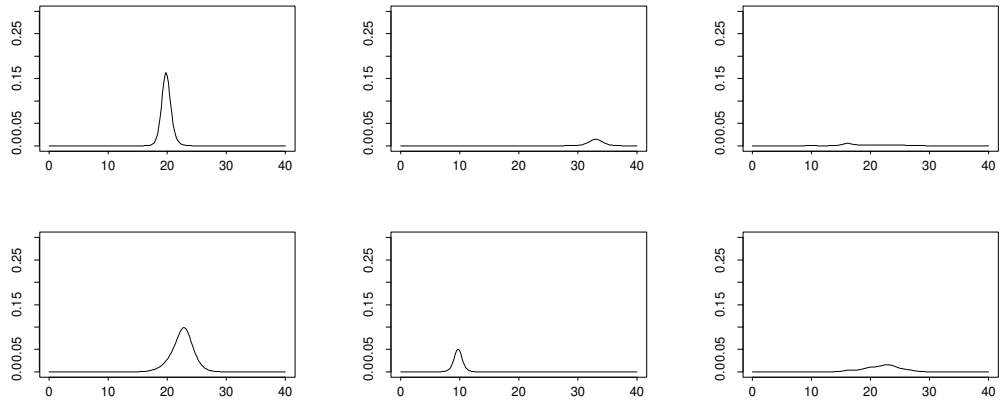


Figure 2.9: Sampled values of the means $(\mu_1, \ldots, \mu_6)$ when fitting $k = 6$ normal distribution to the galaxy data, with label-switching "undone" by methods described in Chapter 3 (Algorithm 3.2, Section 3.2). Compare these with the original sampled values of the means (with label switching present) in Figure 2.6

predictive component densities (2.22) based on the raw output of the Gibbs sampler and on the same output with label-switching "undone". From the results with the label-switching "undone" we see that the run for $k = 6$ has the same components near 10 and 34 as the run for $k = 3$, and has split the component near 20 into 4 distinct components. See Section 3.4.1 for further details and discussion of these results.



(a) Based on raw output of the Gibbs sampler



(b) Based on output of Gibbs sampler with label-switching "undone"

Figure 2.10: Estimates of the six scaled predictive component densities (2.22) when fitting a mixture of $k = 6$ normal distributions to the galaxy data. **Top:** Estimates based on the raw output of the Gibbs sampler, shown in Figure 2.6; **Bottom:** Estimates based on the output of the Gibbs sampler with label-switching "undone" using Algorithm 3.2 (Section 3.2), which is shown in Figure 3.3. Note how the presence of label-switching mixes together any interpretable components to create six similar density estimates, and how "undoing" the label-switching facilitates identification of interpretable individual components being fitted to the data.

2.3 Gibbs sampling for mixtures of univariate t-distributions

It is straightforward to modify the Gibbs sampler described in the previous section to deal with the case where the observations are assumed to arise not from a mixture of normal distributions, but a mixture of t-distributions, with known degrees of freedom p:

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \pi_1 t_p(x; \mu_1, \sigma_1^2) + \cdots + \pi_k t_p(x; \mu_k, \sigma_k^2) \tag{2.23}$$

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 $p\sigma_i^2/(p-2)$. To ensure finite variance we require $p > 2$.

The t-distribution is similar to the normal distribution but has "fatter tails"; the smaller the degrees of freedom, the fatter the tails. Real data often has more observations in the tails than might be expected under the assumption of normality, and so in many cases a mixture of t-distributions will provide a better model than a mixture of normal distributions.

The t-distribution has a well-known representation in terms of the normal and χ_p^2 distributions:

If $Y \sim N(0,1)$ and $q \sim \chi_p^2/p$ independently of Y then $X = \mu + \dfrac{\sigma Y}{\sqrt{q}} \sim t_p(\mu, \sigma^2)$.

We can thus introduce latent variables $q_j \sim \chi_p^2/p$ $(j = 1, \ldots, n)$ and rewrite the mixture model (2.23):

$$p(x_j \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}, q_j) = \pi_1 \mathcal{N}(x_j; \mu_1, \sigma_1^2/q_j) + \cdots + \pi_k \mathcal{N}(x_j; \mu_k, \sigma_k^2/q_j) \quad (j = 1, \ldots, n). \tag{2.24}$$

We now consider the parameter vector augmented by both z^n and q^n, $\theta = (z^n, q^n, \beta, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$. Partitioning θ into its 6 natural component parts, and using the same priors as in the normal case, we obtain the full conditional posteriors:

$$p(z_j = i \mid \cdots) \propto \pi_i \mathcal{N}(x_j; \mu_i, \sigma_i^2/q_j) \tag{2.25}$$

$$q_j \mid \cdots \sim \Gamma\left(\tfrac{1}{2}[p+1], \tfrac{1}{2}[p + \sigma_{z_j}^{-2}(x_j - \mu_{z_j})^2]\right) \tag{2.26}$$

$$\beta \mid \cdots \sim \Gamma\left(g + k\alpha, h + \sum_i \sigma_i^{-2}\right) \tag{2.27}$$

$$\boldsymbol{\pi} \mid \cdots \sim \mathcal{D}(\delta + n_1, \ldots, \delta + n_k) \tag{2.28}$$

$$\mu_i \mid \cdots \sim \mathcal{N}\left(\frac{\sigma_i^{-2}\sum_{j:z_j=i} q_j x_j + \kappa\xi}{\sigma_i^{-2}\sum_{j:z_j=i} q_j + \kappa}, \left(\sigma_i^{-2}\sum_{j:z_j=i} q_j + \kappa\right)^{-1}\right) \tag{2.29}$$

$$\sigma_i^{-2} \mid \cdots \sim \Gamma\left(\alpha + \tfrac{1}{2}n_i, \beta + \tfrac{1}{2}\sum_{j:z_j=i} q_j(x_j - \mu_i)^2\right) \tag{2.30}$$

for $j = 1, \ldots, n$ and $i = 1, \ldots, k$. An alternative is to partition θ into 5 component parts, $\theta = ((z^n, q^n), \beta, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$ and sample from the joint distribution of (z^n, q^n) given the other parameters, thus replacing (2.25) and (2.26) with

$$p(z_j = i \mid \cdots) \propto \pi_i t_p(x_j; \mu_i, \sigma_i^2) \tag{2.31}$$

$$q_j \mid z^n, \cdots \sim \Gamma\big(\tfrac{1}{2}[p+1], \tfrac{1}{2}[p + \sigma_{z_j}^{-2}(x_j - \mu_{z_j})^2]\big). \tag{2.32}$$

We used this second version of the Gibbs sampler to fit a mixture of $k = 3$ t_4 components to the galaxy data, using the same prior distributions as for the normal case (Section 2.2.1). Figure 2.11 shows the resulting estimated predictive density (2.21) and Figure 2.12 shows the corresponding estimate of the cumulative distribution function obtained by numerical integration. The Gibbs sampler was started from a point drawn at random from the prior, and was run for $20\,000$ iterations, the first $10\,000$ being discarded as burn-in. The sampler moved quickly from the random starting point to an area of higher likelihood. However, the mixing behaviour of the sampler was poor. Figure 2.13 shows the sampled values of the means (the first $10\,000$ samples having being discarded as burn-in). Most of the sampled points consist of a mean near 20, a mean near 23 and a mean near 10, with label-switching behaviour occurring between the first two. However, for about 300 iterations around iteration $14\,000$ the means near 20 and 23 are replaced by means near 30 and 21. This is an example of "genuine" multimodality (as distinct from the "symmetric" multimodality which causes label-switching) in the posterior distribution of the means, and the results indicate that the sampler finds it difficult to move between these different modes. Longer runs of the Gibbs sampler will therefore be required in this case for accurate inference, unless we can improve the mixing behaviour of the sampler. We will see a possible way of improving the mixing behaviour in Chapter 4, where we allow the number of components k to vary.

The presence of both "genuine" and "symmetric" multimodality in the posterior distribution of the parameters means that performing sensible inference for the individual components of the mixture is not straightforward, and is deferred to the following chapter (Section 3.4.2).

2.4 Gibbs sampling for mixtures of multivariate normals

We now consider data $x^n = (x_1, \ldots, x_n)$ which are assumed to be independent observations from a mixture of k (k assumed known) *multivariate* normal distributions in r dimensions, with density

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \pi_1 \mathcal{N}_r(x; \mu_1, \Sigma_1) + \cdots + \pi_k \mathcal{N}_r(x; \mu_k, \Sigma_k) \tag{2.33}$$

where $\mathcal{N}_r(x; \mu_i, \Sigma_i)$ denotes the density function of the multivariate normal distribution with mean μ_i and variance-covariance matrix Σ_i.

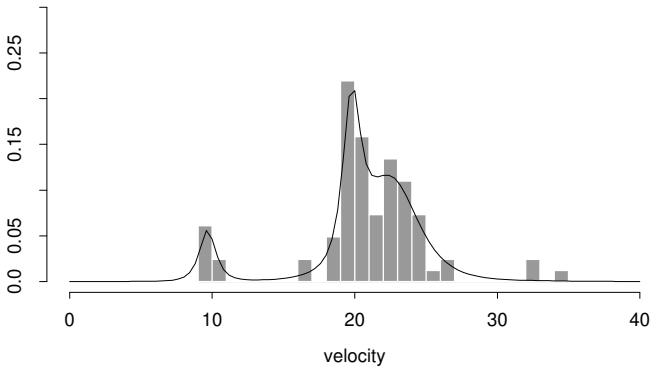


Figure 2.11: Histogram of the galaxy data overlaid with estimate of predictive density obtained by fitting a mixture of $k = 3$ t_4 distributions by Gibbs sampling.

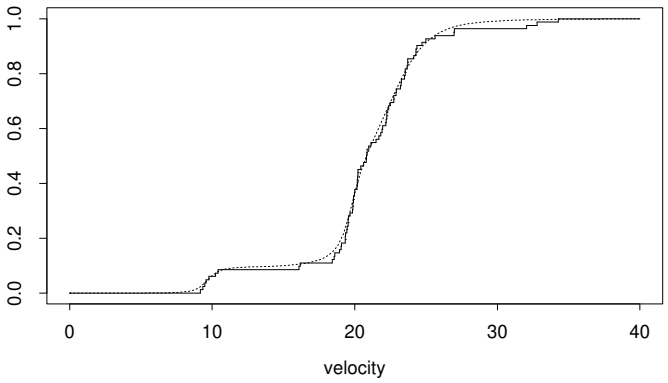


Figure 2.12: Empirical cumulative distribution function for the galaxy data (solid line) overlaid with fitted cumulative density function (dashed line) obtained by numerically integrating the estimated predictive density shown in Figure 2.11, based on fitting a mixture of $k = 3$ t_4 distributions to the data .

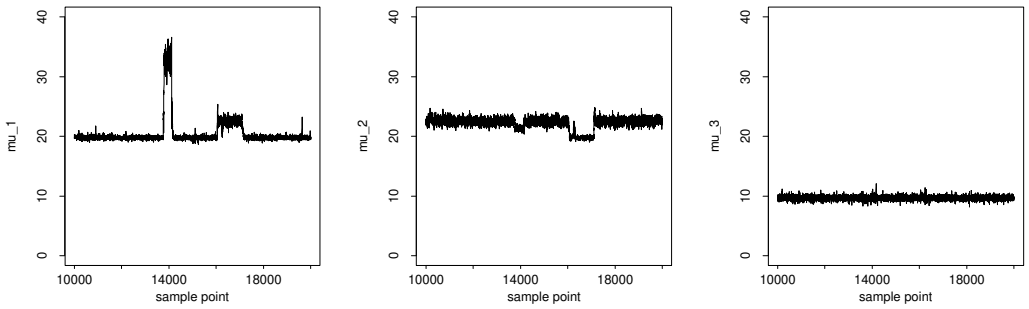


Figure 2.13: Sampled values of means for the three components when using the Gibbs sampler to fit $k = 3$ t_4 distributions to the galaxy data, the first $10\,000$ iterations having been discarded as burn-in. Label-switching occurs between the first and second components around iterations $16\,000$ and $17\,000$. However, the change in behaviour for about 300 iterations around iteration $14\,000$ is not label-switching, but is due to "genuine" multimodality in the posterior distribution of the means, as the means near 20 and 23 are replaced by means near 34 and 21. The sampler does not move freely between these modes, and so is not mixing well.

2.4.1 Prior distributions

In a multivariate setting, the Wishart distribution replaces the gamma distribution as the conjugate prior for the inverse variances. The Wishart distribution in r dimensions with parameters m and A, which we will denote by $\mathcal{W}_r(m, A)$, 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

$$\mathcal{W}_r(V; m, A) = K|A|^{-\frac{m}{2}}|V|^{\frac{m-r-1}{2}} \exp\left\{-\tfrac{1}{2}\text{tr}(A^{-1}V)\right\} I(V \text{ positive definite}) \tag{2.34}$$

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, (2.34) 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 (2.34). 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.

Noting (from the form of (2.34)) that the gamma distribution with parameters (α, β) is the same as the Wishart distribution in 1 dimension with parameters

$(2\alpha, (2\beta)^{-1})$, we consider the following to be the r-dimensional analogue of the priors (2.6)–(2.9) used in the univariate case.

$$\mu_i \sim \mathcal{N}_r(\xi, \kappa^{-1}) \tag{2.35}$$

$$\Sigma_i^{-1} \mid \beta \sim \mathcal{W}_r(2\alpha, (2\beta)^{-1}) \tag{2.36}$$

$$\beta \sim \mathcal{W}_r(2g, (2h)^{-1}) \tag{2.37}$$

$$\boldsymbol{\pi} \sim \mathcal{D}(\boldsymbol{\pi}; \delta, \ldots, \delta) \tag{2.38}$$

for $i = 1, \ldots, k$, where κ, β and h are $r \times r$ matrices, ξ is an $r \times 1$ vector, and α, δ and g are scalars.

We consider only two-dimensional data, and used the following formulae to calculate values for the constants:

$$\xi = (\xi_1, \xi_2) \tag{2.39}$$

$$\kappa = \begin{pmatrix} \frac{1}{R_1^2} & 0 \\ 0 & \frac{1}{R_2^2} \end{pmatrix} \tag{2.40}$$

$$\alpha = 3 \tag{2.41}$$

$$g = 0.3 \tag{2.42}$$

$$h = \begin{pmatrix} \frac{100g}{\alpha R_1^2} & 0 \\ 0 & \frac{100g}{\alpha R_2^2} \end{pmatrix} \tag{2.43}$$

$$\delta = 1 \tag{2.44}$$

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.

These formulae were chosen by analogy with the univariate case (equations (2.10)–(2.15)). Recall that in the univariate case the choice of $\alpha = 2$ was made to express the belief that the variances of the components are similar, without restricting them to be equal. In two dimensions we felt that a slightly stronger constraint would be appropriate, and so increased α to 3, making a corresponding change in g from 0.2 to 0.3. We note that the prior on β

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

is an improper distribution, but careful checking of the necessary integrals shows that the posterior distributions are proper. We emphasise that these priors should be viewed merely as convenient for the purposes of illustration. We feel that further work is required on defining appropriate priors for the case of multivariate data, particularly for the covariance matrices which might benefit from an eigen-decomposition, enabling the specification of priors which favour components which are a similar size, shape, orientation, or some combination of these three (such a decomposition is considered by Banfield and Raftery, 1993).

2.4.2 Full conditional posterior distributions

The full conditional posterior distributions of the parameters and hyperparameters, for use in the Gibbs sampler, are as follows:

$$p(z_j = i \mid \cdots) \propto \pi_i \mathcal{N}_r(x_j; \mu_i, \Sigma_i) \tag{2.45}$$

$$\beta \mid \cdots \sim \mathcal{W}_r\Big(2g + 2k\alpha,\, \Big[2h + 2\sum_i \Sigma_i^{-1}\Big]^{-1}\Big) \tag{2.46}$$

$$\boldsymbol{\pi} \mid \cdots \sim \mathcal{D}(\delta + n_1, \ldots, \delta + n_k) \tag{2.47}$$

$$\mu_i \mid \cdots \sim \mathcal{N}_r\Big((n_i\Sigma_i^{-1} + \kappa)^{-1}(n_i\Sigma_i^{-1}\bar{x}_i + \kappa\xi), (n_i\Sigma_i^{-1} + \kappa)^{-1}\Big) \tag{2.48}$$

$$\Sigma_i^{-1} \mid \cdots \sim \mathcal{W}_r\Big(2\alpha + n_i,\, \Big[2\beta + \sum_{j:z_j=i} (x_j - \mu_i)(x_j - \mu_i)^T\Big]^{-1}\Big) \tag{2.49}$$

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:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j:z_j=i} x_j.$$

2.4.3 Example: The Old Faithful data

The Old Faithful data (the version from Härdle, 1991, also considered by Venables and Ripley, 1994) 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. The data are given in the appendix to this thesis (Section A.2), and a scatter plot of the data is shown in Figure 2.14. The data is clearly multimodal, and would appear to be suitably modelled as a mixture of two bivariate normal distributions. Figure 2.15 shows an estimate of the predictive density based on the output of the Gibbs sampler when fitting a mixture of $k = 2$ bivariate normal distributions to the data using the priors described in Section 2.4.1. The Gibbs sampler was started from a point drawn at random from the prior, and was run for $20\,000$ iterations (taking 193 seconds), the first $10\,000$ iterations being discarded as burn-in. The sampler moved quickly from the random starting point to an area of higher likelihood, and appeared to mix well, with no evidence of problems due to trapping states.

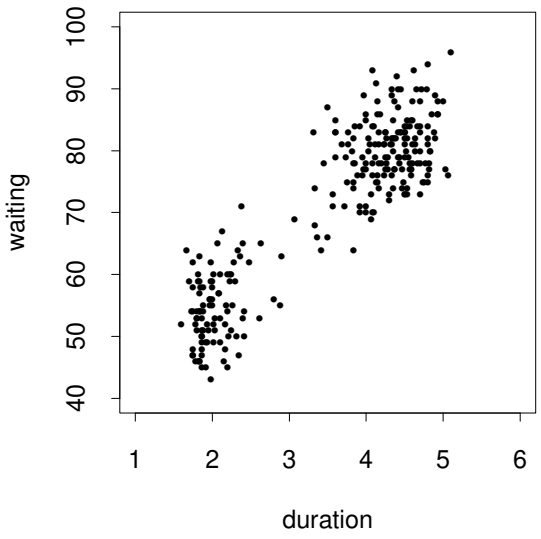


Figure 2.14: Scatter plot of the Old Faithful data (from Härdle, 1991). The x axis shows the duration (in minutes) of the eruption, and the y axis shows the waiting time (in minutes) before the next eruption.

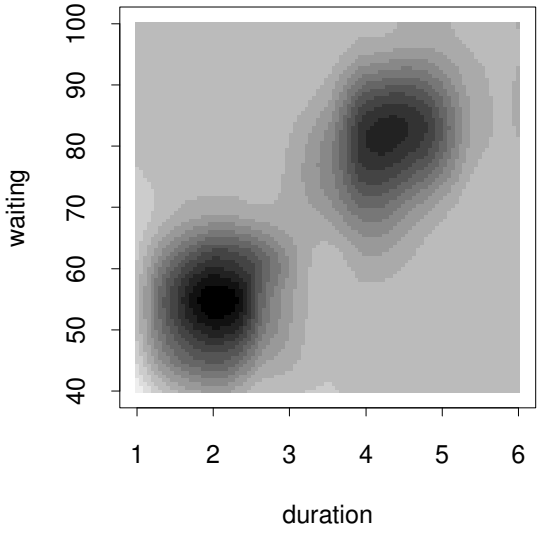


Figure 2.15: Estimated predictive density (2.21) for Old Faithful data, obtained by fitting a mixture of $k = 2$ multivariate normal distributions to the data shown in Figure 2.14, using the priors given in Section 2.4.1 and the Gibbs sampler steps given in Section 2.4.2.

2.4.4 Example: Correlations in the duration for the Old Faithful data

Another interesting two-dimensional view of the Old Faithful data can be obtained by considering the relationship between successive values of the duration, which are not independent as can be seen by examining a scatter plot of the duration of the tth eruption against the duration of the $t+1$-th eruption (Figure 2.16). Ignoring the time-series structure of the data (and so using the wrong likelihood) we tried fitting a mixture of $k = 4$ bivariate normal distributions to this data, using the priors described in section 2.4.1. The Gibbs sampler was started from a point drawn at random from the prior, and was run for $20\,000$ iterations (taking 288 seconds), the first $10\,000$ iterations being discarded as burn-in; Figure 2.17 shows the resulting predictive density estimate. The sampler moved quickly from the random starting point to an area of higher likelihood, and appeared to mix well, with no evidence of problems due to trapping states. Both views of the Old Faithful data are analysed under the assumption that k is *unknown* in Chapter 4.

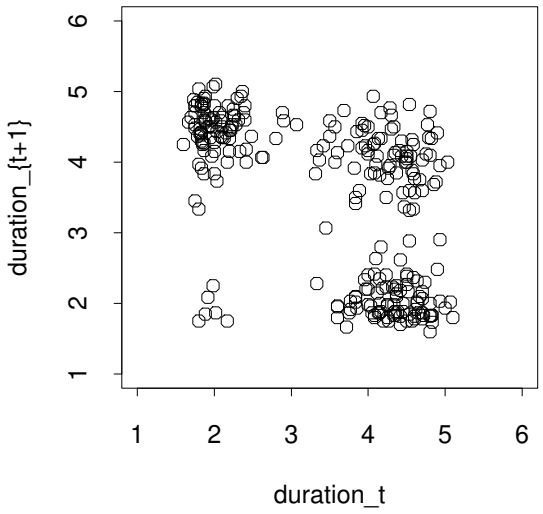


Figure 2.16: Scatter plot showing the non-independence of successive values of the duration for the Old Faithful data from Härdle (1991).

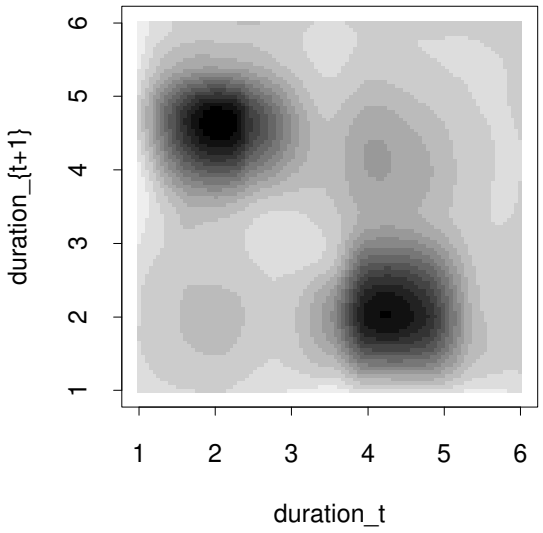


Figure 2.17: Estimated predictive density (2.21) based on fitting a mixture of $k = 4$ bivariate normal distributions to the data shown in Figure 2.16, using the priors given in Section 2.4.1 and the Gibbs sampler steps given in Section 2.4.2.

Chapter 3

Label-switching and Bayesian clustering

In Chapter 1 we saw how the symmetry of the posterior distribution of the mixture model parameters causes problems when attempting to apply Bayesian methods to mixture models in the clustering context, and in particular how the symmetry causes the scaled predictive component densities, and the classification probabilities, to be the same for all components of the mixture (see Section 1.1.5, page 13).

In our analysis of the galaxy data in the previous chapter (Section 2.2) we saw how the symmetry of the posterior distribution can cause label-switching to occur in the Gibbs sampler as the sampler moves between symmetric sets of modes, and how this can hinder interpretation of the individual components being fitted to the data. We also saw how in some cases (in particular in fitting $k = 3$ normal distributions to the galaxy data, Section 2.2) the posterior distribution possesses symmetric sets of modes which are separated by regions of such low probability that label-switching is very unlikely to occur in the Gibbs sampler, and how in this case the output of the sampler appears to provide sensible estimates of the scaled predictive component densities (Figure 2.8). This suggests that we might seek a method of "undoing" the label-switching where it occurs, and then base inference on the output of the Gibbs sampler with the label-switching "undone".

We note that label-switching may also be a problem in applications for which interpretation of the individual components of the mixture is of no interest, as label-switching movements between modes due to symmetry can hide movement between "genuinely" different modes. It is important to see how the sampler moves between "genuinely" different modes in order to assess its mixing properties. Methods of "undoing" the label-switching may then be of interest even in applications where the components themselves have no physical interpretation.

In this chapter we examine previous approaches with dealing with this problem, and explain why they are inadequate. We then describe alternative methods of dealing with the problem, and demonstrate them on some univariate and bi-

variate data, including the galaxy data considered in the previous chapter. The methods are applied to more problems in Chapter 4.

Initially we consider the problem in the context of a mixture of (possibly multivariate) normal distributions, parameterised by $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, with density

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \pi_1 \mathcal{N}(x; \mu_1, \Sigma_1) + \cdots + \pi_k \mathcal{N}(x; \mu_k, \Sigma_k) \tag{3.1}$$

which is invariant under *permutations of* θ defined by

$$\nu(\theta) = \big((\pi_{\nu(1)}, \ldots, \pi_{\nu(k)}), (\mu_{\nu(1)}, \ldots, \mu_{\nu(k)}), (\Sigma_{\nu(1)}, \ldots, \Sigma_{\nu(k)})\big)$$

where ν is any permutation of $1, \ldots, k$. We consider the problem in a more general context in Section 3.3.

3.1 Previous approaches

Previous attempts to deal with the problem of label-switching have centred around the idea of imposing *identifiability constraints* on the parameter space, such as $\pi_1 < \cdots < \pi_k$ or (when considering mixtures of univariate distributions) $\mu_1 < \cdots < \mu_k$, which can be satisfied by only one permutation of θ for each θ. Imposing an identifiability constraint on the parameters breaks the symmetry of the posterior distribution of the parameters, and so we might hope that it would allow us to perform sensible inference for the individual components of the mixture. Inference conditional on such a constraint may be performed by *post-processing* the sample $(\theta^{(1)}, \ldots, \theta^{(N)})$ produced by the Gibbs sampler, by applying permutations $\nu_1, \ldots, \nu_N$ such that the constraint is satisfied by the *permuted* or *relabelled* sample points $\nu_1(\theta^{(1)}), \ldots, \nu_N(\theta^{(N)})$. This is made formal by the following Proposition and Corollary.

Proposition 3.1. *Consider identifiability constraints $\theta \in A$, where A is a set such that for all θ, $\nu(\theta) \in A$ for exactly one permutation $\nu = \nu_\theta$ say. Let g be the function defined by $g(\theta) = \nu_\theta(\theta)$. Then*

$$E\big(F(g(\Theta)) \,|\, x^n\big) = E\big(F(\Theta) \,|\, x^n, \Theta \in A\big).$$

Proof.

$$\begin{aligned}
E\big(F(g(\Theta)) \,|\, x^n\big) &= \int F(g(\theta)) p(\theta \mid x^n)\, d\theta \\
&= \int F(g(\theta)) p(g(\theta) \mid x^n)\, d\theta \\
&\qquad\qquad [p(\theta \mid x^n) = p(g(\theta) \mid x^n) \text{ by symmetry of } p] \\
&= k! \int_A F(\theta) p(\theta \mid x^n)\, d\theta \quad \text{[symmetry]} \\
&= k! \int F(\theta) p(\theta \mid x^n) I(\theta \in A)\, d\theta \\
&= E\big(F(\Theta) \,|\, x^n, \Theta \in A\big).
\end{aligned}$$

□

Corollary 3.2. *Let random variables $\Theta^{(1)}, \Theta^{(2)}, \ldots$ be such that for any function of the parameters, F,*

$$\frac{1}{N}\sum_{t=1}^{N} F(\Theta^{(t)}) \to E\big(F(\Theta)\,|\,x^n\big) \text{ almost surely as } N \to \infty \tag{3.2}$$

as will be the case for the Markov chain constructed using the Gibbs sampler, for example. Then for A and g as in Proposition 3.1

$$\frac{1}{N}\sum_{t=1}^{N} F(g(\Theta^{(t)})) \to E\big(F(\Theta)\,|\,x^n, \Theta \in A\big) \text{ almost surely as } N \to \infty \tag{3.3}$$

and so the sample path average of the permuted sample is an appropriate estimate of $E\big(F(\Theta)\,|\,x^n, \Theta \in A\big)$.

Proof.

$$\begin{aligned}\frac{1}{N}\sum_{t=1}^{N} F(g(\Theta^{(t)})) &\to E\big(F(g(\Theta))\,|\,x^n\big) \quad \text{[by equation (3.2)]} \\ &= E\big(F(\Theta)\,|\,x^n, \Theta \in A\big) \quad \text{[by Proposition 3.1.]} \end{aligned} \tag{3.4}$$

□

3.1.1 Inference with the permuted sample

Proposition 3.1 and Corollary 3.2 show that inference conditional on an identifiability constraint $\theta \in A$ may be performed by finding permutations $\nu_1, \ldots, \nu_N$ such that the constraint is satisfied by the *permuted sample* points $\nu_1(\theta^{(1)}), \ldots, \nu_N(\theta^{(N)})$. Quantities of interest may then be estimated by their average over the permuted sample. For example, the scaled component densities given $\theta \in A$ may be estimated by:

$$\begin{aligned}\int \pi_i \mathcal{N}(x;\mu_i,\Sigma_i) p(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,x^n,\theta\in A)\,d\boldsymbol{\pi}\,d\boldsymbol{\mu}\,d\boldsymbol{\Sigma} \approx \\ \frac{1}{N-m}\sum_{t=m+1}^{N} \pi^{(t)}_{\nu_t(i)} \mathcal{N}(x;\mu^{(t)}_{\nu_t(i)},\Sigma^{(t)}_{\nu_t(i)}).\end{aligned} \tag{3.5}$$

Similarly we can obtain estimates of the classification probabilities of the observed data (see equation (1.10)):

$$\begin{aligned}\Pr(Z_j = i\,|\,x^n,\theta\in A) &= \int \frac{\pi_i \mathcal{N}(x_j;\mu_i,\Sigma_i)}{\sum_l \pi_l \mathcal{N}(x_j;\mu_l,\Sigma_l)} p(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,x^n,\theta\in A)\,d\boldsymbol{\pi}\,d\boldsymbol{\mu}\,d\boldsymbol{\Sigma} \\ &\approx \frac{1}{N-m}\sum_{t=m+1}^{N} \frac{\pi^{(t)}_{\nu_t(i)} \mathcal{N}\big(x_j;\mu^{(t)}_{\nu_t(i)},\Sigma^{(t)}_{\nu_t(i)}\big)}{\sum_l \pi^{(t)}_l \mathcal{N}(x_j;\mu^{(t)}_l,\Sigma^{(t)}_l)}\end{aligned} \tag{3.6}$$

and the predictive classification probabilities for a future data point x_{n+1} (see equation (1.11)):

$$\Pr(Z_{n+1} = i \mid x^{n+1}, \theta \in A) \propto \int \pi_i \mathcal{N}(x_{n+1}; \mu_i, \Sigma_i) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid x^n, \theta \in A)\, d\boldsymbol{\pi}\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma} \tag{3.7}$$

which is approximated by (3.5).

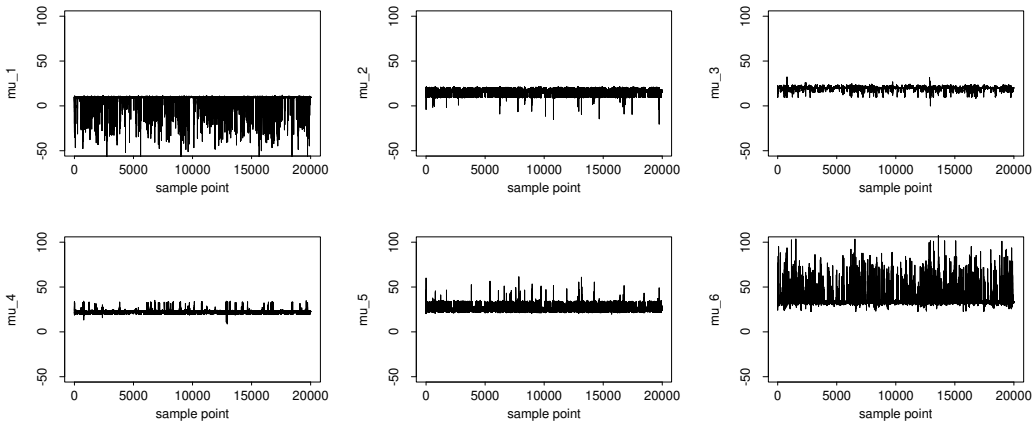
3.1.2 Problems with this approach

Unfortunately, insufficient care in the choice of suitable identifiability constraints can lead to much of the symmetry in the posterior distribution of the parameters remaining, and as a result inference for the individual components of the mixture may continue to make little sense. For example, consider the output of the Gibbs sampler used in the previous chapter to fit a mixture of $k = 6$ normal distributions to the galaxy data (Section 2.2), in which label-switching occurred frequently (Figure 2.6, page 28). Suppose we choose to impose the constraint $\mu_1 < \cdots < \mu_k$ on the parameter space, and so post-process the output of the Gibbs sampler by applying permutations so that this constraint is satisfied. The graphs of the means of the permuted sample (Figure 3.1a) show that the relatively isolated components centred near 10 and 34 (top left and bottom right in the figure) have suffered slightly from the addition of some "noise", although estimates of the scaled predictive component densities (Figure 3.1b) suggest that this noise carries little weight. There appears to be some evidence of continued label-switching behaviour in the graphs of the means of the other 4 components, presumably due to symmetries in the posterior distribution which remain despite imposing the constraint. This is reflected in the corresponding scaled predictive component density estimates, which continue to exhibit either multimodality or skewness (Figure 3.1b) making them difficult to interpret physically in the context of fitting normal distributions to the data.

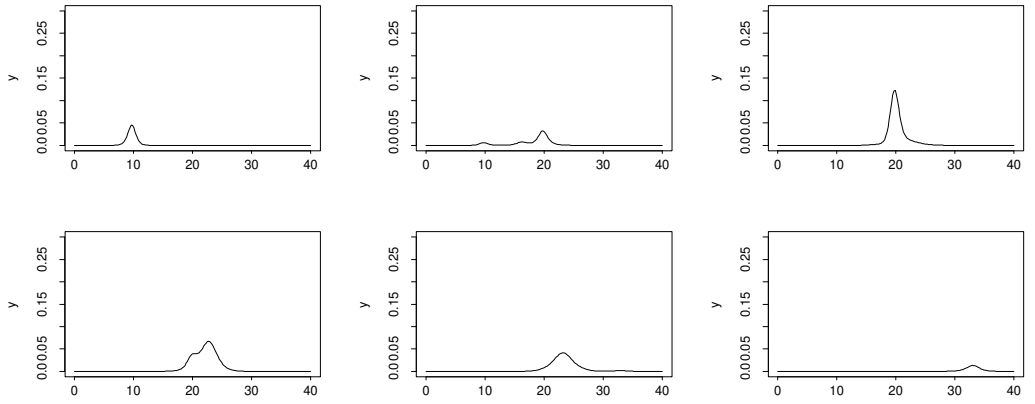
Despite the problems with this approach, few authors have tried anything more sophisticated. Richardson and Green (1997) suggest labellings be chosen to order on the means, variances, weights or "some combination of all three", without being more specific on what that might mean. Celeux *et al.* (1995) suggest various relabelling schemes, but they all rely on the "true" values of the parameters being known, which makes it difficult to imagine circumstances in which they might be useful.

3.2 A possible solution

Our initial approach is to "undo" the label-switching by seeking permutations $\nu_1, \ldots, \nu_n$ such that the relabelled sample points $\nu_1(\theta^{(1)}), \ldots, \nu_N(\theta^{(N)})$ are all labelled in the "same way" in that they agree well on the (ordered) scaled component



(a) Sampled values of the means for the permuted sample.



(b) Scaled predictive component density estimates based on permuted sample.

Figure 3.1: Graphs showing the result of imposing the identifiability constraint $\mu_1 < \cdots < \mu_k$ when fitting a mixture of $k = 6$ normal distributions to the galaxy data. The output $\theta^{(1)}, \ldots, \theta^{(N)}$ of the Gibbs sampler used to fit this model (Section 2.2) was post-processed by applying permutations $\nu_1, \ldots, \nu_N$ so that the permuted sample $\nu_1(\theta^{(1)}), \ldots, \nu_N(\theta^{(N)})$ satisfied the constraint. **Top:** The means of the six components, from the permuted sample; **Bottom:** Estimates of the scaled predictive component densities based on the permuted sample with the first $10\,000$ sampled values having been discarded as burn-in (equation (3.5)).

densities:

$$\big(\pi_1 \mathcal{N}(\cdot; \mu_1, \Sigma_1), \ldots, \pi_k \mathcal{N}(\cdot; \mu_k, \Sigma_k)\big). \tag{3.8}$$

We consider two sample points $\theta^{(1)}$ and $\theta^{(2)}$ to be labelled in the "same way" if

$$D\big[\theta^{(1)} \,\|\, \theta^{(2)}\big] = \sum_{i=1}^{k} \Delta\big[\pi_i^{(1)} \mathcal{N}(\cdot; \mu_i^{(1)}, \Sigma_i^{(1)}) \,\|\, \pi_i^{(2)} \mathcal{N}(\cdot; \mu_i^{(2)}, \Sigma_i^{(2)})\big]$$

is small, where $\Delta[\cdot \,\|\, \cdot]$ is a suitably chosen measure of divergence from one scaled density to another. We attempt to ensure that the relabelled sample points all agree well with each other by seeking permutations $\nu_1, \ldots, \nu_N$ and a parameter value $\widehat{\theta}$ to minimise

$$\mathcal{D} = \sum_{t=1}^{N} D\big[\nu_t(\theta^{(t)}) \,\|\, \widehat{\theta}\big]. \tag{3.9}$$

For suitable choice of $\Delta[\cdot \,\|\, \cdot]$ (an example of which we give in Section 3.2.2) the following gives a computationally tractable algorithm for attempting to minimise $\mathcal{D}$.

Algorithm 3.1. Starting with some initial values for $\nu_1, \ldots, \nu_N$ (setting them all to the identity permutation for example), iterate the following steps until a fixed point is reached:

Step 1: Choose $\widehat{\theta}$ to minimise

$$\sum_{t=1}^{N} D\big[\nu_t(\theta^{(t)}) \,\|\, \widehat{\theta}\big].$$

Step 2: For $t = 1, \ldots, N$ choose ν_t to minimise

$$D\big[\nu_t(\theta^{(t)}) \,\|\, \widehat{\theta}\big] = \sum_{i=1}^{k} \Delta\big[\pi_{\nu_t(i)}^{(t)} \mathcal{N}(\cdot; \mu_{\nu_t(i)}^{(t)}, \Sigma_{\nu_t(i)}^{(t)}) \,\|\, \widehat{\pi}_i \mathcal{N}(\cdot; \widehat{\mu}_i, \widehat{\Sigma}_i)\big].$$

3.2.1 Notes on Algorithm 3.1

Algorithm 3.1 can be viewed as a k-means type clustering algorithm (Lloyd, 1957), with $k!$ clusters corresponding to the $k!$ different ways of labelling the components. The idea is to cluster together those sample points ($t = 1, \ldots, N$) which have been labelled in the "same way", in that they give similar scaled density estimates for the scaled components. Step 1 estimates the $k!$ symmetric cluster centres, and Step 2 allocates each sample point to the cluster with the nearest centre.

The algorithm is guaranteed to reach a fixed point, as each step decreases $\mathcal{D}$ and there are only finitely many possible values for the permutations $\nu_1, \ldots, \nu_N$. As with all algorithms which are monotonic in the fit criterion, the solution reached may depend on the starting point, and there is no guarantee that the algorithm will converge to the global optimal solution. It is suggested that the algorithm be run from several starting points. If different runs give different optimal permutation and parameter values then they will provide competing views of the individual components; the optimal value of the fit criterion for each run gives an informal impression of the relative validity of these competing views.

The calculations required for Step 1 depend on the choice of $\Delta[\cdot \,\|\, \cdot]$. Step 2 can be seen as N minimisation problems of the form

$$\text{Choose } \nu \text{ to minimise } \sum_{i=1}^{k} c(i, \nu(i)) \tag{3.10}$$

where $c(i,l)$ is of the form

$$c(i,l) = \Delta\left[\pi_l^{(1)} \mathcal{N}(\cdot; \mu_l^{(1)}, \Sigma_l^{(1)}) \,\|\, \pi_i^{(2)} \mathcal{N}(\cdot; \mu_i^{(2)}, \Sigma_i^{(2)})\right].$$

Problem (3.10) is equivalent to the integer programming problem:

$$\begin{aligned} &\text{Choose } \{y_{il}\} (i = 1, \ldots, k; l = 1, \ldots, k) \text{ to minimise } \sum_{i=1}^{k} \sum_{l=1}^{k} y_{il}\, c(i,l) \\ &\qquad \text{subject to } y_{il} = 0 \text{ or } 1 \text{ and } \sum_{i=1}^{k} y_{il} = \sum_{l=1}^{k} y_{il} = 1 \end{aligned} \tag{3.11}$$

with the correspondence between the problems being:

> If $\{\widehat{y}_{il}\}$ is an optimal solution to problem (3.11) then the corresponding optimal solution to problem (3.10) is $\widehat{\nu}(i) = l$ if and only if $\widehat{y}_{il} = 1$.

Problem (3.11) is a special version of the Transportation problem, known as the *assignment problem*, for which efficient algorithms exist (see for example Taha, 1989, page 195). We used a NAG Fortran routine to solve this problem.

3.2.2 A possible choice of $\Delta[\cdot \,\|\, \cdot]$

We motivate our definition of $\Delta[\cdot \,\|\, \cdot]$ by recalling the definition of the Kullback–Leibler divergence from a density f to a density g:

$$KL\big[f(\cdot) \,\|\, g(\cdot)\big] = \int f(x) \log \frac{f(x)}{g(x)}\, dx$$

and the Kullback–Leibler divergence from probabilities $(p, 1-p)$ to $(q, 1-q)$

$$KL[p \,\|\, q] = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

We then define the divergence from a weighted density function $pf(\cdot)$ to a weighted density function $qg(\cdot)$ by

$$\Delta\big[pf(\cdot) \,\|\, qg(\cdot)\big] = KL[p \,\|\, q] + pKL\big[f(\cdot) \,\|\, g(\cdot)\big] \tag{3.12}$$

which gives

$$\begin{aligned}\Delta\big[\pi_i^{(1)}\mathcal{N}(\cdot;\mu_i^{(1)},\Sigma_i^{(1)}) \,\|\, \pi_i^{(2)}\mathcal{N}(\cdot;\mu_i^{(2)},\Sigma_i^{(2)})\big] = \\ \pi_i^{(1)} \log \frac{\pi_i^{(1)}}{\pi_i^{(2)}} + (1-\pi_i^{(1)}) \log \frac{1-\pi_i^{(1)}}{1-\pi_i^{(2)}} \\ + \pi_i^{(1)} \int \mathcal{N}(x;\mu_i^{(1)},\Sigma_i^{(1)}) \log \frac{\mathcal{N}(x;\mu_i^{(1)},\Sigma_i^{(1)})}{\mathcal{N}(x;\mu_i^{(2)},\Sigma_i^{(2)})}\, dx.\end{aligned} \tag{3.13}$$

With this choice of $\Delta[\cdot \,\|\, \cdot]$ Algorithm 3.1 translates into Algorithm 3.2 below. A proof of this is given in the appendix to this chapter (Proposition 3.3, Section 3.7).

Algorithm 3.2. Starting with some initial values for $\nu_1, \ldots, \nu_N$ (setting them all to the identity permutation for example), iterate the following steps until a fixed point is reached:

Step 1: Let $\widehat{\theta}$ be given by:

$$\widehat{\pi}_i = \frac{1}{N} \sum_t \pi_{\nu_t(i)}^{(t)} \tag{3.14}$$

$$\widehat{\mu}_i = \sum_t \pi_{\nu_t(i)}^{(t)} \mu_{\nu_t(i)}^{(t)} \Big/ \sum_t \pi_{\nu_t(i)}^{(t)} \tag{3.15}$$

$$\widehat{\Sigma}_i = \sum_t \pi_{\nu_t(i)}^{(t)} \big(\Sigma_{\nu_t(i)}^{(t)} + (\mu_{\nu_t(i)}^{(t)} - \widehat{\mu}_i)(\mu_{\nu_t(i)}^{(t)} - \widehat{\mu}_i)^T\big) \Big/ \sum_t \pi_{\nu_t(i)}^{(t)} \tag{3.16}$$

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

Step 2: For $t = 1, \ldots, N$ choose ν_t to minimise

$$\begin{aligned}\sum_{i=1}^{k} \Big\{ \pi_{\nu_t(i)}^{(t)} \tfrac{1}{2} \log |\widehat{\Sigma}_i| + \pi_{\nu_t(i)}^{(t)} \tfrac{1}{2} \mathrm{tr}\big[\widehat{\Sigma}_i^{-1} \big(\Sigma_{\nu_t(i)}^{(t)} + (\mu_{\nu_t(i)}^{(t)} - \widehat{\mu}_i)(\mu_{\nu_t(i)}^{(t)} - \widehat{\mu}_i)^T\big)\big] \\ - \pi_{\nu_t(i)}^{(t)} \log \widehat{\pi}_i - \big(1 - \pi_{\nu_t(i)}^{(t)}\big) \log\big(1 - \widehat{\pi}_i\big) \Big\}.\end{aligned}$$

Step 1 is computationally straightforward. Step 2 can be solved as a version of the Transportation problem, as described in Section 3.2.1, with

$$c(i,l) = \Big\{ \pi_l^{(t)} \tfrac{1}{2} \log |\widehat{\Sigma}_i| + \pi_l^{(t)} \tfrac{1}{2} \mathrm{tr}\big[\widehat{\Sigma}_i^{-1} \big(\Sigma_l^{(t)} + (\mu_l^{(t)} - \widehat{\mu}_i)(\mu_l^{(t)} - \widehat{\mu}_i)^T\big)\big] \\ - \pi_l^{(t)} \log \widehat{\pi}_i - \big(1 - \pi_l^{(t)}\big) \log\big(1 - \widehat{\pi}_i\big) \Big\}.$$

We demonstrate the use of Algorithm 3.2 on some examples in Section 3.4.

3.3 An alternative more generally applicable solution

Consider now the more general mixture model, with parameters $\theta = (\boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ and density

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \pi_1 f(x; \phi_1, \eta) + \cdots + \pi_k f(x; \phi_k, \eta). \tag{3.17}$$

The likelihood for θ is once again the same for all permutations of θ

$$\nu(\theta) = \big((\pi_{\nu(1)}, \ldots, \pi_{\nu(k)}), (\phi_{\nu(1)}, \ldots, \phi_{\nu(k)}), \eta\big)$$

and the same problems with symmetry arise.

In some cases we may be able to adapt Algorithm 3.1 to this situation, attempting to "undo" the label-switching by making the permuted sample points agree on the scaled component densities

$$\big(\pi_1 f(\cdot; \phi_1, \eta), \ldots, \pi_k f(\cdot; \phi_k, \eta)\big). \tag{3.18}$$

However, in some cases the optimisations required in Steps 1 and 2 of the algorithm may be rather computationally demanding, and the resulting algorithm may be too slow to be of practical use.

In an attempt to find an algorithm which is very generally applicable we consider "undoing" the label-switching by making the permuted sample points agree on the $k \times n$ matrix of classification probabilities $P = (p_{ij})$ where

$$p_{ij} = \Pr(Z_j = i \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \frac{\pi_i f(x_j; \phi_i, \eta)}{\sum_l \pi_l f(x_j; \phi_l, \eta)}. \tag{3.19}$$

A natural measure of the divergence from one matrix of classification probabilities $P = (p_{ij})$ to another such matrix $Q = (q_{ij})$ is given by

$$D[P \parallel Q] = \sum_{j=1}^{n} \sum_{i=1}^{k} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{3.20}$$

Let $P^{(t)} = (p_{ij}^{(t)})$ be the matrix of classification probabilities corresponding to the sample point $\theta^{(t)}$, and let $\nu_l(P^{(t)})$ be the matrix of classification probabilities

corresponding to the permuted sample point $\nu_t(\theta^{(t)})$. Then the (i,j)th element of $\nu_t(P^{(t)})$ is $p^{(t)}_{\nu_t(i)j}$. In order to ensure that the permuted sample points agree on the classification probabilities we seek permutations $\nu_1, \ldots, \nu_N$ and a matrix of classification probabilities $\widehat{P} = (\widehat{p}_{ij})$ (whose columns are constrained to sum to unity) to minimise

$$\mathcal{D} = \sum_{t=1}^{N} D[\nu_t(P^{(t)}) \parallel \widehat{P}]. \tag{3.21}$$

Each step of the following algorithm reduces $\mathcal{D}$:

Algorithm 3.3. Starting with some initial values for $\nu_1, \ldots, \nu_N$ (setting them all to the identity permutation for example), iterate the following steps until a fixed point is reached:

Step 1: Choose $\widehat{P}$ to minimise

$$\sum_{t=1}^{N} D[\nu_t(P^{(t)}) \parallel \widehat{P}].$$

Step 2: For $t = 1, \ldots, N$ choose ν_t to minimise

$$D[\nu_t(P^{(t)}) \parallel \widehat{P}] = \sum_{i=1}^{k} \sum_{j=1}^{n} p^{(t)}_{\nu_t(i)j} \log \frac{p^{(t)}_{\nu_t(i)j}}{\widehat{p}_{ij}}.$$

Step 1 is solved easily by setting

$$\widehat{p}_{ij} = \frac{1}{N} \sum_{t=1}^{N} p^{(t)}_{\nu_t(i)j}.$$

Step 2 can be solved as a version of the Transportation problem, as described in Section 3.2.1, with

$$c(i,l) = \sum_{j=1}^{n} p^{(t)}_{lj} \log \frac{p^{(t)}_{lj}}{\widehat{p}_{ij}}. \tag{3.22}$$

Algorithm 3.3 has obvious parallels with Algorithm 3.1, and the remarks made in Section 3.2.1 regarding convergence and choice of starting point apply equally here. We note that the algorithm may be computationally quite demanding on memory, as there are N matrices of classification probabilities, each of which consist of nk numbers. It may not be possible to store all these in memory at one time, which will slow the algorithm down. From the point of view of searching for interpretations for the data this algorithm is also not so attractive as Algorithm 3.2 as it does not work directly with the shapes of the components. However, we will see in our examples that the fact that the algorithm assumes no particular shape for the components is very useful in some circumstances.

3.4 Examples

We now demonstrate Algorithms 3.2 and 3.3 on both univariate and bivariate data.

3.4.1 Fitting $k = 6$ normal components to the galaxy data

For our first example we consider the output of the Gibbs sampler used to fit $k = 6$ normal components to the galaxy data, described in Section 2.2. In this case, for the purposes of illustration, we applied Algorithms 3.2 and 3.3 to post-process all $20\,000$ sample points generated by the Gibbs sampler, but in general we recommend that the first m burn-in samples are discarded before the algorithms are applied, in order to avoid problems which may be caused by atypical initial sample points.

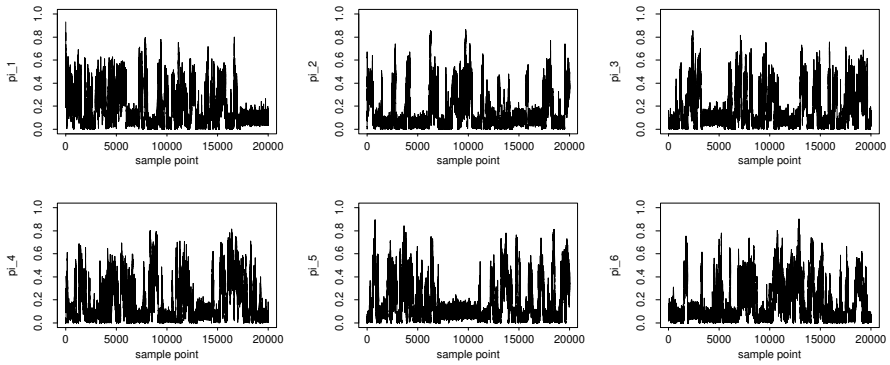
We ran each algorithm using 10 different starting points: the first chosen by initialising all permutations to the identity (corresponding to using the raw output of the Gibbs sampler) and 9 others chosen by selecting the initial permutations at random. Algorithm 3.2 gave the same optimal permutations for all 10 starting points; Algorithm 3.3 gave several different optimal solutions depending on the starting point used, but they all gave qualitatively very similar results. Algorithm 3.2 was generally much quicker than Algorithm 3.3, though it often took more iterations to converge to a fixed point. For example, for the runs which took the raw output of the Gibbs sampler as their starting point Algorithm 3.2 took 15 iterations and 377 seconds to converge, while Algorithm 3.3 took 9 iterations and 1080 seconds[1]. Further experience with the algorithms has shown that while in some cases the optimal permutations found do depend on the starting point chosen, these differences seldom make a significant difference to inference, and all future examples are based on running the algorithm with the raw output of the Gibbs sampler used as a starting point.

The label-switching in evidence as the parameter plots move between several distinct levels in the raw output of the Gibbs sampler (Figure 3.2), appears to have been eliminated from the permuted samples obtained using both algorithms (Figures 3.3 and 3.4). As a result, the estimates of the scaled predictive component densities based on the permuted samples provide a more sensible picture of the components fitted to the data than those based on the raw output of the Gibbs sampler (see Figure 3.5). Both algorithms appear to have succeeded in identifying four normal components quite clearly, and two (on the right-hand side of the figures) whose positions are less clear. We interpret the slightly bumpy nature of these components as being due to "genuine" multimodality in the posterior distribution of the parameters. The two algorithms differ slightly in the way in which they deal with this "genuine" multimodality. We note that Algorithm 3.2 also provides (through equations (3.14)–(3.16)) a natural way of finding estimates $(\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ for the parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Estimates based on the permuted sample

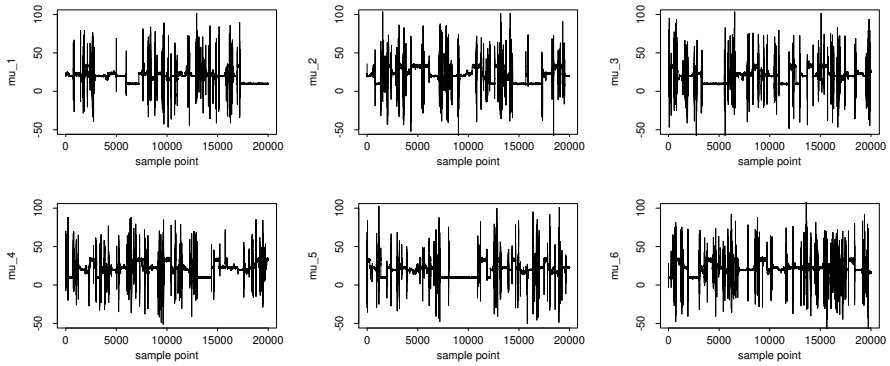
[1] CPU times on a Sun UltraSparc 200 workstation

are shown under the scaled predictive component density estimates in Figure 3.5b.

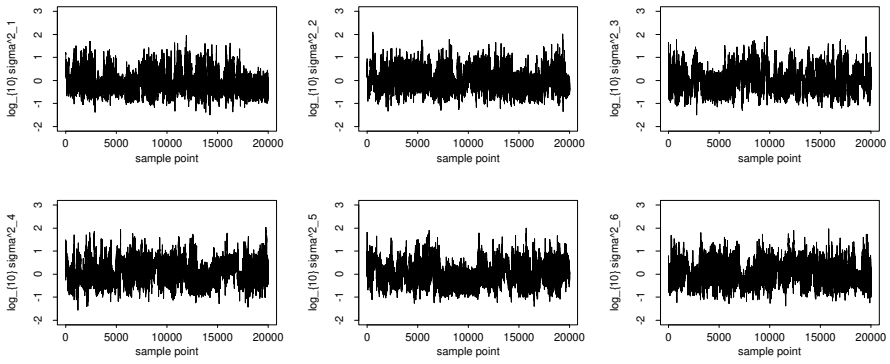
The data points can be clustered into groups by choosing the allocation variables z_j $(j = 1, \ldots, n)$ which maximise estimates of either the classification probabilities (3.6) or the predictive classification probabilities (3.7). The latter is computationally convenient, as it is equivalent to maximising the estimates of the scaled predictive component densities which we have already calculated for Figure 3.5; the resulting clusterings of the data are shown in Figure 3.6. The algorithms differ in their choice of allocation variable z_j for only one point, and in each case the clustering groups the data into only 5 distinct groups — the sixth component has no data points allocated to it.



(a) Mixture proportions

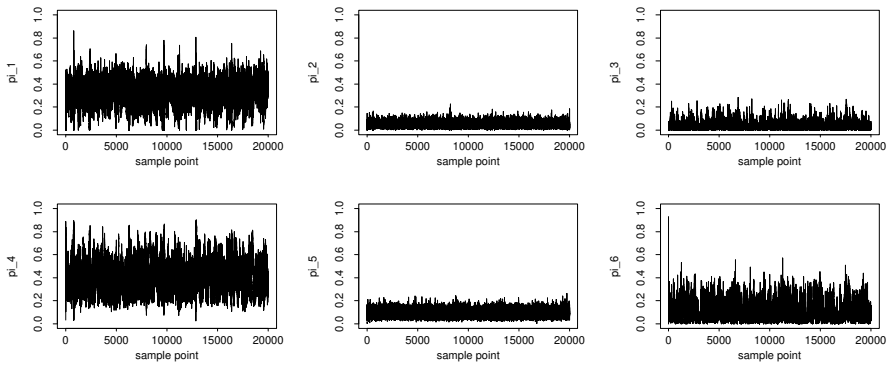


(b) Means

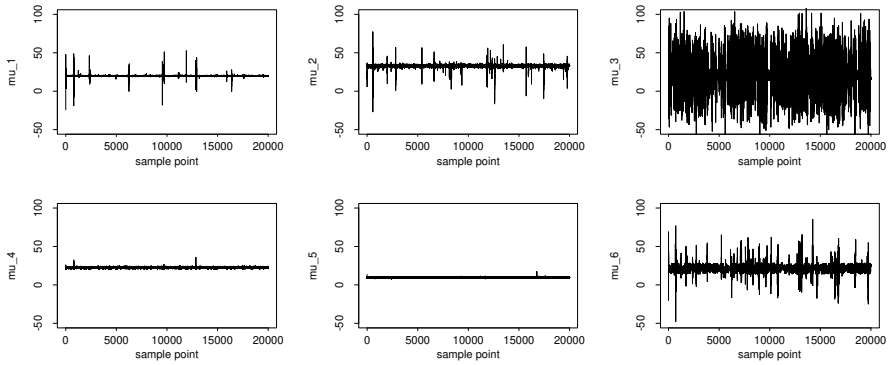


(c) $\log_{10}$ (variances)

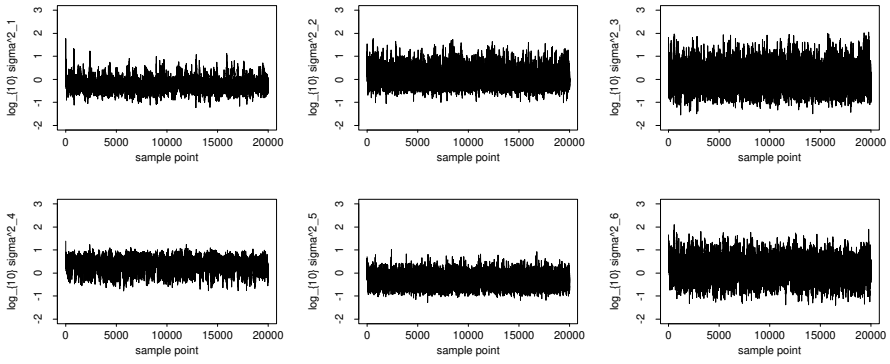
Figure 3.2: Sampled values of mixture proportions, means and $\log_{10}$(variances) when fitting $k = 6$ normal distributions to the galaxy data, using the Gibbs sampler described in Section 2.2.



(a) Mixture proportions

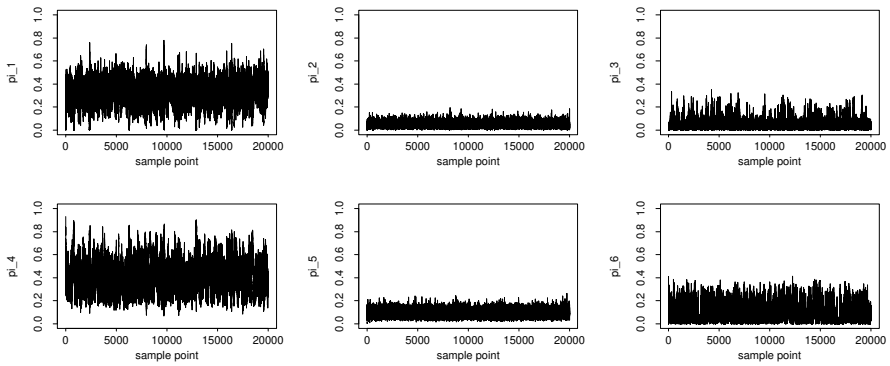


(b) Means

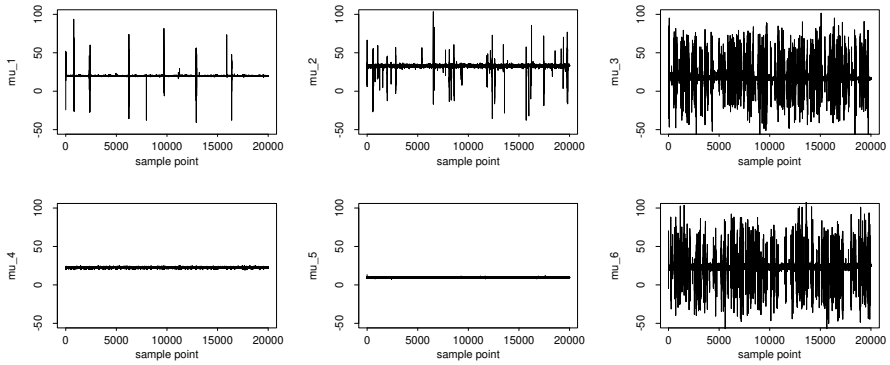


(c) $\log_{10}$ (variances)

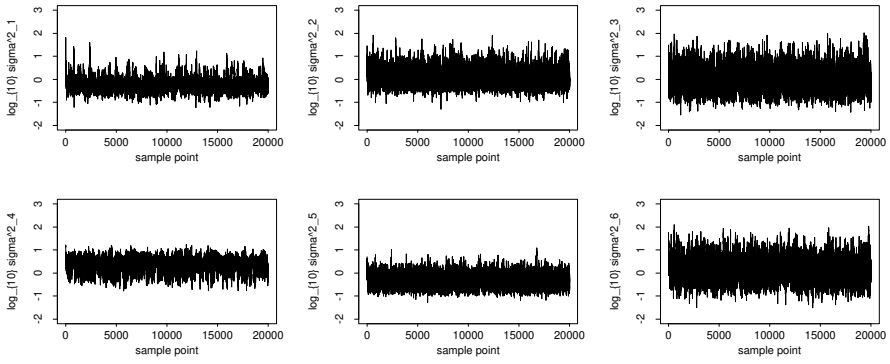
Figure 3.3: Permuted sample values of mixture proportions, means and $\log_{10}$(variances) when fitting $k = 6$ normal distributions to the galaxy data, obtained by applying Algorithm 3.2 to the sample shown in Figure 3.2. The algorithm appears to have been successful in simultaneously removing the label-switching behaviour from the mixture proportions, means, and variances.



(a) Mixture proportions

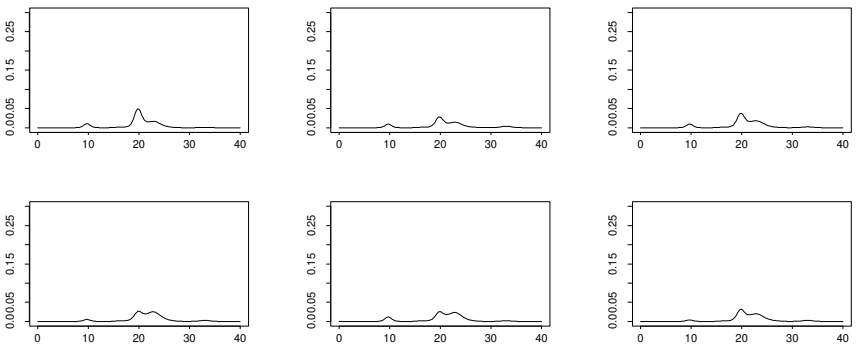


(b) Means

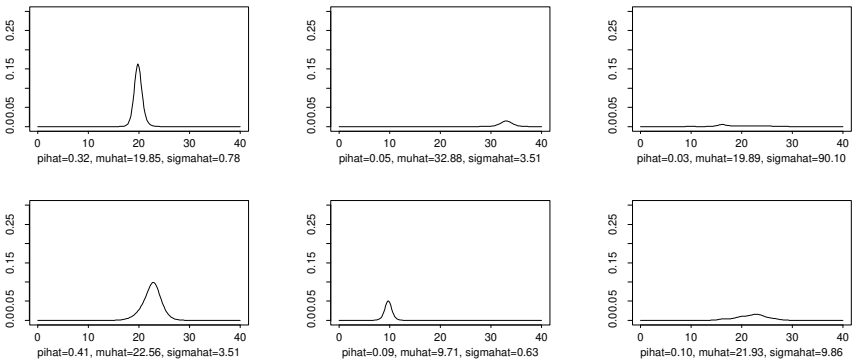


(c) $\log_{10}$ (variances)

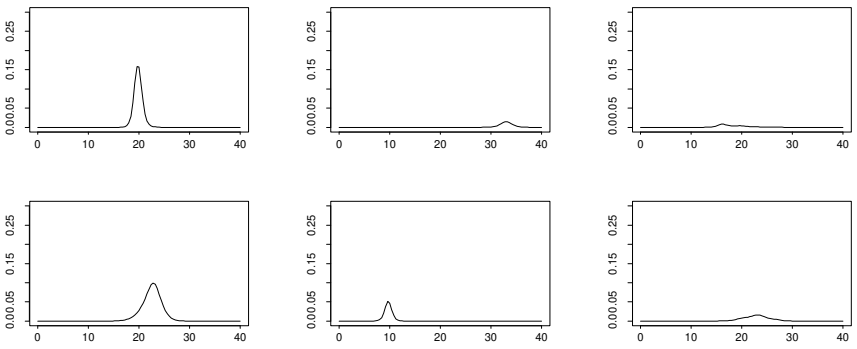
Figure 3.4: Permuted sample values of mixture proportions, means and $\log_{10}$(variances) when fitting $k = 6$ normal distributions to the galaxy data, obtained by applying Algorithm 3.3 to the sample shown in Figure 3.2. The algorithm appears to have been successful in simultaneously removing the label-switching behaviour from the mixture proportions, means, and variances.



(a) Based on raw output of Gibbs sampler (this raw output is shown in Figure 3.2).

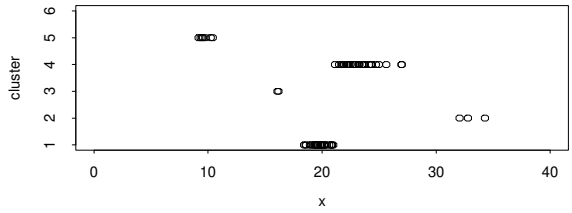


(b) Based on the permuted sample obtained by applying Algorithm 3.2 to the output of the Gibbs sampler (this permuted sample is shown in Figure 3.3).

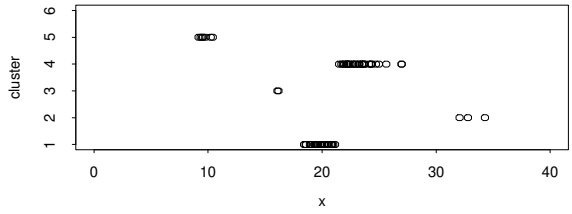


(c) Based on the permuted sample obtained by applying Algorithm 3.3 to the output of the Gibbs sampler (this permuted sample is shown in Figure 3.4).

Figure 3.5: Estimates of the scaled predictive component densities (3.5) when fitting $k = 6$ normal distributions to the galaxy data. In each case the estimates were formed after having discarded the first $m = 10\,000$ sample points as burn-in.



(a)



(b)

Figure 3.6: Clusterings of galaxy data, based on fitting $k = 6$ normal components. The clusterings were obtained by maximising the estimated predictive classification probabilities (3.7), based on the permuted samples obtained from Algorithms 3.2 and 3.3, which is equivalent to maximising the corresponding scaled predictive component density estimates shown in Figure 3.5b) and c). Note that the clusterings differ in their allocation of only one data point ($x = 21.137$), and that in each case the clustering splits the data into only five distinct groups, as no data points are allocated to the sixth component.

3.4.2 Fitting $k = 3$ t_4 components to the galaxy data.

Although Algorithm 3.2 uses distances defined between scaled *normal* component densities, we hoped that it could be successfully applied in the context of using a Gibbs sampler to fit a mixture of t distributions to data, by applying the algorithm to the sampled values of the parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ treating them exactly as if they were the results of a Gibbs sampler for fitting a mixture of normal distributions. Having found the optimal permutations, inference is then performed by reverting to treating the parameters as the parameters of t distributions. For example, estimates of the scaled predictive component densities are made by replacing the normal densities in equation (3.5) with t densities.

We used both Algorithms 3.2 and 3.3 to post-process the results from the Gibbs sampler described in Section 2.3 to fit a mixture of $k = 3$ t_4 components to the galaxy data. The first $m = 10\,000$ sample points were discarded as burn-in, and the algorithms were applied to the remaining $10\,000$ sample points, using the raw output of the Gibbs sampler as a starting point.

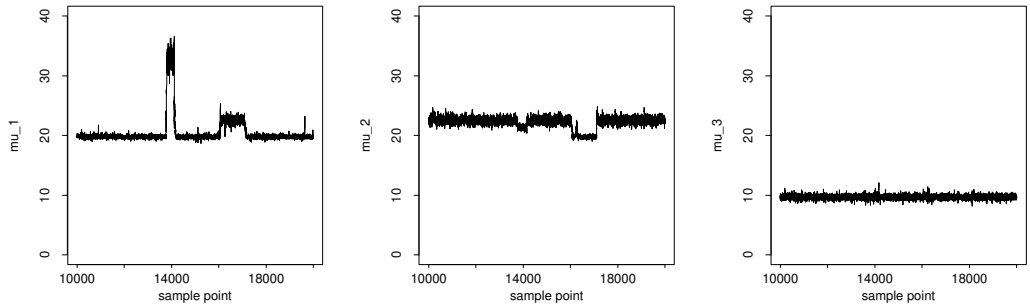
The sampled values of the means for the three components, from the raw output of the Gibbs sampler, and the permuted samples obtained from the two algorithms are shown in Figure 3.7. It seems that both algorithms have succeeded in "undoing" the label-switching which occurred between the first and second component around iterations $16\,000$ and $17\,000$ in the raw output of the Gibbs sampler. However, the algorithms differ in their treatment of the "genuine" multimodality in the posterior distribution of the means, which is exemplified by the change in the means of the first two components of the raw output of the Gibbs sampler, from means near 20 and 23 to means near 34 and 21, for about 300 iterations around iteration $14\,000$ (Figure 3.7a).

Algorithm 3.2 "shares" the sampled values of the mean near 34 between the first two components, whilst Algorithm 3.3 puts them all in the first component. We believe that this difference is not due to the particular starting point chosen; nor is it an artifact of Algorithm 3.2 having been developed in the context of mixtures of normal distributions and then applied to t distributions. Rather it is due to the fact that Algorithm 3.2 works directly with the shapes of the components, which it assumes to be of a given (unimodal) form, causing it to be less sensitive to, or less permissive of, multimodality in the scaled predictive component densities. In contrast, Algorithm 3.3 does not make any assumptions about the shapes of the components, and the permuted sample it produces gives a clearer picture of the two "genuine" modes in the distribution, effectively by "undoing" the label-switching separately for each "genuine" mode. Algorithm 3.2 would presumably behave in this way if it were applied separately to the sample points lying in the two distinct modes.

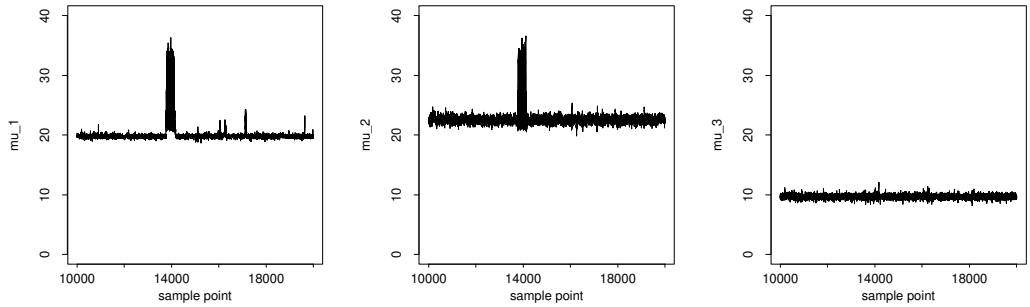
We do not believe that it makes any sense to perform clustering, or inference for the individual components of the mixture, by averaging over such different modes which clearly represent quite different views of the data. We suggest that for the purposes of clustering the two modes should be considered separately,

though it seems that such a division will be somewhat arbitrary, and there may be points which might be said to lie in either (or neither) mode. Notwithstanding this reservation, we divided the sample by eye, based on the means of the permuted sample produced by Algorithm 3.3 (Figure 3.7c), into a *minor mode* which consisted of 326 sample points ($13\,786$—$14\,111$), and a *major mode* which consisted of the remaining $9\,674$ sample points. Estimates of the scaled predictive component densities and corresponding clusterings of the points into $k = 3$ groups, based on the permuted sample obtained from Algorithm 3.3, are shown in Figure 3.8 (for the major mode) and Figure 3.9 (for the minor mode). The probabilities of these two different views can be estimated by the proportion of sample points which lie in each mode, giving 0.0326 for the minor mode, and 0.9674 for the major mode. However, the poor mixing of the sampler between the major and minor modes means that longer runs (or more runs from different starting points) would be required in order to obtain accurate estimates for these quantities. In the next chapter we will see how the mixing of the sampler can be improved by allowing the number of components k to vary (Section 4.4.1).

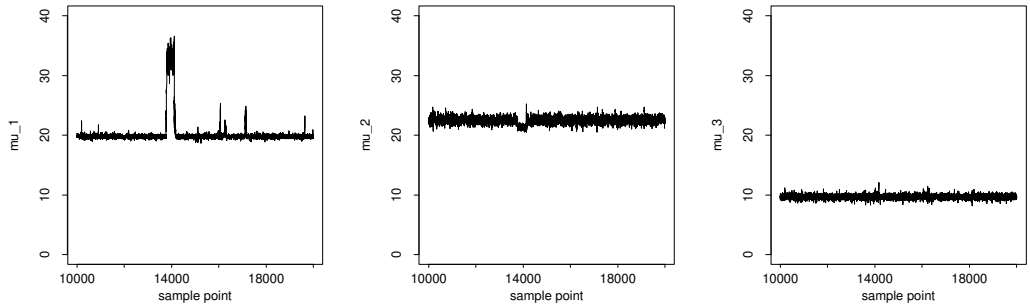
Considering the major mode and minor mode separately raises the question of the validity of the clustering inference we performed in the previous section when fitting $k = 6$ normal components to the galaxy data, since the bumpy nature of the estimates of the scaled predictive component densities (on the right-hand side of Figures 3.5b and 3.5c) indicates the presence of "genuine" multimodality in the posterior distribution of the parameters. How different do modes have to be before they should be considered separately? We currently have no answer to this question, beyond the application of intuition. A more formal method might be obtained by developing Algorithms 3.2 or 3.3 to cluster together those sample points which correspond to the same mode of the posterior distribution of the parameters and therefore give a consistent view of the data.



(a) Raw output of Gibbs sampler

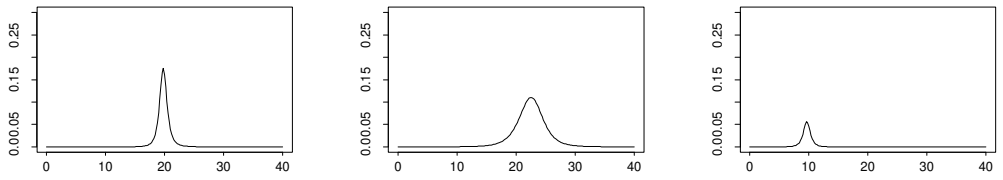


(b) Permuted sample from Algorithm 3.2

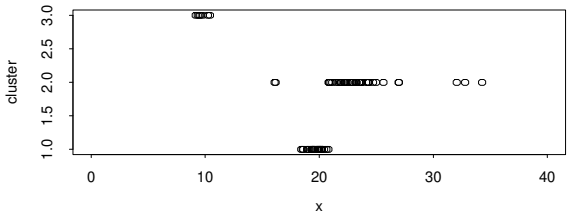


(c) Permuted sample from Algorithm 3.3

Figure 3.7: Sampled values of means of the three components when fitting $k = 3$ t_4 distributions to the galaxy data. Figures b) and c) show the means for the permuted sample obtained by applying Algorithms 3.2 and 3.3 respectively to post-process the raw output of the Gibbs sampler described in Section 2.3.

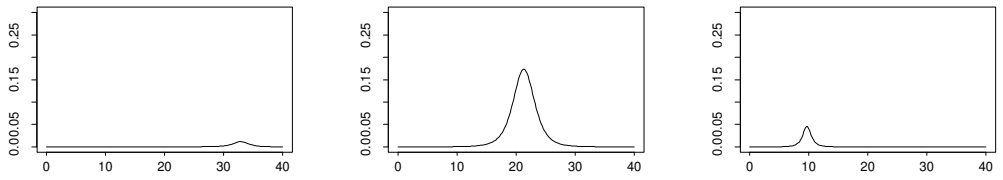


(a) Scaled predictive component densities, based on major mode.

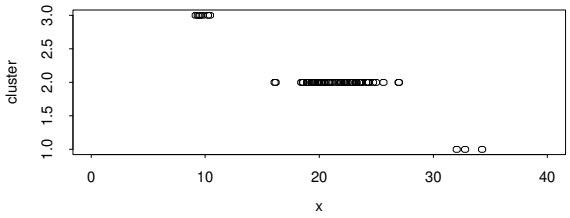


(b) Clustering of data, based on major mode.

Figure 3.8: Estimates of the scaled predictive component densities and corresponding clustering of data when fitting $k = 3$ t_4 distributions to the galaxy data, based on the major mode. The sample points corresponding to the major mode were picked out by eye from the graphs of the means of the sample permuted using Algorithm 3.3 (Figure 3.7c). These sample points were used to form estimates of the scaled predictive component densities using equation (3.5), replacing the normal densities in (3.5) with t_4 densities. The clustering was obtained by choosing allocation variables to maximise the estimated scaled predictive component densities shown in a), which is equivalent to maximising the predictive class probabilities (3.7).



(a) Scaled predictive component densities, based on minor mode.



(b) Clustering of data, based on minor mode.

Figure 3.9: Estimates of the scaled predictive component densities and corresponding clustering of data when fitting $k = 3$ t_4 distributions to the galaxy data, based on the minor mode. The sample points corresponding to the minor mode were picked out by eye from the graphs of the means of the sample permuted using Algorithm 3.3 (Figure 3.7c). These sample points were used to form estimates of the scaled predictive component densities using equation (3.5), replacing the normal densities in (3.5) with t_4 densities. The clustering was obtained by choosing allocation variables to maximise the estimated scaled predictive component densities shown in a), which is equivalent to maximising the predictive class probabilities (3.7).

3.4.3 The *Iris Virginica* data.

As an example of a clustering problem in two dimensions, we 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, giving 150 specimens in all.

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. In the next chapter (Section 4.7) we investigate in a Bayesian context whether data for the *virginica* species is suitably modelled by a single normal distribution, or whether a mixture of two or more normal components is justified. In this chapter we concentrate on using Bayesian methods to produce a clustering of the data for the *virginica* species into two groups, by fitting a mixture of $k = 2$ normal distributions to the data. In order to reduce the data to two dimensions we consider only the measurements of sepal length and petal length for the 50 examples of this species, which are given in the appendix to this thesis (Section A.3). A scatter plot of the data is given in Figure 3.10 where the observations are numbered 1 to 50 in order of their listing in Table 1.1 in Andrews and Herzberg (1985), which is the numbering used by McLachlan (1992) and also the numbering used in the data supplied with S-PLUS.

We used the Gibbs sampler, as described in Section 2.4, to fit a mixture of two bivariate normal distributions to this data, with the priors as in Section 2.4.1. The sampler was run for $20\,000$ iterations, starting from a random starting point, and appeared to mix well. The first $10\,000$ sample points were discarded as burn-in. Label-switching was present in the raw output of the Gibbs sample, and so we tried applying both Algorithm 3.2 and Algorithm 3.3 to the sample. Both algorithms appeared to be successful in "undoing" the label-switching; see for example Figure 3.11. Although the optimal permutations found by the two algorithms were different (of the $10\,000$ permutations, about 500 were different), all plots made based on the results of the two algorithms were indistinguishable by eye (results not shown).

Estimates of the scaled predictive component densities based on the raw output of the Gibbs sampler, and the permuted sample obtained from Algorithm 3.3, are shown in Figure 3.12. The corresponding estimates based on the permuted sample obtained from Algorithm 3.2 were indistinguishable by eye from the results for Algorithm 3.3. The label-switching in the raw output of the Gibbs sampler mixes together the two components to create two very similar estimates of the scaled predictive component densities, which are not suitable for performing clustering of the data. In contrast, the scaled predictive component density estimates based on the permuted sample separate the two components, and allow sensible clustering to be performed. The clustering obtained by choosing the allocation variables to maximise the estimated scaled predictive component densities obtained using Algorithm 3.3 (Figure 3.12b) puts the eight observations numbered

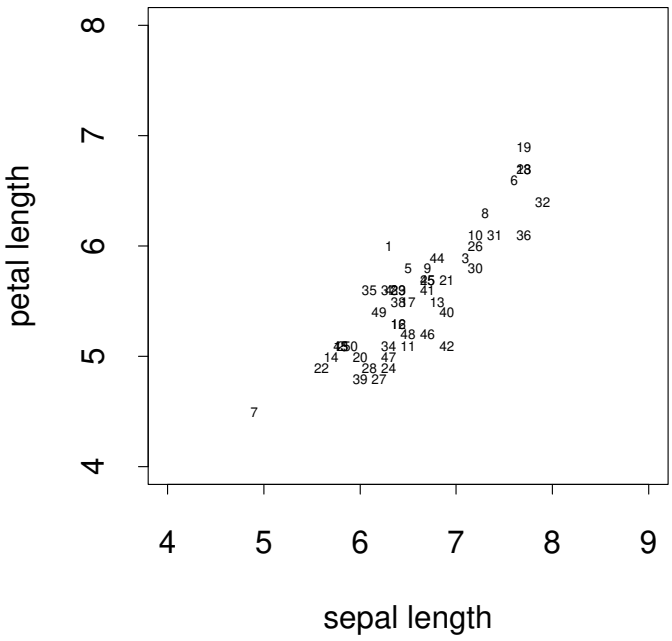


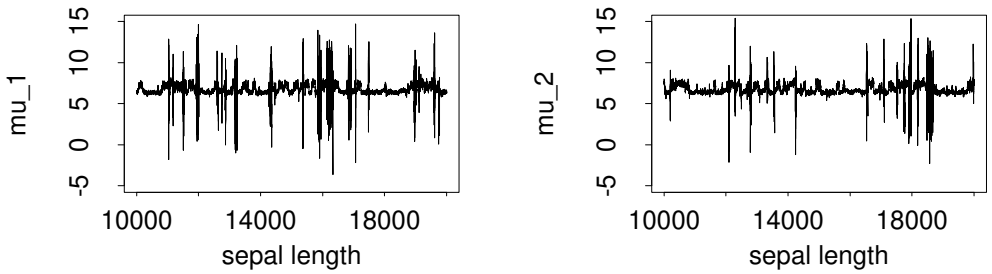
Figure 3.10: Scatter plot of petal length against sepal length for the *Iris Virginica* data. The numbering of the observations is in order of their listing in Table 1.1 in Andrews and Herzberg (1985), which is the numbering used by McLachlan (1992) and also the numbering used in the data supplied with S-PLUS. The superposed observations near the top-right of the figure are numbers 18 and 23.

6,8,18,19,23,31,32 and 36 in one cluster, and the remaining observations in the other cluster (note that observations 18 and 23 are superposed in Figure 3.10). An identical clustering is obtained for the results of Algorithm 3.2. McLachlan (1992) obtains a clustering by fitting a mixture of two normal components to the full four-dimensional data (rather than the two dimensions we consider here) using maximum likelihood methods. The smaller of his clusters contains only five observations: 6,18,19,23 and 31.

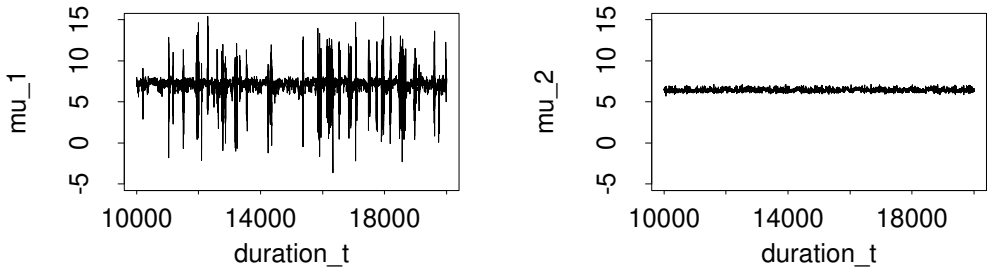
Further illustration of Algorithms 3.2 and 3.3 may be found in Chapter 4.

3.5 Discussion

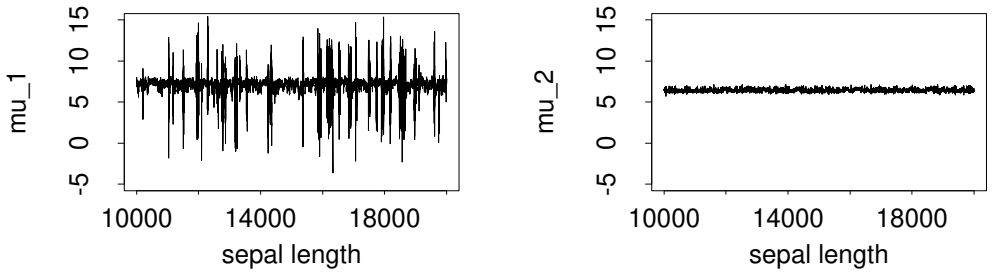
Some readers may be concerned that we are deviating somewhat from a principled Bayesian approach to the problem by post-processing the results in this way. Since relabelling the sample is equivalent to imposing an identifiability constraint on the parameter space it is effectively making a quite substantial change to the



(a) Raw output of Gibbs sampler

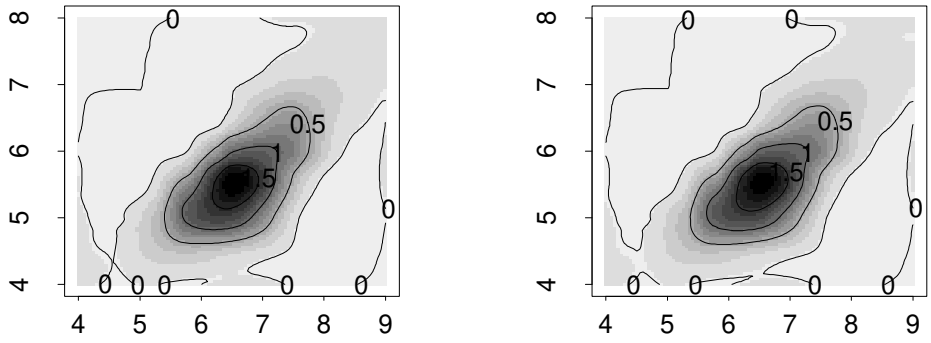


(b) Permuted sample, obtained by applying Algorithm 3.2 to the output of the Gibbs sampler.

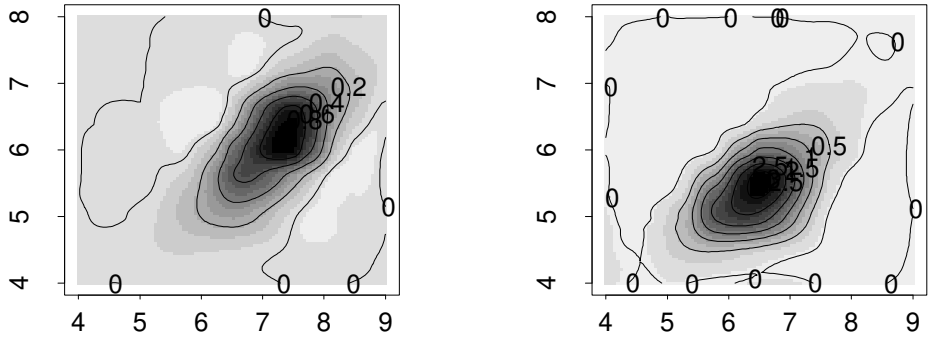


(c) Permuted sample, obtained by applying Algorithm 3.3 to the output of the Gibbs sampler.

Figure 3.11: Sampled value of the sepal length (x-coordinate of the means) for the two components, obtained when using the Gibbs sampler to fit a mixture of $k = 2$ normal distributions to the *Iris Virginica* data. In each case the first $m = 10\,000$ sample points have been discarded as burn-in. Note that the label-switching which clearly occurs between the two components in the raw output of the Gibbs sampler is successfully "undone" by both Algorithms. There are small differences in the graphs b) and c) which cannot be seen by eye on this scale.



(a) Based on raw output of Gibbs sampler.



(b) Based on the permuted sample obtained by applying Algorithm 3.3 to the output of the Gibbs sampler.

Figure 3.12: Estimates of the scaled predictive component densities when fitting 2 normal distributions to the *Iris Virginica* data, based on both the raw output of the Gibbs sampler, and the permuted sample obtained using Algorithm 3.3. Dark areas correspond to areas of higher density, but the figures are not all shaded on the same scale, and in particular the contour lines show that the component on the left of b) has smaller weight than the component on the right. The presence of label-switching in the Gibbs sampler output leads to very similar estimates of the scaled predictive density for the two components in a). Permuting the sample using Algorithm 3.3 succeeds in "undoing" the label-switching and in doing so produces density estimates for the components which are suitable for clustering inference.

prior based on examination of the data, in an apparently unprincipled way. We now present two possible views on the validity of the approach.

3.5.1 The Revisionist Bayesian view

Firstly, we can view the results of different ways of choosing the labelling as the results based on different priors. Suppose we show the results of several different labellings to an expert in the field, who picks out one of the labellings and is able to offer a physical interpretation of the results. In some cases such a physical interpretation may be expressible in terms of a simple constraint such as $\mu_1 < \cdots < \mu_k$, for example if the results prompted the expert to realise that the sample contained data from a number of different year-groups where the mean for each group might be expected to increase with age. However, one might hope (perhaps optimistically) that such a constraint would have been known and mentioned by the expert before the original analysis, and so we anticipate that our method would more often reveal interpretations which cannot be summarised by such a simple constraint. In either case, we can consider the imposition of the constraint as a refinement of the prior prompted by forcing the expert to examine his expert knowledge (in other words, his prior belief) more closely in the area of parameter space which is consistent with the data.

We suspect that some readers will be unhappy with this view, and will prefer to take the Mode-hunter view outlined below. However, there are some fairly natural questions which we feel can only be answered if we are prepared to take the Revisionist view. For example, looking at the component density estimates in Figure 3.5 it seems reasonable to ask the question "What is the probability that a given observation x arose from the component centred near 10 (bottom-middle in the Figure)." This can only be answered if we are prepared to accept that the permutations we have applied (and the components we have thus identified) have some genuine legitimacy in themselves.

3.5.2 The Mode-hunter view

Alternatively, we can view the relabelling methods as a way of searching for modes in the posterior distribution of quantities of interest. For example, Algorithm 3.2 can be see as a way of searching for modes in the posterior distribution of the scaled component densities

$$\big(\pi_1 \mathcal{N}(x; \mu_1, \Sigma_1), \ldots, \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)\big). \tag{3.23}$$

This distribution will be invariant under permutations of the component labels in the same way as the posterior distribution of the parameters, and will therefore possess up to $k!$ symmetrically equivalent sets of "genuine" modes. The posterior

mean of this distribution, which is estimated from the Gibbs sampler by

$$\Big(\frac{1}{N}\sum_{t=1}^{N}\pi_1^{(t)}\mathcal{N}\big(x\,;\mu_1^{(t)},\Sigma_1^{(t)}\big),\ldots,\frac{1}{N}\sum_{t=1}^{N}\pi_k^{(t)}\mathcal{N}\big(x\,;\mu_k^{(t)},\Sigma_k^{(t)}\big)\Big) \tag{3.24}$$

may therefore be a poor summary of the full joint distribution, as Figure 2.10 demonstrates. Relabelling the output of the Gibbs sampler can be seen as a method of ensuring that we obtain representatives of only one of the $k!$ symmetrically equivalent sets of "genuine" modes, thus ensuring that the mean is a better approximation of one of the symmetrically equivalent modes. Figure 3.5 may be viewed then as an estimate of the mean of the distribution taken over only one set of "genuine" modes. Any "genuine" multimodality may of course still cause problems when approximating the mode by a mean, as evidenced by the rather bumpy nature of the component in the bottom right corner of the Figure.

Similarly, Algorithm 3.3 may be regarded as a way of searching for modes in the posterior distribution of the classification probabilities. Thus although the Mode-hunter view does not allow us to answer questions such as "What is the probability that a given observation x arose from the component centred near 10" (or perhaps more accurately, the meaning of such a question is not clear), it *does* allow us to view classification probabilities based on the relabelled sample as an estimate of the mode of the joint distribution of the classification probabilities.

Having drawn a distinction between the two views, we remark that in either case we wish to permute the sample so that the permuted sample points $\nu_1(\theta^{(1)}),\ldots,\nu_k(\theta^{(k)})$ "agree" with each other in a suitably chosen way. The Revisionist because it makes for more interpretable results; the Mode-hunter because it gives a convenient approximation to the mode. We feel that a more model-based method, such as Algorithm 3.2, seems preferable if we wish to take the Revisionist view and seek more interpretable results. However, Algorithm 3.3 appears to provide results which are more helpful to a user attempting to separate "genuinely" different modes in the posterior distribution of the parameters; a process which we feel is important for the meaningful application of Bayesian methods to clustering using mixture models. While both algorithms require care in their application and the interpretation of their results, and neither provides a completely automatic solution to the problems of multimodality encountered in mixture models, we feel that both algorithms provide useful tools for anyone contemplating a Bayesian approach to clustering with mixture models.

3.6 A connection with approximating posterior distributions

In conclusion, we now take a brief look at the connection between the label-switching problem and the problem of analytically approximating the full posterior distribution of the parameters. This was the context in which the problems of symmetry first came to my attention, and led to the attempt at a solution to the problem of label-switching described in Stephens (1996, 1997).

Suppose we attempt to summarise the information contained in the posterior distribution of the parameters by fitting a density from some parametric family to the parameter values $\theta^{(1)}, \ldots, \theta^{(N)}$ obtained from the Gibbs sampler. The fact that the joint distribution of the parameters is invariant under permutations of the component labels suggests fitting a density of the form

$$\frac{1}{k!} \sum_{\nu \in S_k} g(\nu(\theta); a) \tag{3.25}$$

where $g(\cdot; a)$ is some family of densities parameterised by a, and S_k is the set of all permutations of $\{1, \ldots, k\}$. If we rewrite this density as

$$\frac{1}{k!} \sum_{\nu \in S_k} g_\nu(\theta; a) \tag{3.26}$$

then we see it is a mixture of $k!$ equally weighted components, in which the permutations correspond to the allocation variables (z^n in our usual formulation of the mixture model).

If we treat $\theta^{(1)}, \ldots, \theta^{(N)}$ as independent observations from the density (3.25), and we treat $\nu_1, \ldots, \nu_N$ as parameters to be estimated, then each step of the following algorithm increases the joint likelihood

$$L(\nu_1, \ldots, \nu_N; a) = \prod_{t=1}^{N} g(\nu_t(\theta^{(t)}); a). \tag{3.27}$$

Algorithm 3.4. Starting with some initial values for $\nu_1, \ldots, \nu_N$ (setting them all to the identity permutation for example), iterate the following steps until a fixed point is reached:

Step 1: Choose $\widehat{a}$ to maximise

$$\prod_{t=1}^{N} g(\nu_t(\theta^{(t)}); \widehat{a}).$$

Step 2: For $t = 1, \ldots, N$ choose ν_t to maximise

$$g(\nu_t(\theta^{(t)}); \widehat{a}).$$

This algorithm has obvious parallels with Algorithm 3.2, and is guaranteed to converge for the same reasons. The implementation of Step 1 will depend on the choice of g, and Step 2 is once again solved by treating it as a version of the transportation problem.

We note that when viewed as a method of estimating a this algorithm is the "clustering" approach to parameter estimation for mixture models, in which the allocation variables (in this case, the permutations) of the mixture are treated as

parameters to be estimated, and we search for a joint maximum likelihood estimate of the parameters and allocation variables. This approach is difficult to justify theoretically, as in general the estimates for a will be inconsistent and tend to underestimate the variance of the components (Marriott, 1975). This tendency may be exacerbated by the fact that $\theta^{(1)}, \ldots, \theta^{(N)}$ are not actually independent observations. However, it has the advantage of leading to a computationally tractable algorithm for suitable choice of g. With N observations (N possibly very large) from $k!$ components, more theoretically sound algorithms for estimating a, such as EM-type algorithms or Bayesian methods, are computationally daunting for moderate values of k (greater than 5 say).

In our initial work on the label-switching problem (Stephens, 1996, 1997) we considered an algorithm of the same form as Algorithm 3.4 with our choice of g being

$$g(\theta\,; a) = \mathcal{D}(\boldsymbol{\pi}\,; a_1) \prod_{i=1}^{k} \mathcal{N}(\mu_i; a_2, a_3)\Gamma(\sigma_i^{-2}; a_4, a_5).$$

Parts of the maximisation in Step 1 of Algorithm 3.4 must then be done numerically, leading to an algorithm which is computationally more expensive than Algorithm 3.2. For this reason, and also the fact that we believe inference for the component densities to be of more interest than inference for the parameters themselves, we have abandoned this algorithm in our current implementation. For suitable choice of g though this may provide a computationally attractive method of "undoing" label-switching in some cases, and certainly provides a useful tool when seeking an analytic approximation of the form (3.25) to the posterior distribution of the parameters.

3.7 Appendix

Proposition 3.3. *Let*

$$D\big[\theta^{(1)} \,\|\, \theta^{(2)}\big] = \sum_{i=1}^{k} \Delta\big[\pi_i^{(1)}\mathcal{N}(\cdot; \mu_i^{(1)}, \Sigma_i^{(1)}) \,\|\, \pi_i^{(2)}\mathcal{N}(\cdot; \mu_i^{(2)}, \Sigma_i^{(2)})\big] \tag{3.28}$$

where

$$\begin{aligned}\Delta\big[\pi_i^{(1)}\mathcal{N}(\cdot; \mu_i^{(1)}, \Sigma_i^{(1)}) \,\|\, \pi_i^{(2)}\mathcal{N}(\cdot; \mu_i^{(2)}, \Sigma_i^{(2)})\big] = \\ \pi_i^{(1)} \log \frac{\pi_i^{(1)}}{\pi_i^{(2)}} + (1-\pi_i^{(1)}) \log \frac{1-\pi_i^{(1)}}{1-\pi_i^{(2)}} \\ + \pi_i^{(1)} \int \mathcal{N}(x; \mu_i^{(1)}, \Sigma_i^{(1)}) \log \frac{\mathcal{N}(x; \mu_i^{(1)}, \Sigma_i^{(1)})}{\mathcal{N}(x; \mu_i^{(2)}, \Sigma_i^{(2)})}\, dx.\end{aligned} \tag{3.29}$$

Recall Algorithm 3.1 which consists of 2 steps:

Step 1: Choose $\widehat{\theta}$ to minimise

$$\sum_{t=1}^{N} D\big[\nu_t(\theta^{(t)}) \,\|\, \widehat{\theta}\big].$$

Step 2: For $t = 1, \ldots, N$ choose ν_t to minimise

$$D\big[\nu_t(\theta^{(t)}) \,\|\, \widehat{\theta}\big] = \sum_{i=1}^{k} \Delta\big[\pi^{(t)}_{\nu_t(i)} \mathcal{N}(\cdot; \mu^{(t)}_{\nu_t(i)}, \Sigma^{(t)}_{\nu_t(i)}) \,\|\, \widehat{\pi}_i \mathcal{N}(\cdot; \widehat{\mu}_i, \widehat{\Sigma}_i)\big].$$

With $\Delta[\cdot \,\|\, \cdot]$ as given by equation (3.29) these steps are equivalent to

Step 1: Let $\widehat{\theta}$ be given by:

$$\widehat{\pi}_i = \frac{1}{N} \sum_t \pi^{(t)}_{\nu_t(i)} \tag{3.30}$$

$$\widehat{\mu}_i = \sum_t \pi^{(t)}_{\nu_t(i)} \mu^{(t)}_{\nu_t(i)} \Big/ \sum_t \pi^{(t)}_{\nu_t(i)} \tag{3.31}$$

$$\widehat{\Sigma}_i = \sum_t \pi^{(t)}_{\nu_t(i)} \big(\Sigma^{(t)}_{\nu_t(i)} + (\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)(\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)^T\big) \Big/ \sum_t \pi^{(t)}_{\nu_t(i)} \tag{3.32}$$

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

Step 2: For $t = 1, \ldots, N$ choose ν_t to minimise

$$\sum_{i=1}^{k} \Big\{ \pi^{(t)}_{\nu_t(i)} \tfrac{1}{2} \log|\widehat{\Sigma}_i| + \pi^{(t)}_{\nu_t(i)} \tfrac{1}{2} \mathrm{tr}\big[\widehat{\Sigma}_i^{-1} \big(\Sigma^{(t)}_{\nu_t(i)} + (\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)(\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)^T\big)\big] - \pi^{(t)}_{\nu_t(i)} \log \widehat{\pi}_i - \big(1 - \pi^{(t)}_{\nu_t(i)}\big) \log\big(1 - \widehat{\pi}_i\big) \Big\}$$

and so Algorithm 3.1 translates to Algorithm 3.2.

Proof. The translation of Step 1 is given by Proposition 3.6 below. The translation of Step 2 follows from Lemma 3.5 below. □

Lemma 3.5 and Proposition 3.6 follow Lemma 3.4, which they will require.

Lemma 3.4. *The value of the integral*

$$I := \int \mathcal{N}(x; \mu_l, \Sigma_l) \log\big(\mathcal{N}(x; \mu_i, \Sigma_i)\big)\, dx$$

is

$$-\tfrac{1}{2} \log|2\pi\Sigma_i| - \tfrac{1}{2} \mathrm{tr}\big[\Sigma_i^{-1} \big(\Sigma_l + (\mu_l - \mu_i)(\mu_l - \mu_i)^T\big)\big]$$

Proof.

$$
\begin{aligned}
I &= \int \mathcal{N}(x;\mu_l,\Sigma_l)\log\big(\mathcal{N}(x;\mu_i,\Sigma_i)\big)\,dx \\
&= E\Big\{-\tfrac{1}{2}\log|2\pi\Sigma_i| - \tfrac{1}{2}\mathrm{tr}\big[\Sigma_i^{-1}(X_l-\mu_i)(X_l-\mu_i)^T\big]\Big\}
\end{aligned}
$$

where $X_l \sim \mathcal{N}(\mu_l,\Sigma_l)$

$$
\begin{aligned}
&= -\tfrac{1}{2}\log|2\pi\Sigma_i| - \tfrac{1}{2}E\Big\{\mathrm{tr}\big[\Sigma_i^{-1}\big((X_l-\mu_l)(X_l-\mu_l)^T + (\mu_l-\mu_i)(\mu_l-\mu_i)^T\big)\big]\Big\} \\
&= -\tfrac{1}{2}\log|2\pi\Sigma_i| - \tfrac{1}{2}\mathrm{tr}\big[\Sigma_i^{-1}\big(\Sigma_l + (\mu_l-\mu_i)(\mu_l-\mu_i)^T\big)\big].
\end{aligned}
$$

□

Lemma 3.5.

$$
\begin{aligned}
D[\nu(\theta)\,\|\,\widehat{\theta}] = C - \sum_i \Big\{ &\pi_{\nu(i)}\log\widehat{\pi}_i + (1-\pi_{\nu(i)})\log(1-\widehat{\pi}_i) - \pi_{\nu(i)}\tfrac{1}{2}\log|\widehat{\Sigma}_i| \\
&- \pi_{\nu(i)}\tfrac{1}{2}\mathrm{tr}\big[\widehat{\Sigma}_i^{-1}\big(\Sigma_{\nu(i)} + (\mu_{\nu(i)}-\widehat{\mu}_i)(\mu_{\nu(i)}-\widehat{\mu}_i)^T\big)\big]\Big\}
\end{aligned}
$$

where C is a function which does not depend on the permutation ν or on $\widehat{\theta} = (\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}})$.

Proof.

$$
\begin{aligned}
LHS &= D[\nu(\theta)\,\|\,\widehat{\theta}] \\
&= \sum_i \Delta\big[\pi_{\nu(i)}\mathcal{N}(\cdot;\mu_{\nu(i)},\Sigma_{\nu(i)})\,\|\,\widehat{\pi}_i\mathcal{N}(\cdot;\widehat{\mu}_i,\widehat{\Sigma}_i)\big] \\
&= \sum_i \Big\{ \pi_{\nu(i)}\log\frac{\pi_{\nu(i)}}{\widehat{\pi}_i} + (1-\pi_{\nu(i)})\log\frac{1-\pi_{\nu(i)}}{1-\widehat{\pi}_i} \\
&\qquad + \pi_{\nu(i)}\int \mathcal{N}(x;\mu_{\nu(i)},\Sigma_{\nu(i)})\log\frac{\mathcal{N}(x;\mu_{\nu(i)},\Sigma_{\nu(i)})}{\mathcal{N}(x;\widehat{\mu}_i,\widehat{\Sigma}_i)}\,dx\Big\} \\
&= \sum_i \Big\{ \pi_{\nu(i)}\log\pi_{\nu(i)} + (1-\pi_{\nu(i)})\log(1-\pi_{\nu(i)}) \\
&\qquad + \pi_{\nu(i)}\int \mathcal{N}(x;\mu_{\nu(i)},\Sigma_{\nu(i)})\log\mathcal{N}(x;\mu_{\nu(i)},\Sigma_{\nu(i)})\,dx\Big\} \\
&\qquad - \sum_i \Big\{ \pi_{\nu(i)}\log\widehat{\pi}_i + (1-\pi_{\nu(i)})\log(1-\widehat{\pi}_i) \\
&\qquad\qquad + \pi_{\nu(i)}\int \mathcal{N}(x;\mu_{\nu(i)},\Sigma_{\nu(i)})\log\mathcal{N}(x;\widehat{\mu}_i,\widehat{\Sigma}_i)\,dx\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_i \Big\{ \pi_i \log \pi_i + (1-\pi_i)\log(1-\pi_i) \\
&\quad + \pi_i \int \mathcal{N}(x;\mu_i,\Sigma_i)\log \mathcal{N}(x;\mu_i,\Sigma_i)\,dx \Big\} \\
&\qquad - \sum_i \Big\{ \pi_{\nu(i)} \log \widehat{\pi}_i + (1-\pi_{\nu(i)})\log(1-\widehat{\pi}_i) \\
&\qquad\quad - \pi_{\nu(i)} \big\{ \tfrac{1}{2}\log|2\pi\widehat{\Sigma}_i| + \tfrac{1}{2}\mathrm{tr}\big[\widehat{\Sigma}_i^{-1}\big(\Sigma_{\nu(i)} + (\mu_{\nu(i)}-\widehat{\mu}_i)(\mu_{\nu(i)}-\widehat{\mu}_i)^T\big)\big]\big\} \Big\} \\
&\qquad\qquad \text{[By Lemma 3.4]} \\
&= C - \sum_i \Big\{ \pi_{\nu(i)} \log \widehat{\pi}_i + (1-\pi_{\nu(i)})\log(1-\widehat{\pi}_i) - \pi_{\nu(i)}\tfrac{1}{2}\log|\widehat{\Sigma}_i| \\
&\qquad\quad - \pi_{\nu(i)}\tfrac{1}{2}\mathrm{tr}\big[\widehat{\Sigma}_i^{-1}\big(\Sigma_{\nu(i)} + (\mu_{\nu(i)}-\widehat{\mu}_i)(\mu_{\nu(i)}-\widehat{\mu}_i)^T\big)\big] \Big\} \\
&= RHS.
\end{aligned}
$$

□

Proposition 3.6. *The minimum over* $\widehat{\theta} = (\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ *of*

$$\mathcal{D} = \sum_{t=1}^{N} D[\nu_t(\theta^{(t)}) \,\|\, \widehat{\theta}]$$

is achieved at

$$\widehat{\pi}_i = \frac{1}{N}\sum_t \pi^{(t)}_{\nu_t(i)}$$

$$\widehat{\mu}_i = \sum_t \pi^{(t)}_{\nu_t(i)} \mu^{(t)}_{\nu_t(i)} \Big/ \sum_t \pi^{(t)}_{\nu_t(i)}$$

$$\widehat{\Sigma}_i = \sum_t \pi^{(t)}_{\nu_t(i)} \big(\Sigma^{(t)}_{\nu_t(i)} + (\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)(\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)^T\big) \Big/ \sum_t \pi^{(t)}_{\nu_t(i)}.$$

Proof. By Lemma 3.5 we can split the problem up into:

P1: Choose $\widehat{\pi}_i$ $(i = 1, \ldots, k)$ to maximise $\displaystyle\sum_{t=1}^{N} \Big\{ \pi^{(t)}_{\nu_t(i)} \log \widehat{\pi}_i + \big(1 - \pi^{(t)}_{\nu_t(i)}\big) \log\big(1 - \widehat{\pi}_i\big) \Big\}$

and

P2: Choose $\widehat{\mu}_i$ and $\widehat{\Sigma}_i$ $(i = 1, \ldots, k)$ to minimise

$$\sum_{t=1}^{N} \Big\{ \pi^{(t)}_{\nu_t(i)} \tfrac{1}{2} \log|\widehat{\Sigma}_i| + \pi^{(t)}_{\nu_t(i)} \tfrac{1}{2} \mathrm{tr}\big[\widehat{\Sigma}_i^{-1}\big(\Sigma^{(t)}_{\nu_t(i)} + (\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)(\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)^T\big)\big] \Big\}.$$

Dividing P1 through by N we obtain

$$\text{Choose } \widehat{\pi}_i \text{ to maximise } \frac{\sum_t \pi^{(t)}_{\nu_t(i)}}{N} \log \widehat{\pi}_i + \left(1 - \frac{\sum_t \pi^{(t)}_{\nu_t(i)}}{N}\right) \log(1 - \widehat{\pi}_i)$$

for which the solution is well known to be

$$\widehat{\pi}_i = \frac{\sum_t \pi^{(t)}_{\nu_t(i)}}{N}.$$

P2 can be solved for $\widehat{\mu}_i$ by differentiating, and setting the derivative to zero, which gives

$$\widehat{\mu}_i = \sum_t \pi^{(t)}_{\nu_t(i)} \mu^{(t)}_{\nu_t(i)} \Big/ \sum_t \pi^{(t)}_{\nu_t(i)}.$$

Eaton and Olkin (1987, page 1642) show that for positive definite symmetric matrices A and B, the minimum of

$$\text{tr}(AB) - \log |B|$$

is achieved at $B = A^{-1}$, or equivalently, the minimum of

$$\text{tr}(B^{-1}A) + \log |B|$$

is at $B = A$. From this we deduce

$$\widehat{\Sigma}_i = \sum_t \pi^{(t)}_{\nu_t(i)} \big(\Sigma^{(t)}_{\nu_t(i)} + (\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)(\mu^{(t)}_{\nu_t(i)} - \widehat{\mu}_i)^T\big) \Big/ \sum_t \pi^{(t)}_{\nu_t(i)}.$$

□

Chapter 4

Bayesian analysis of mixtures with an unknown number of components

We now consider the case where we wish to model data $x^n = (x_1, \ldots, x_n)$ as independent observations from a mixture with an *unknown* number of components k, with density

$$p(x \mid k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \pi_1 f(x; \phi_1, \eta) + \cdots + \pi_k f(x; \phi_k, \eta) \tag{4.1}$$

where the mixture proportions $\pi_1, \ldots, \pi_k$ are constrained to be non-negative and sum to unity, $\phi_1, \ldots, \phi_k$ are vectors of component-specific parameters, and η is a vector of parameters common to all components. If the components of the mixture have a physical interpretation then inference for k may be one of the main aims of our analysis. Alternatively, if the components of the mixture have no obvious physical interpretation and the mixture model is being used as a convenient representation of an unknown density, then the choice of k will affect the flexibility of the model. Density estimates based on small values of k will tend to be smoother than those based on larger k — compare for example the density estimates for the galaxy data shown in Figure 2.2 based on $k = 3$ and $k = 6$. Procedures which allow k to vary may therefore be of interest whether or not the components have a physical interpretation.

A classical approach to this problem is to perform an analysis of the data using a range of different values of k, and then attempt to select k using some model-choice criterion such as Akaike's AIC (Akaike, 1973), or by hypothesis testing. However, the absence of a single dominating local maximum in the likelihood (see Section 1.1.3) means that the theory underlying AIC does not apply in the mixture model context. Hypothesis testing is also not straightforward in this context, as the theory used to construct the null distribution of the likelihood ratio statistic does not hold (see for example McLachlan and Basford, 1988, pages 21–29). Furthermore, these methods restrict us to choosing a particular value for k, and do not

provide us with a coherent method of combining our results for different values of k; something we may wish to do in the context of predictive density estimation for example.

In the Bayesian paradigm, inference is based on the posterior distribution of the parameters $p(\theta \mid x^n)$ where $\theta = (k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$. Inference for k is then based on the marginal posterior distribution

$$\Pr(k = i \mid x^n) \quad i = 1, 2, 3, \ldots. \tag{4.2}$$

Models $k = i$ and $k = l$ may be compared via the ratio of their posterior probabilities

$$\begin{aligned} \frac{\Pr(k = i \mid x^n)}{\Pr(k = l \mid x^n)} &= \frac{p(x^n \mid k = i) \Pr(k = i)}{p(x^n \mid k = l) \Pr(k = l)} \\ &= B_{il} \frac{\Pr(k = i)}{\Pr(k = l)} \end{aligned} \tag{4.3}$$

where

$$B_{il} = \frac{p(x^n \mid k = i)}{p(x^n \mid k = l)} \tag{4.4}$$

is known as the *Bayes factor* for comparing $k = i$ with $k = l$. The Bayes factor does not depend on the prior distribution of k, and as a result researchers often report Bayes factors rather than posterior probabilities.

Work on this problem has been fairly recent. Escobar and West (1995) consider an approach which assumes a prior structure based on the Dirichlet process, which (according to West, 1997) is

> "more geared towards density estimation than deconvolution or parameter estimation; here the number of components is treated simply as a nuisance parameter to be averaged away."

Other approaches include Chib (1995), Carlin and Chib (1995) and Phillips and Smith (1996).

More recently Richardson and Green (1997) consider the problem in the context of mixture of univariate normal distributions with parameters $\theta = (k, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, though the methods they develop are more generally applicable. Assuming a suitable prior distribution of the form

$$p(\theta) = p(k) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mid k) \tag{4.5}$$

they describe a method of constructing an irreducible Markov chain with $p(\theta \mid x^n)$ as its stationary distribution using "reversible jump" Markov Chain Monte Carlo methods (Green, 1994, 1995). If $\theta^{(1)}, \ldots, \theta^{(N)}$ is the realisation of such a chain

then (by Theorem 2.1, Section 2.1) we may estimate quantities of interest by an appropriate sample path average. For example,

$$\begin{aligned}\Pr(k = i \mid x^n) &= E(I(k = i) \mid x^n) \\ &\approx \frac{1}{N} \sum_{t=1}^{N} I(k^{(t)} = i) \\ &= \frac{1}{N} \#\{t : k^{(t)} = i\}. \end{aligned} \tag{4.6}$$

and the predictive density for a future observation may be estimated by averaging over different values of k:

$$\begin{aligned} p(x_{n+1} \mid x^n) &= \int p(x_{n+1} \mid \theta) p(\theta \mid x^n)\, d\theta \\ &\approx \frac{1}{N} \sum_{t=1}^{N} p(x_{n+1} \mid \theta^{(t)}). \end{aligned} \tag{4.7}$$

This approach has a number of advantages over considering models for different fixed values of k separately. In particular it can lead to improved mixing over the mixture model parameters (we see an example of this in Section 4.4.1), and it removes the need to calculate the marginal densities $p(x^n \mid k)$ (which is not altogether straightforward) when finding the posterior distribution for k.

In this chapter we present an alternative method of constructing a Markov chain with $p(\theta \mid x^n)$ as its stationary distribution. 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, extending work by Ripley (1977) who first applied this idea to the simulation of point processes. The method we present appears to be computationally slightly more expensive that that of Richardson and Green (1997) in the context of mixtures of univariate normal distributions, though direct comparisons are difficult and the efficiency of both methods may be improved by future modification. Both methods certainly give computationally tractable solutions to the problem, with rough results available in a matter of minutes. However, the mathematics required to perform our algorithm is much simpler, particularly in the case of bivariate data; work is still in progress on extending the method of Richardson and Green to this case (Posse, 1997). Our method is illustrated by fitting mixtures of normal (and t) distributions to univariate and bivariate data in Sections 4.4–4.9.

4.1 Construction of the birth-death process

Consider the model (4.1) with parameters $(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$. We will assume that given *hyperparameters* ω, and common component parameters η, the prior distribution of $(k, \boldsymbol{\pi}, \boldsymbol{\phi})$ is of the form

$$p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \mid \omega, \eta) = p(k \mid \omega, \eta) p(\boldsymbol{\pi}, \boldsymbol{\phi} \mid k, \omega, \eta) \tag{4.8}$$

where $p(\boldsymbol{\pi}, \boldsymbol{\phi} \,|\, k, \omega, \eta)$ is a distribution in which $\boldsymbol{\pi}$ and $\phi_1, \ldots, \phi_k$ are independent with

$$\boldsymbol{\pi} \,|\, k, \omega, \eta \sim \mathcal{D}(1, \ldots, 1) \tag{4.9}$$

and $\phi_1, \ldots, \phi_k$ independently distributed on a space Φ according to density

$$\widetilde{p}(\cdot \,|\, \omega, \eta). \tag{4.10}$$

We will further assume that $\widetilde{p}(\cdot \,|\, \omega, \eta)$ is a density we can simulate from. In this section we will treat ω and η as fixed, and describe the construction of an irreducible Markov chain with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \,|\, x^n, \omega, \eta)$. In Section 4.2 we will describe how this can be combined with Gibbs sampling steps to create an irreducible Markov chain with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \omega, \eta \,|\, x^n)$. Quantities of interest may then be estimated by forming sample path averages (by Theorem 2.1).

4.1.1 The connection with point processes

We note that the prior distribution of $(\boldsymbol{\pi}, \boldsymbol{\phi})$ defined above does not depend on the labelling of the components, in that

$$p\big((\pi_1, \ldots, \pi_k), (\phi_1, \ldots, \phi_k) \,|\, k, \omega, \eta\big) = p\big((\pi_{\nu(1)}, \ldots, \pi_{\nu(k)}), (\phi_{\nu(1)}, \ldots, \phi_{\nu(k)}) \,|\, k, \omega, \eta\big)$$

for all permutations ν of $1, \ldots, k$. Since the likelihood

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

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

$$p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \,|\, x^n, \omega, \eta) \propto L(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \,|\, \omega, \eta) \tag{4.11}$$

will be similarly invariant. If we ignore the labelling of the components, we can consider any set of k parameter values $\big\{(\pi_1, \phi_1), \ldots, (\pi_k, \phi_k)\big\}$ as a set of k points in $[0, 1] \times \Phi$, with the constraint that $\pi_1 + \cdots + \pi_k = 1$. The posterior distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \,|\, x^n)$ can thus be seen as a (suitably constrained) distribution of points in $[0, 1] \times \Phi$, or in other words a *point process* on $[0, 1] \times \Phi$. (Alternatively the posterior distribution can be seen as a distribution of points 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 is known as a *marked point process*.)

Methods for simulating point processes often rely on the construction of a continuous time Markov process with appropriate stationary distribution; an idea which originated with Ripley (1977). A continuous time Markov process is a sequence of random variables $\{\Theta^{(t)}\}$ which obeys the Markov property: for any

fixed t_0 the distribution of $\{\Theta^{(t_0+s)} : s > 0\}$ given $\Theta^{(t_0)}$ is independent of $\{\Theta^{(t)} : t < t_0\}$. $\{\Theta^{(t)}\}$ is said to have stationary distribution π if

$$\Theta^{(t)} \sim \pi \Rightarrow \Theta^{(t+s)} \sim \pi \text{ for all } s > 0.$$

It is clear that for any fixed time period t_0, $\{\Theta^{(t_0)}, \Theta^{(2t_0)}, \Theta^{(3t_0)}, \dots\}$ is then a Markov chain in discrete time with stationary distribution π. We describe the construction of a continuous time Markov process with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \mid x^n, \omega, \eta)$, from which we construct a discrete time Markov chain with the same stationary distribution.

4.1.2 Introduction to general Markov birth-death processes

Preston (1976, Section 5) describes a general Markov birth-death process on a state space $\Omega = \bigcup_k \Omega_k$ where the Ω_k are disjoint. Briefly, the idea is that 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. The time to the next birth/death event is therefore exponentially distributed, with mean $1/\big(\beta(y) + \delta(y)\big)$, and it will be a birth with probability

$$\Pr(\text{birth}) = \frac{\beta(y)}{\beta(y) + \delta(y)}$$

and a death with probability

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

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 $K_\delta^{(k)}(y; G)$.

By the memoryless property of the exponential distribution this process has the Markov property. It is necessary to impose conditions on β and δ to ensure that the process exists, and has a stationary distribution. Suitable conditions, and a more formal construction are given by Preston (1976).

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

We will construct a birth-death process on the state space $\Omega = \bigcup_{k \geq 1} \Omega_k$, where Ω_k is the parameter space of the mixture model with k components, ignoring

the labelling of the components. A birth corresponds to increasing the number of components by one, while a death corresponds to decreasing the number of components by one. 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

$$p(x) = \pi_1 f(x; \phi_1, \eta) + \cdots + \pi_k f(x; \phi_k, \eta), \tag{4.12}$$

and we may write $(\pi_i, \phi_i) \in y$ for $i = 1, \ldots, k$. (Recall that we are treating η as fixed.)

We restrict births and deaths to be of the following form:

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

$$y \cup (\pi, \phi) := \Big\{ \big(\pi_1(1-\pi), \phi_1\big), \ldots, \big(\pi_k(1-\pi), \phi_k\big), \big(\pi, \phi\big) \Big\} \in \Omega_{k+1}. \tag{4.13}$$

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

$$\begin{aligned} y \setminus (\pi_i, \phi_i) := \Big\{ &\Big(\frac{\pi_1}{1-\pi_i}, \phi_1\Big), \ldots, \Big(\frac{\pi_{i-1}}{1-\pi_i}, \phi_{i-1}\Big), \\ &\Big(\frac{\pi_{i+1}}{1-\pi_i}, \phi_{i+1}\Big), \ldots, \Big(\frac{\pi_k}{1-\pi_i}, \phi_k\Big) \Big\} \in \Omega_{k-1}. \end{aligned} \tag{4.14}$$

These definitions have been chosen to ensure that the constraint $\pi_1 + \cdots + \pi_k = 1$ remains satisfied after a birth or death.

The following proposition defines a general Markov birth-death process with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \,|\, x^n, \omega, \eta)$. The definition may look complicated, but it is very straightforward to create a simple process of this form which is easily simulated from, as we explain in Section 4.1.4.

Proposition 4.1. *Assume that given hyperparameters ω, and common component parameters η, the prior distribution of $(k, \boldsymbol{\pi}, \boldsymbol{\phi})$ is of the form*

$$p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \,|\, \omega, \eta) = p(k \,|\, \omega, \eta) p(\boldsymbol{\pi}, \boldsymbol{\phi} \,|\, k, \omega, \eta)$$

where $p(\boldsymbol{\pi}, \boldsymbol{\phi} \,|\, k, \omega, \eta)$ is a distribution in which $\boldsymbol{\pi}$ and $\phi_1, \ldots, \phi_k$ are independent with

$$\boldsymbol{\pi} \,|\, k, \omega, \eta \sim \mathcal{D}(1, \ldots, 1) \tag{4.15}$$

and $\phi_1, \ldots, \phi_k$ independently distributed on a space Φ according to density

$$\widetilde{p}(\cdot \,|\, \omega, \eta) \tag{4.16}$$

which we assume we can simulate from.

Then 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

$$\beta(y)K_\beta^{(k)}(y;F) = \int_{(\pi,\phi):y\cup(\pi,\phi)\in F} b\big(y;(\pi,\phi)\big)k(1-\pi)^{k-1}\widetilde{p}(\phi\,|\,\omega,\eta)\,d\pi\,d\phi \tag{4.17}$$

and

$$\delta(y)K_\delta^{(k)}(y;F) = \sum_{(\pi,\phi)\in y:y\setminus(\pi,\phi)\in F} d\big(y\setminus(\pi,\phi);(\pi,\phi)\big) \tag{4.18}$$

for some $b:\Omega\times([0,1]\times\Phi)\to R^+$ and $d:\Omega\times([0,1]\times\Phi)\to R^+$, has stationary distribution $p(k,\boldsymbol{\pi},\boldsymbol{\phi}\,|\,x^n,\omega,\eta)$, provided

$$(k+1)d\big(y;(\pi,\phi)\big)L\big(y\cup(\pi,\phi)\big)\frac{p(k+1\,|\,\omega,\eta)}{p(k\,|\,\omega,\eta)} = b\big(y;(\pi,\phi)\big)L(y) \tag{4.19}$$

for all $y\in\Omega_k$ and $(\pi,\phi)\in[0,1]\times\Phi$, where $L(y)$ is the likelihood of the model represented by y:

$$L\big(\{(\pi_1,\phi_1),\ldots,(\pi_k,\phi_k)\}\big) = \prod_{j=1}^{n}\big[\pi_1 f(x_j;\phi_1,\eta)+\cdots+\pi_k f(x_j;\phi_k,\eta)\big].$$

Proof. The proof is deferred to the appendix to this chapter (Section 4.11.1). □

4.1.4 An easily simulated process

Consider the process obtained by setting

$$\beta(y) = b\big(y;(\pi,\phi)\big) = \lambda_b \text{ (a constant)} \tag{4.20}$$

in Proposition 4.1. In this process births and deaths occur as follows:

Births: Births occur at a constant rate λ_b, and when a birth occurs it occurs at a point $(\pi,\phi)\in[0,1]\times\Phi$, chosen according to density

$$q(\pi,\phi) = k(1-\pi)^{k-1}\widetilde{p}(\phi\,|\,\omega,\eta).$$

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

$$d\big(y\setminus(\pi_j,\phi_j)\,;(\pi_j,\phi_j)\big),$$

and so we have

$$\delta(y) = \sum_{(\pi,\phi)\in y} d\big(y\setminus(\pi,\phi)\,;(\pi,\phi)\big)$$

and when a death occurs it occurs at $(\pi_j, \phi_j) \in y$ with probability

$$\frac{d\big(y\backslash(\pi_j,\phi_j)\,;(\pi_j,\phi_j)\big)}{\delta(y)}.$$

Algorithm 4.1 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 4.1. 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, given by:

$$\begin{aligned}\delta_j(y) &= d\big(y\backslash(\pi_j,\phi_j)\,;(\pi_j,\phi_j)\big)\\ &= \lambda_b \frac{L\big(y\backslash(\pi_j,\phi_j)\big)}{L(y)} \frac{p(k\,|\,\omega,\eta)}{(k+1)p(k+1\,|\,\omega,\eta)} \quad (j=1,\ldots,k)\end{aligned}$$

[from (4.19).]

3. Calculate the total death rate $\delta(y) = \sum_j \delta_j(y)$.

4. Simulate the time s 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

$$\Pr(\text{birth}) = \frac{\beta(y)}{\beta(y)+\delta(y)}$$

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

6. Adjust y to reflect the birth or death:

   Birth: Simulate the point (π,ϕ) at which a birth takes place 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 $\pi = Y_1/(Y_1+Y_2)$.

   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 1.

4.2 A Markov chain with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \omega, \eta \mid x^n)$

Algorithm 4.1 describes the simulation of a continuous time Markov process with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \mid x^n, \omega, \eta)$. If we consider this process only at fixed time intervals $t_0, 2t_0, \ldots$ then this defines a discrete time Markov chain with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \mid x^n, \omega, \eta)$. By considering the parameter vector augmented by the allocation variables, $\theta = (k, \boldsymbol{\pi}, \boldsymbol{\phi}, \omega, \eta, z^n)$ and combining Algorithm 4.1 with steps of the Gibbs sampler we can simulate from a Markov chain with stationary distribution $p(\theta \mid x^n)$, as described in Algorithm 4.2 below. This Markov chain will be irreducible (and so satisfy the conditions of Theorem 2.1) provided the full conditional posterior distributions for each parameter give support to the full parameter space, as will often be the case.

Algorithm 4.2. 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 $(\omega^{(t)}, \eta^{(t)})$.

Step 2: Set $k^{(t+1)} = k^{(t)'}$.

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

Step 4: Perform Gibbs sampling steps to obtain $\eta^{(t+1)}$ and $\omega^{(t+1)}$.

Step 5: Perform Gibbs sampling steps to obtain $\boldsymbol{\pi}^{(t+1)}$ and $\boldsymbol{\phi}^{(t+1)}$.

Proposition 4.2. *Algorithm 4.2 defines a Markov chain with stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \omega, \eta, z^n \mid x^n)$, which will be irreducible provided the full conditional posterior distributions for each parameter give support to all parts of the parameter space.*

Proof. Suppose $\theta^{(t)} \sim p(\theta \mid x^n)$. We wish to show that $\theta^{(t+1)} \sim p(\theta \mid x^n)$ and so this is the stationary distribution of the Markov chain. We note that

$$p(\theta \mid x^n) = p(\omega, \eta \mid x^n) p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \mid \omega, \eta, x^n) p(z^n \mid k, \boldsymbol{\pi}, \boldsymbol{\phi}, \omega, \eta, x^n). \qquad (4.21)$$

Step 1 has stationary distribution $p(k, \boldsymbol{\pi}, \boldsymbol{\phi} \mid \eta^{(t)}, \omega^{(t)}, x^n)$ and so after Step 1

$$(k^{(t)'}, \boldsymbol{\pi}^{(t)'}, \boldsymbol{\phi}^{(t)'}, \eta^{(t)}, \omega^{(t)}) \sim p(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \omega, \eta \mid x^n).$$

Then after Step 3,

$$(k^{(t+1)}, \boldsymbol{\pi}^{(t)'}, \boldsymbol{\phi}^{(t)'}, \eta^{(t)}, \omega^{(t)}, (z^n)^{(t+1)}) \sim p(k, \boldsymbol{\pi}, \boldsymbol{\phi}, \omega, \eta, z^n \mid x^n),$$

and the result follows as in the proof of Proposition 2.2 (page 19). We note that Step 5 could be replaced by setting $\boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi}^{(t)'}$ and $\boldsymbol{\phi}^{(t+1)} = \boldsymbol{\phi}^{(t)'}$ without affecting the irreducibility or stationary distribution of the chain. However, the mixing of the chain will generally be improved by the Gibbs sampling step. □

4.3 Examples: prior distributions and values for (t_0, λ_b)

Our examples (Sections 4.4 to 4.9) demonstrate the use of Algorithm 4.2 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. In the case of univariate data we also consider modelling the data as arising from a mixture of an unknown (finite) number of t distributions (with fixed degrees of freedom).

4.3.1 Prior distributions

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

$$p(k) \propto \frac{\lambda^k}{k!} \quad k = 1, \ldots, k_{max} = 100 \tag{4.22}$$

where λ is a constant; we will perform analyses with several different values of λ.

Conditional on k we use the following priors for the model parameters:

1. The *fixed-κ* prior, which is the name we give to the hierarchical prior described in Chapter 2 (Section 2.4, page 2.4), which was also used by Richardson and Green (1997). In the notation of the previous section we have

$$\phi_i = (\mu_i, \Sigma_i) \tag{4.23}$$
$$\eta \text{ unused} \tag{4.24}$$
$$\omega = \beta \tag{4.25}$$
$$\widetilde{p}(\phi \,|\, \omega, \eta) = p(\mu, \Sigma \,|\, \beta) \tag{4.26}$$

   with

$$\mu_i \sim \mathcal{N}_r(\xi, \kappa^{-1}) \quad (i = 1, \ldots, k) \tag{4.27}$$
$$\Sigma_i^{-1} \,|\, \beta \sim \mathcal{W}_r(2\alpha, (2\beta)^{-1}) \quad (i = 1, \ldots, k) \tag{4.28}$$
$$\beta \sim \mathcal{W}(2g, (2h)^{-1}) \tag{4.29}$$

   with the constants given by equations (2.39)–(2.43) on page 37. The full conditional posterior distributions required for the Gibbs sampling steps in Algorithm 4.2 are then as given in equations (2.46) to (2.49). 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 (so $\omega = (\beta, \xi, \kappa)$) 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. 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.

The full conditional posteriors used by the Gibbs sampling steps of Algorithm 4.2 are then as for the fixed-κ prior, with the addition of:

$$\xi \mid \cdots \sim \mathcal{N}_r(\bar{\mu}, (k\kappa)^{-1}) \tag{4.30}$$

$$\kappa \mid \cdots \sim \mathcal{W}_r(\kappa; l+k+1, (lI_r + SS)^{-1}) \tag{4.31}$$

where $\bar{\mu} = \sum_i \mu_i / k$ and $SS = \sum_i (\mu_i - \xi)^2$. The Gibbs sampling updates were performed in the order $(\beta, \kappa, \xi, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

These priors were chosen for their computational convenience. We will see that inference for k can be highly sensitive to the priors used, and warn against considering these priors as "non-informative" or "weakly" informative. Further discussion is deferred to Section 4.10, after we have examined some examples of the use of these priors.

4.3.2 Values for (t_0, λ_b)

Algorithm 4.1 requires the specification of a birth-rate λ_b, and Algorithm 4.2 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. Larger values of λ_b will result in better mixing over k at the cost of more computation time per iteration of Algorithm 4.2. We found that setting $\lambda_b = \lambda$ (the parameter of the Poisson prior in (4.22)) generally gave reasonable performance in the examples we considered.

4.4 Example 1: Galaxy data

As our first example of the use of Algorithm 4.2, we use it to fit the following mixture models to the galaxy data:

a) A mixture of an unknown number of normal distributions using the fixed-κ prior described in Section 4.3.1.

b) A mixture of an unknown number of normal distributions using the variable-κ prior described in Section 4.3.1.

c) A mixture of an unknown number of t distributions on 4 degrees of freedom, with the fixed-κ prior.

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 a)-c) above 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

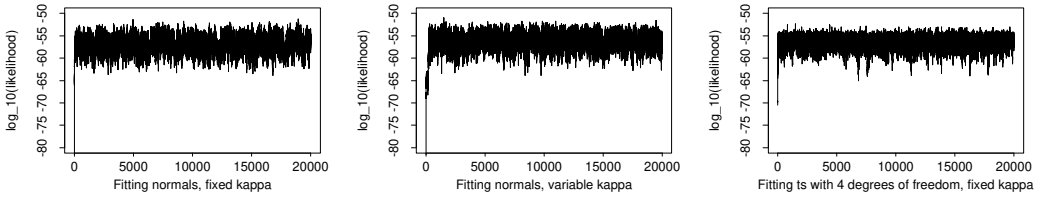


Figure 4.1: $\log_{10}$(likelihood) of the parameter values sampled using Algorithm 4.2 to fit the three different models a)-c) to the galaxy data using $\lambda = 3$. **Left:** Normals, fixed-κ; **Middle:** Normals, variable-κ; **Right:** t_4s, fixed-κ. In each case the sampler moves quickly to an area of higher likelihood and does not appear to get stuck in areas of low or high likelihood.

idea of how the method would behave as λ was allowed to get very large. None of the runs reached the upper limit of $k_{max} = 100$, and so the results were not materially affected by the truncation of the prior on k at this limit.

4.4.1 Starting points, computational expense, and mixing behaviour

Each run consisted of $20\,000$ iterations of Algorithm 4.2, 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. Figure 4.1 shows graphs of the $\log_{10}$(likelihood) for the runs of the sampler with $\lambda = 3$. We see that the sampler moves quickly from the low likelihood of the starting point to an area of parameter space with higher likelihood, and there is no evidence that the sampler gets stuck for prolonged periods in "trapping states".

The runs for $\lambda = 3$ for the three models a)-c) took 239 seconds, 230 seconds and 267 seconds[1] respectively, which corresponds to about 80 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). The time spent performing birth-death calculations could also be decreased by reducing the (virtual) time t_0 for which the birth-death process is run at each iteration, but this would be at the expense of poorer mixing. It is not easy to see how an optimal balance between computational expense *per* iteration and the quality of mixing should be achieved.

Mixing over k

Figure 4.2 shows the sampled values of k for the runs with $\lambda = 3$. It can be seen that the value of k changes rapidly; in fact the percentage of iterations of the algorithm in which k changed for the three models a)-c) were 36%, 52%, and 38%

[1]CPU times on a Sun UltraSparc 200 workstation

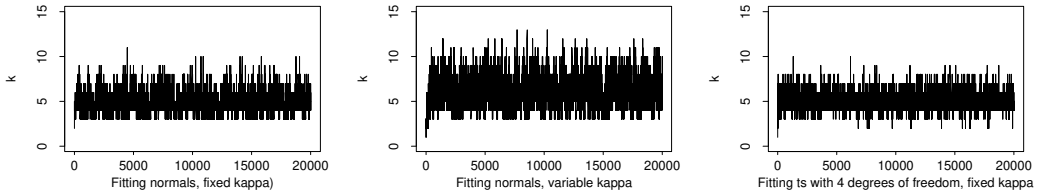


Figure 4.2: Values of k sampled using Algorithm 4.2 when fitting the three different models to the galaxy data using $\lambda = 3$. The three columns show results for **Left:** Normals, fixed-κ; **Middle:** Normals, variable-κ; **Right:** t_4s, fixed-κ. In each case k can be seen to vary rapidly, suggesting that the sampler mixes well over k, and so (4.6) will give a reliable estimate of $p(k \mid x^n)$ for values of k visited reasonably often.

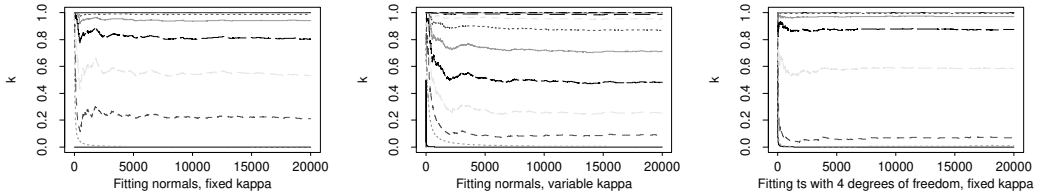


Figure 4.3: Graphs for the galaxy data of $\widehat{p}_T(k \leq j)$ (given by equation (4.32)) against T, for increasing values of j. These are based on the sampled values of k shown in Figure 4.2, and obtained using Algorithm 4.2 when fitting the three different models to the galaxy data using $\lambda = 3$. The columns show results for **Left:** Normals, fixed-κ; **Middle:** Normals, variable-κ; **Right:** t_4s, fixed-κ. The gap between the lines for j and $j - 1$ gives an estimate for $\Pr(k = j)$ based on the first T samples, and these estimates can be seen to stabilise quickly.

respectively. Figure 4.3 shows graphs against T of

$$\widehat{p}_T(k \leq j) = \frac{\#\{t : k^{(t)} \leq j; t = 1, \ldots, T\}}{T} \tag{4.32}$$

for increasing values of j. The gaps between the lines for j and $j-1$ give an estimate for $\Pr(k = j)$ based on the first T samples ($T = 1, \ldots, 20\,000$), and we see that the rapid mixing leads to these estimates stabilising fairly quickly.

Mixing over the mixture model parameters (within k)

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. For example, we saw in previous chapters (Sections 2.3 and 3.4.2) how when fitting $k = 3$ t_4 distributions to the galaxy data with the fixed-κ prior, there are two well-separated modes, which the Gibbs sampler (with k fixed) struggles to move between (the minor mode is visited only once in $10\,000$ iterations). We applied Algorithm 4.2 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 4.4). 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 with means near 10, 20, and 23 to the minor mode with means near 10, 21 and 34 *via* the 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. In an attempt to gain efficiency we tried truncating the prior on k to allow only $k = 3$ and $k = 4$, using $\lambda = 0.1$ to favour the 3 component model strongly. In $10\,000$ iterations of the resulting sampler there were 7371 visits to $k = 3$, with about 6 separate visits to the minor mode (results not shown). Alternative strategies for obtaining a sample from the birth-death process conditional on a fixed value of k are given by Ripley (1977).

4.4.2 Inference for k

In the Bayesian paradigm, inference for k is based on the posterior distribution of k given the data. Figure 4.5 shows histograms which enable us to see how using different models (t_4 versus normal) and different priors on k ($\lambda = 1, 3, 6$) and on the parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ (fixed-κ or variable-κ) affects the posterior distribution of k. The histograms show the total proportion of sampled values of each value of k, after having discarded the first $10\,000$ samples of each run to ensure that the algorithm had reached an equilibrium. We see that the posterior distribution of k is sensitive to the prior used, both in terms of choice of λ and the prior (variable-κ or fixed-κ) used on the parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$.

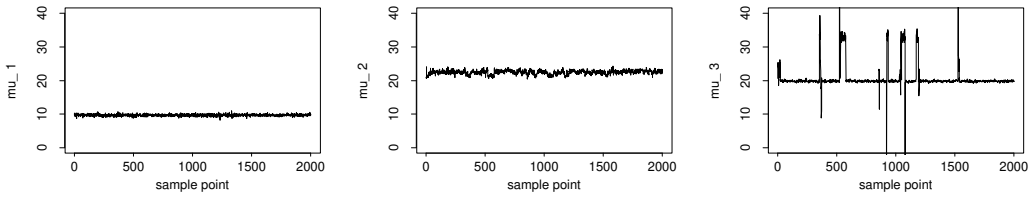


Figure 4.4: Sampled values of means for three components, sampled using Algorithm 4.2 when fitting a variable number of t_4 components to the galaxy data, with fixed-κ prior, $\lambda = 1$, and conditioning the resulting output on $k = 3$. The resulting sample of 1913 sample points was permuted by applying Algorithm 3.3 (see previous chapter). Comparing this figure with the corresponding results for the fixed k sampler (Figure 3.7c) we see that the variable k sampler visits the minor mode at least 6 separate times in 1913 iterations, compared with once in $10\,000$ iterations for the fixed k sampler.

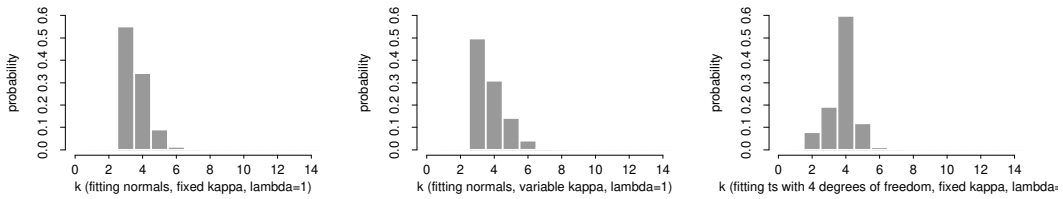
4.4.3 Predictive density estimation

Figure 4.6 shows the predictive density estimates (4.7) based on the different runs. We see that these estimates depend more on the value of λ chosen than on the choice of normal or t_4, and fixed-κ or variable-κ. 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 overfitting badly.

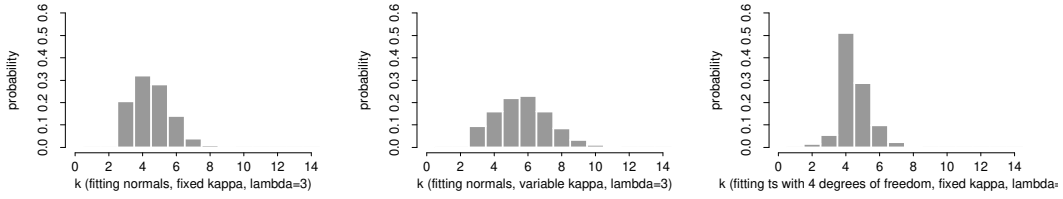
4.4.4 Interpreting components via label-switching

The variable-k sampler developed here has ignored the labelling of the components of the mixture, and so produces a sample which is essentially "unlabelled". Our work in Chapter 3 suggests that in order to interpret the results of our analysis in terms of the individual components fitted to the data, we might seek a labelling of the sample which gives interpretable predictive density estimates for the individual components. This can be done by considering the subset of the sample obtained by conditioning on a fixed value of k, and applying Algorithm 3.2 or Algorithm 3.3 (from the previous chapter) to this subset. For example, consider our results for the Normal, variable-κ model with $\lambda = 3$. If we apply Algorithm 3.2 successively to the 6 subsets of the sample obtained by conditioning on $k = 3, 4, 5, 6, 7$ and 8 (the union of which covers over 95% of the sample) then the resulting predictive density estimates for the individual components are as shown in Figure 4.7.

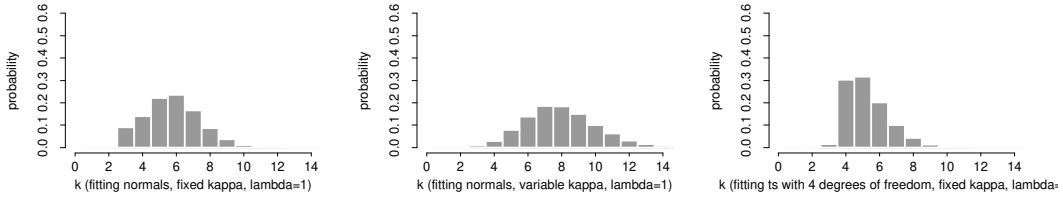
If we examine the components corresponding to $k = 5$ we see four fairly normal components, and a fifth which is multimodal, which we interpret as being due to genuine multimodality in the posterior distribution of the parameters. This fifth component is split roughly into 2 when $k = 6$, into 3 for $k = 7$, and into 4 for $k = 8$, while the other 4 components remain relatively unchanged, except for a steady decrease in the weight and variance of the component centred near 23. This raises the question of whether we might be able to obtain an alternative view of



(a) $\lambda = 1$

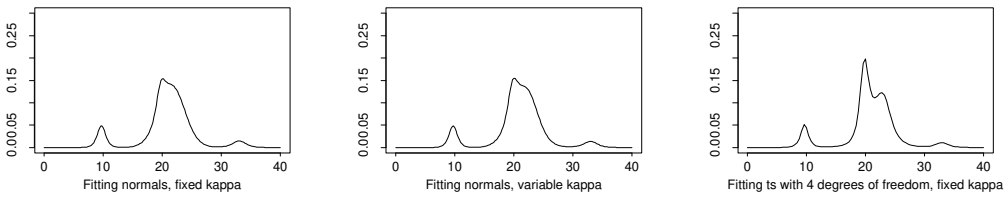


(b) $\lambda = 3$

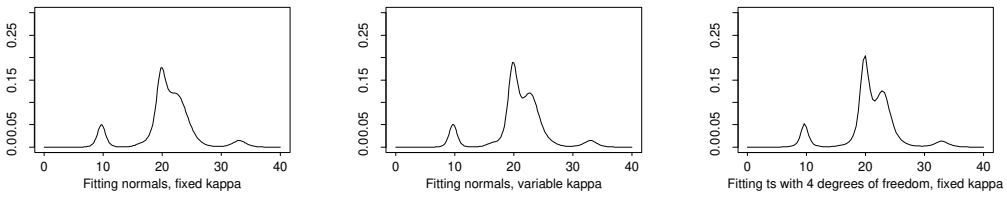


(c) $\lambda = 6$

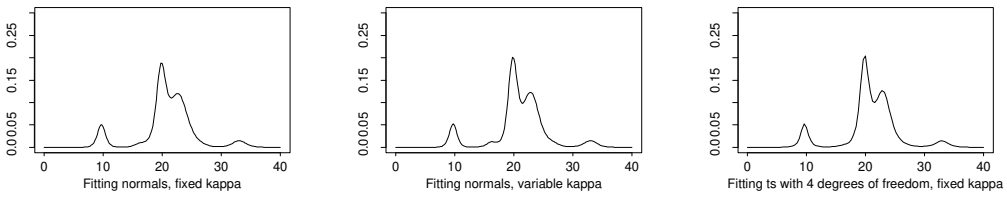
Figure 4.5: Graphs showing estimates (4.6) of $\Pr(k = i)$ for $i = 1, 2, \ldots$, for the galaxy data. These estimates are based on the values of k sampled using Algorithm 4.2 when fitting the three different models to the galaxy data with $\lambda = 1, 3, 6$, with in each case the first $10\,000$ samples 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).



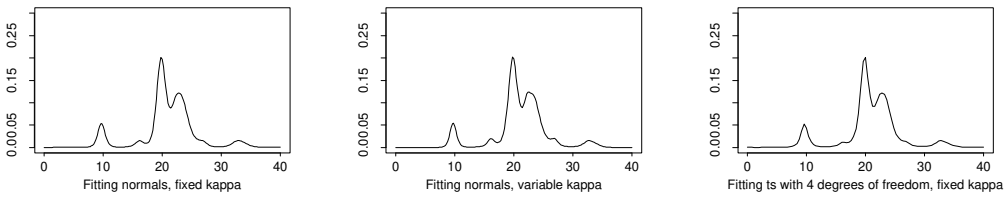
(a) $\lambda = 1$



(b) $\lambda = 3$



(c) $\lambda = 6$



(d) $\lambda = 25$

Figure 4.6: Predictive density estimates (4.7) for the galaxy data. These are based on the output of Algorithm 4.2 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.

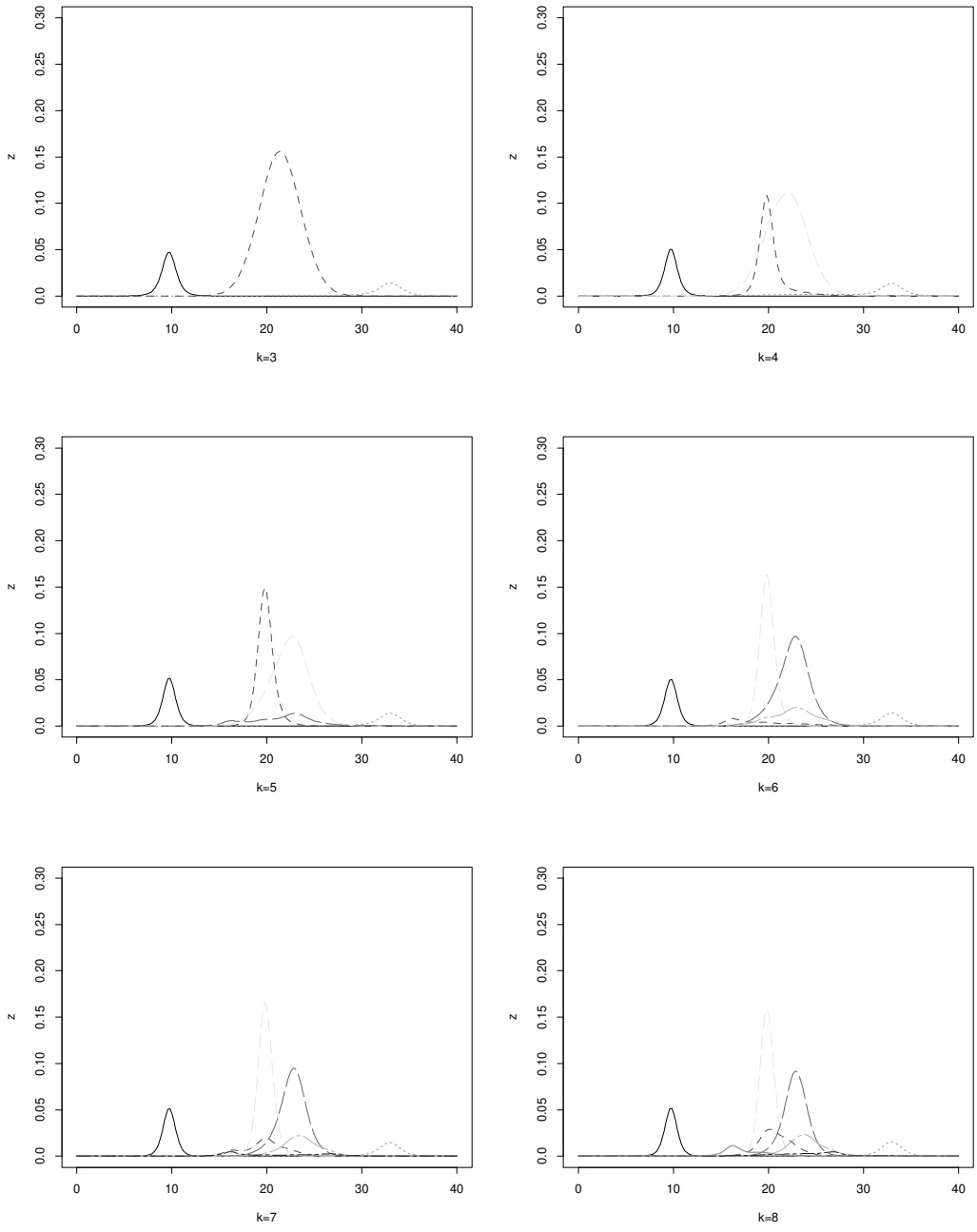


Figure 4.7: Predictive density estimates of individual components fitted to the galaxy data, conditional on $k = 3, 4, 5, 6, 7$ and 8 (reading from left-right, top-bottom). These were obtained by applying Algorithm 4.2 to fit normal distributions to the galaxy data with the variable-κ prior on $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ and with $\lambda = 3$. The resulting sample was split into 6 separate samples according to $k = 3, 4, 5, 6, 7$ and 8, and these samples were then permuted by applying Algorithm 3.2 (from the previous chapter) to each of them in turn.

the posterior by combining the results for all different ks, and grouping together components which are "similar", in that they have similar predictive density estimates. However, attempts to do this have failed to produce any easily interpretable results.

4.4.5 t distributions or normal distributions?

We turn now to the problem of whether we should model the data with t distributions or normal distributions. A fully Bayesian approach to this problem might be to fit t distributions with an unknown number of degrees of freedom p (with $p = \infty$ corresponding to the normal distribution), which is treated as a random quantity with a prior distribution. Inference would then be based on the posterior distribution of p. However, the lack of a natural conjugate prior for p means that this is not entirely straightforward. We therefore use an alternative method; the following analysis is the result of in-depth discussions with Mark Mathieson.

We assumed the data has arisen from a mixture of normals or a mixture of t_4s, and use the prior distributions

$$p(t_4) = p(\text{normal}) = 0.5 \tag{4.33}$$

$$p(k \mid t_4) = p(k \mid \text{normal}) \propto 1/k! \quad (k = 1, \ldots, k_{max} = 100) \tag{4.34}$$

with (conditional on the type of mixture and the number of components) the fixed-κ prior given in Section 4.3.1. We then used Algorithm 4.2 to fit a mixture of t_4 distributions to the data and estimate the posterior distribution of k

$$p(k \mid t_4, x^n). \tag{4.35}$$

We performed five separate runs of $20\,000$ iterations, the first $10\,000$ iterations of each run being discarded as burn-in, and found the mean and standard error of the resulting estimates, which are shown in Table 4.1. We obtained estimates for

$$p(k \mid \text{normal}, x^n) \tag{4.36}$$

in a similar way, and the results are also shown in Table 4.1.

By Bayes theorem we have

$$p(k \mid t_4, x^n) = \frac{p(k, t_4 \mid x^n)}{p(t_4 \mid x^n)} \quad \text{for all } k \tag{4.37}$$

and so

$$p(t_4 \mid x^n) = \frac{p(k, t_4 \mid x^n)}{p(k \mid t_4, x^n)} = \frac{p(x^n \mid k, t_4)p(k, t_4)}{p(k \mid t_4, x^n)p(x^n)} \quad \text{for all } k, \tag{4.38}$$

and similarly

$$p(\text{normal} \mid x^n) = \frac{p(x^n \mid k, \text{normal})p(k, \text{normal})}{p(k \mid \text{normal}, x^n)p(x^n)} \quad \text{for all } k. \tag{4.39}$$

$k =$	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 4.1: Estimates of the posterior probabilities $p(k \mid t_4, x^n)$ and $p(k \mid \text{normal}, x^n)$. These are the means of the estimates from five separate runs of Algorithm 4.2, 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 4.2: Estimates of the posterior probabilities $p(t_4, k \mid x^n)$ and $p(\text{normal}, k \mid x^n)$. These were obtained from the estimates of $p(x^n \mid k = 3, t_4)$ and $p(x^n \mid k = 3, \text{normal})$ given by Mathieson (1997) and the results shown in Table 4.1, as described in the text.

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 (which he refers to as THM) of obtaining estimates for $p(x^n \mid k, t_4)$ and $p(x^n \mid k, \text{normal})$ using importance sampling, and uses this method to obtain the estimates

$$-\log p(x^n \mid k = 3, t_4) \approx 227.64 \tag{4.40}$$

$$-\log p(x^n \mid k = 3, \text{normal}) \approx 229.08 \tag{4.41}$$

giving (using equations (4.38) and (4.39))

$$p(t_4 \mid x^n) \approx 0.916 \tag{4.42}$$

$$p(\text{normal} \mid x^n) \approx 0.084 \tag{4.43}$$

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 4.2. We conclude that for the prior distributions used, mixtures of t_4 distributions are heavily favoured over mixtures of normal distributions, with $k = 4$ t_4 components having the highest posterior probability.

4.5 Example 2: Old Faithful data

For our second example, we return to the first view of the Old Faithful data we considered in Chapter 2 (Section 2.4.3), which consists of 272 observations in two dimensions, with two moderately separated groups (Figure 2.14). We used

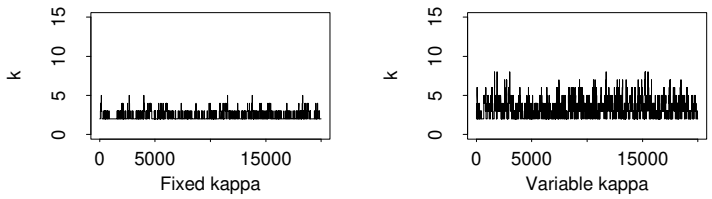
Algorithm 4.2 to fit a mixture of an unknown number of bivariate normal distributions to the data, performing separate runs for $\lambda = 1$ and $\lambda = 3$ with both the fixed-κ and variable-κ priors (detailed in Section 4.3.1).

Each run consisted of $20\,000$ iterations of Algorithm 4.2, 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 run for fixed-κ with $\lambda = 3$ took about 13 minutes. Figure 4.8 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 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. Thus the mixing over k appears to be poorer than for the galaxy data. This is presumably partly due to the tighter posterior distribution on k for this data, and partly 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 $p(k \mid x^n)$.

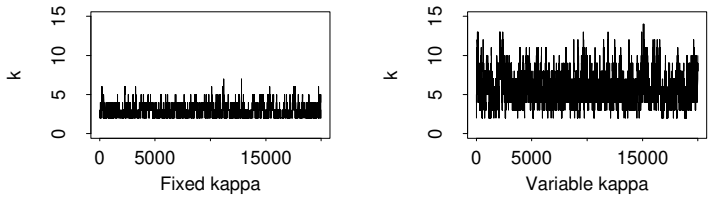
Figure 4.3 shows graphs against T ($T = 1, \ldots, 20\,000$) of $\widehat{p}_T(k \leq j)$ (equation (4.32)) for increasing values of j, and we see that these estimates stabilise fairly quickly. Figure 4.10 shows histograms representing estimates of the posterior distribution of k for the different runs, based on the sample with the first $10\,000$ iterations discarded as burn-in. Once again the posterior distribution for k depends heavily on the prior used. However, the predictive density estimates shown in Figure 4.11 appear less sensitive to changes in the prior.

4.6 Example 3: Old Faithful revisited

Recall that successive durations for the Old Faithful data are not independent, and that a scatter plot of the duration of the tth eruption against the duration of the $(t + 1)$th eruption (Figure 2.16, Section 2.4.4) gives an impression of four or more groups. As before, we ignored the time-series structure of the data, and used Algorithm 4.2 to fit a mixture of an unknown number of bivariate normal distributions to the data, performing separate runs for $\lambda = 1$ and $\lambda = 3$ with both the fixed-κ and variable-κ priors (detailed in Section 4.3.1). The sampler was run for $20\,000$ iterations from a random starting point, and appeared to mix well both between k and within k. The run for fixed-κ with $\lambda = 3$ took about 15 minutes. The first $10\,000$ iterations of each run were discarded as burn-in, and estimates (4.6) of the posterior distribution of k for the different runs are shown in Figure 4.12. Once again the posterior distribution for k depends heavily on the prior used, while the predictive density estimates shown in Figure 4.13 appear less sensitive to changes in the prior.

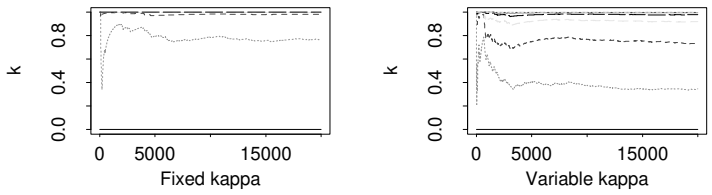


(a) $\lambda = 1$

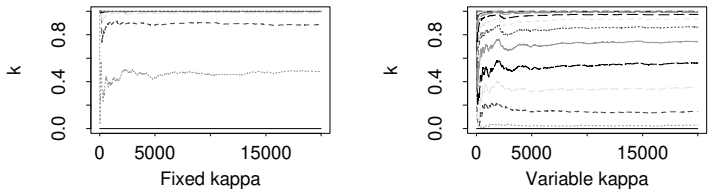


(b) $\lambda = 3$

Figure 4.8: Values of k sampled using Algorithm 4.2 when fitting a normal mixture to the first view of the Old Faithful data (Example 2) using $\lambda = 1, 3$. The two columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior.

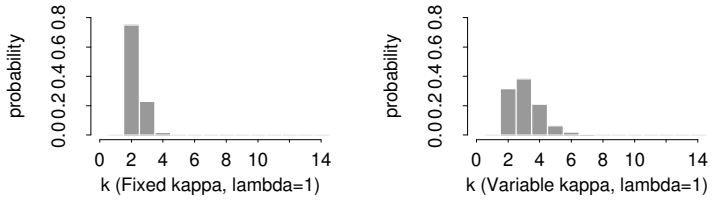


(a) $\lambda = 1$

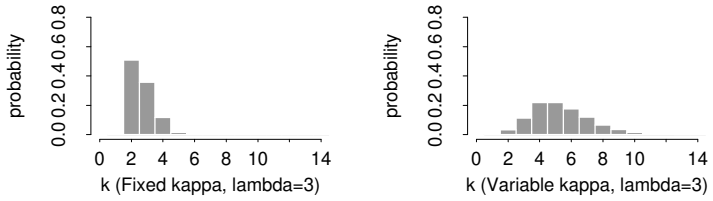


(b) $\lambda = 3$

Figure 4.9: Graphs for the first view of the Old Faithful data of $\widehat{p}_T(k \leq j)$ (given by equation (4.32)) against T, for increasing values of j. These are based on the sampled values of k shown in Figure 4.8, and obtained using Algorithm 4.2 when fitting mixtures of normal distributions to the data using $\lambda = 1, 3$. The columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior. The gap between the lines for j and $j - 1$ gives an estimate for $\Pr(k = j)$ based on the first T samples, and these estimates can be seen to stabilise fairly quickly.

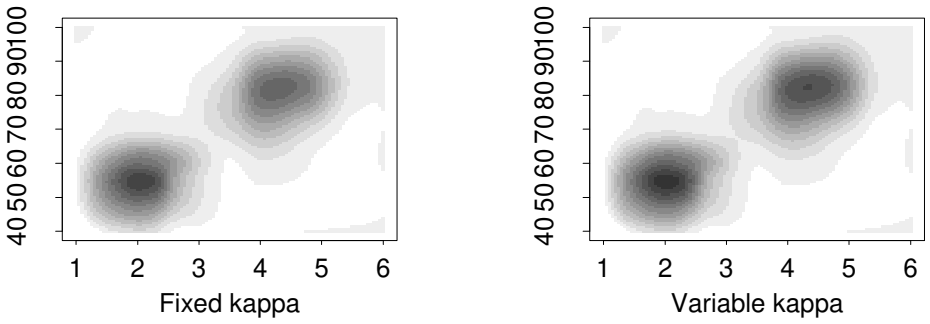


(a) $\lambda = 1$

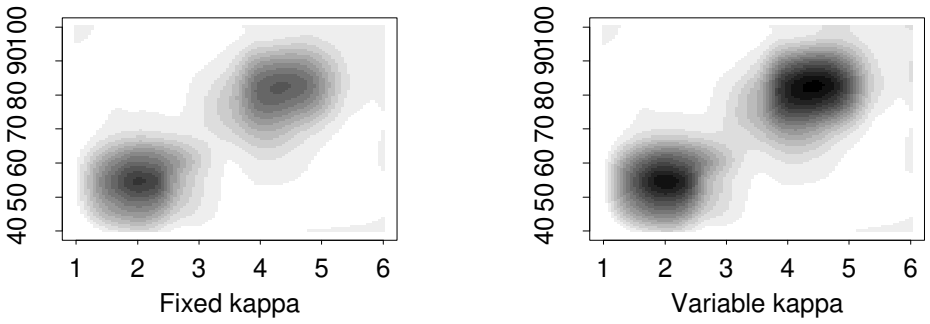


(b) $\lambda = 3$

Figure 4.10: Graphs showing estimates (4.6) of $\Pr(k = i)$ for $i = 1, 2, \ldots$, for the Old Faithful data. These estimates are based on the values of k sampled using Algorithm 4.2 when fitting normal distributions to the data, using $\lambda = 1, 3$. The two columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior. The posterior distribution of k can be seen to depend on the prior distribution for k (value of λ), and the prior distribution for $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ (variable-κ or fixed-κ).

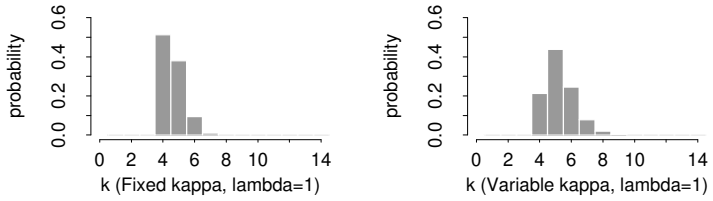


(a) $\lambda = 1$

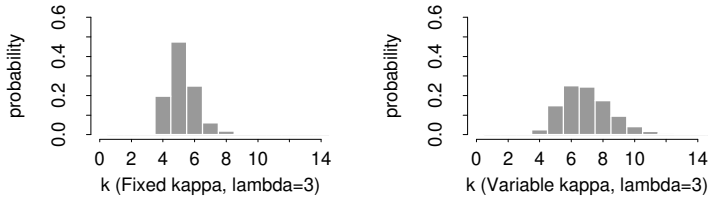


(b) $\lambda = 3$

Figure 4.11: Predictive density estimates (4.7) for the Old Faithful data. These are based on the output of Algorithm 4.2 when fitting normal distributions to the data using $\lambda = 1, 3$. The two columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior. Dark shading corresponds to regions of high density, and the figures are all shaded on the same scale. The density estimates appear to be less sensitive than the posterior distribution of k to choice of prior.

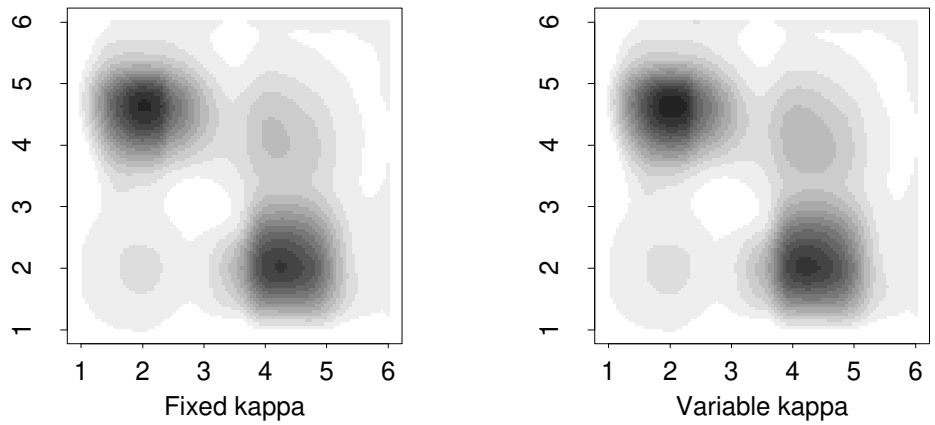


(a) $\lambda = 1$

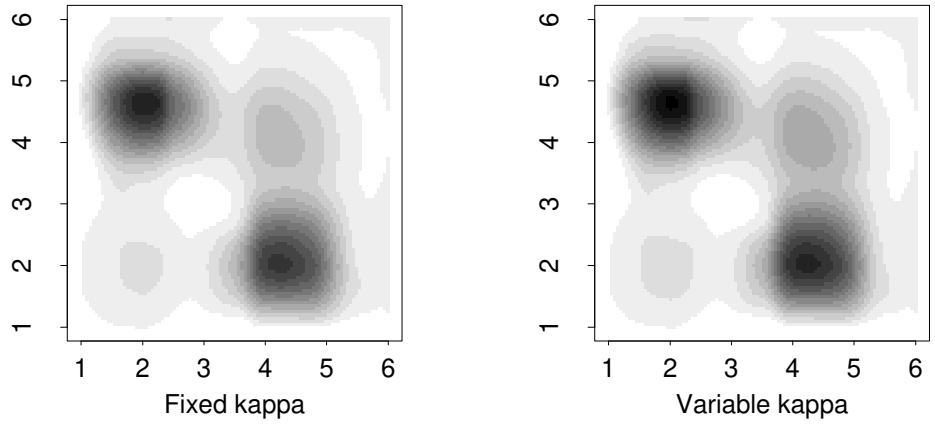


(b) $\lambda = 3$

Figure 4.12: Graphs showing estimates (4.6) of $\Pr(k = i)$ for $i = 1, 2, \ldots$, for the Old Faithful data revisited. These estimates are based on the values of k sampled using Algorithm 4.2 when fitting normal distributions to the data, using $\lambda = 1, 3$. The two columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior. The posterior distribution of k can be seen to depend on the prior distribution for k (value of λ), and the prior distribution for $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ (variable-κ or fixed-κ).

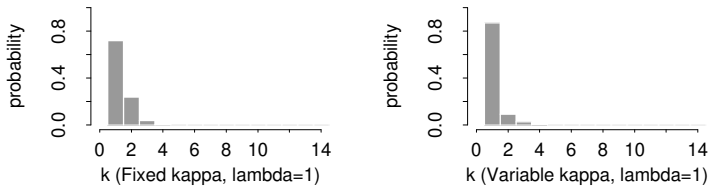


(a) $\lambda = 1$

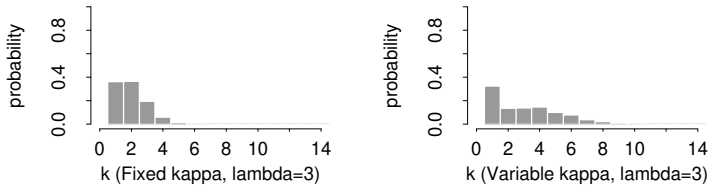


(b) $\lambda = 3$

Figure 4.13: Predictive density estimates (4.7) for the Old Faithful data revisited. These are based on the output of Algorithm 4.2 when fitting normal distributions to the data using $\lambda = 1, 3$. The two columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior. Dark shading corresponds to regions of high density, and the figures are all shaded on the same scale. The density estimates appear to be less sensitive than the posterior distribution of k to choice of prior.



(a) $\lambda = 1$



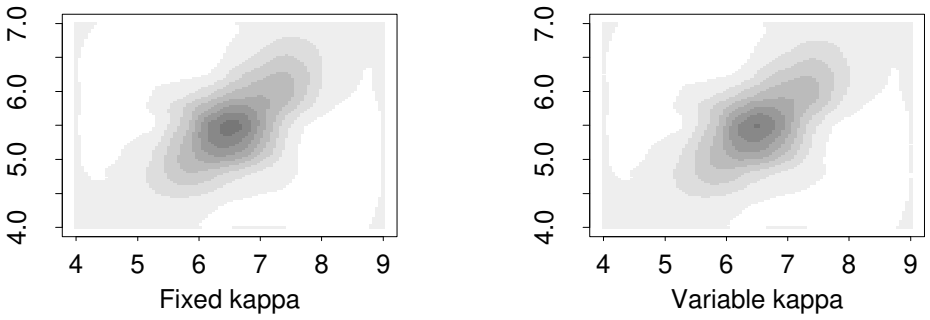
(b) $\lambda = 3$

Figure 4.14: Graphs showing estimates (4.6) of $\Pr(k = i)$ for $i = 1, 2, \ldots$, for the *Iris Virginica* data. These estimates are based on the values of k sampled using Algorithm 4.2 when fitting normal distributions to the data, using $\lambda = 1, 3$. The two columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior. The mode of these estimates 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.

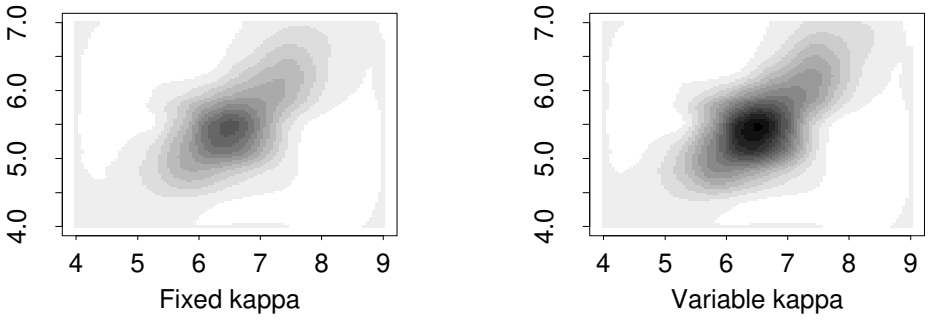
4.7 Example 4: *Iris Virginica* data

For our fourth example we return to the *Iris Virginica* data considered in the previous chapter (Section 3.4.3). We recall that there have been suggestions that this data supports the division of the *virginica* species into subspecies (Wilson, 1982), and we wish to investigate this claim by fitting a mixture of an unknown number of bivariate normal distributions to the 50 observations of sepal length and petal length (shown in Figure 3.10), using Bayesian methods.

Our analysis was performed with $\lambda = 1, 3$ and with both fixed-κ and variable-κ priors (see Section 4.3.1). We applied Algorithm 4.2 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 4.14) 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 for this data are shown in Figure 4.15.



(a) $\lambda = 1$



(b) $\lambda = 3$

Figure 4.15: Predictive density estimates (4.7) for the *Iris Virginica* data. These are based on the output of Algorithm 4.2 when fitting normal distributions to the data using $\lambda = 1, 3$. The two columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior. Dark shading corresponds to regions of high density, and the figures are all shaded on the same scale.

4.8 Example 5: Pima data

For our penultimate example we consider a dataset discussed by Ripley (1996) which consists of the readings of 9 variables for 768 women of Pima Indian heritage, living near Phoenix, Arizona. These women were tested for diabetes according to World Health Organisation criteria, and we consider only those who tested positive. There is a suggestion of multimodality in the diabetic group (see for example Ripley, 1996, page 99) and in order to investigate this further we attempted to fit a mixture of bivariate normal distributions to two of the measured variables: *plasma glucose concentration* and *diastolic blood pressure* (in mm Hg). We discarded records which had one or both of these variables missing, leaving 250 records for analysis (see Section A.4 in the appendix to this thesis). A scatter plot of the data is shown in Figure 4.16.

As before, we performed four separate analyses, using $\lambda = 1, 3$ with both fixed-κ and variable-κ priors (see section 4.3.1). In each case we used Algorithm 4.2 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 4% ($\lambda = 1$) and 12% ($\lambda = 3$) for the fixed-κ prior, and 27% ($\lambda = 1$) and 59% ($\lambda = 3$) for the variable-κ prior. Traces of the sampled value of k for $\lambda = 3$ with the variable-κ prior (not shown) showed a trend to fitting increasing values of k throughout the run, indicating that the chain had not converged and would have to be run for longer (or started from a more carefully chosen start point) if accurate inference was required. Estimates for the posterior distribution of k (Figure 4.17) show that the data supports models with more that one component, with the two component model being heavily favoured by the posteriors corresponding to the fixed-κ priors. With the variable-κ priors the posterior for k depends heavily on the prior used for k (that is, the value of λ chosen) which indicates that for this choice of prior on $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the data is not very informative for k.

In order to obtain a sensible clustering of the observations into two groups we considered the results of our run with $\lambda = 3$ and the fixed-κ prior conditioned on $k = 2$. We applied Algorithm 3.2 from the previous chapter to obtain a labelled sample which was suitable for clustering inference, and the clustering obtained by choosing allocation variables to maximise the estimated predictive classification probabilities (3.5) is shown in Figure 4.18.

The full predictive density estimates for this data are shown in Figure 4.19, and the estimate corresponding to $\lambda = 3$ with the variable-κ prior can be seen to be less smooth than the others, due to the large number of components being fitted to the data in this case.

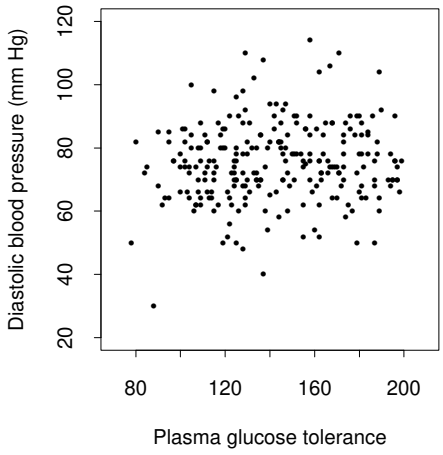
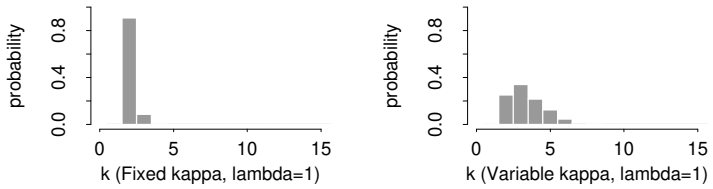
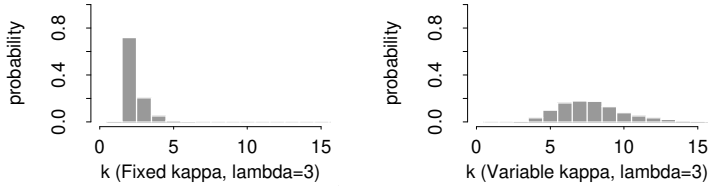


Figure 4.16: Scatter plot of the Pima data.



(a) $\lambda = 1$



(b) $\lambda = 3$

Figure 4.17: Graphs showing estimates (4.6) of $\Pr(k = i)$ for $i = 1, 2, \ldots$, for the Pima data. These estimates are based on the values of k sampled using Algorithm 4.2 when fitting normal distributions to the data, using $\lambda = 1, 3$. The two columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior. These estimates support a model with two or more components. The estimates for the variable-κ prior are highly dependent on the value of λ chosen.

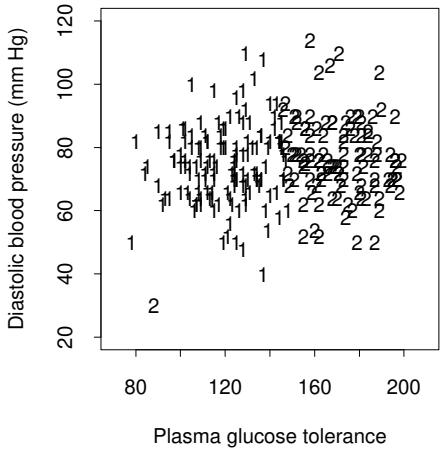
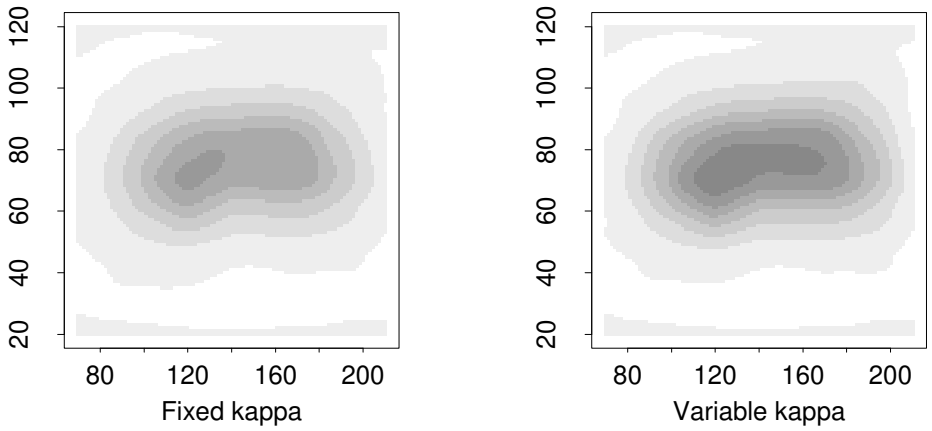
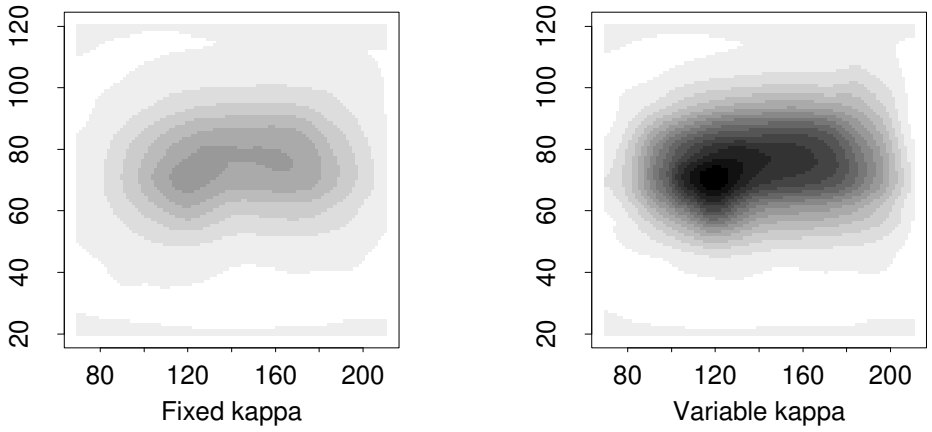


Figure 4.18: Clustering based on fitting 2 normal distributions to the Pima data. The clustering was obtained by conditioning the results for $\lambda = 3$ with fixed-κ on $k = 2$, and permuting the resulting sample using Algorithm 3.3. The allocation variables were then chosen to maximise the estimated scaled predictive component densities based on the permuted sample (equation (3.5)), which is equivalent to maximising the estimated predictive classification probabilities (see Section 1.1.4, Chapter 1).



(a) $\lambda = 1$



(b) $\lambda = 3$

Figure 4.19: Predictive density estimates (4.7) for the Pima data. These are based on the output of Algorithm 4.2 when fitting normal distributions to the data using $\lambda = 1, 3$. The two columns show results for **Left:** Fixed-κ prior; **Right:** Variable-κ prior. Dark shading corresponds to regions of high density, and the figures are all shaded on the same scale.

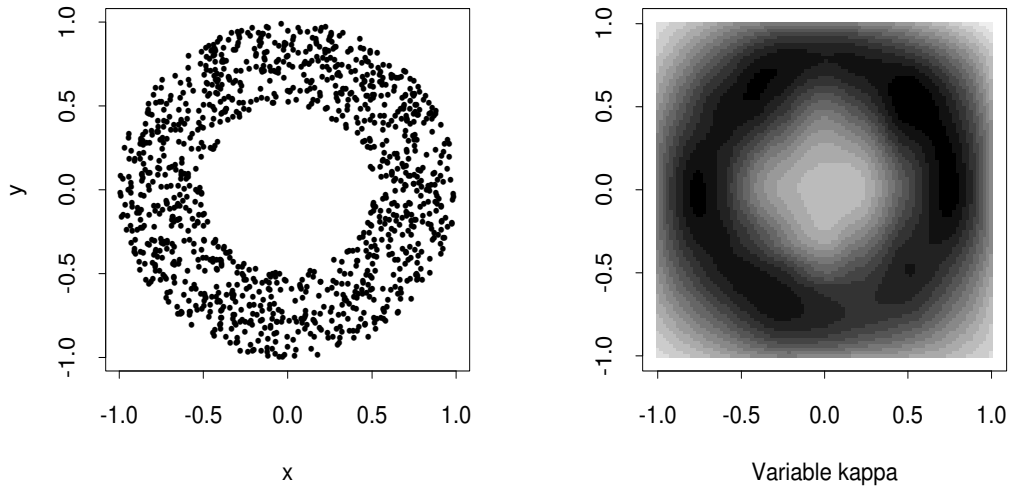


Figure 4.20: **Left:** Scatter plot showing the Simulated Ring data. **Right:** Predictive density estimate for Simulated Ring data, obtained by fitting normal distributions to the data, using $\lambda = 3$ and the variable-κ prior. The posterior distribution for the number of components k had most of its support between 20 and 30.

4.9 Example 6: Simulated Ring data

As a final example we present the results of a density estimation problem for data which is not well modelled by a mixture of a reasonable number of normal components. We simulated a uniform "ring" of data, by simulating 2000 random points uniformly in the square of side-length 2 about the origin, and then removing all those points within a radius of 0.5 and outside a radius of 1. A scatter plot of the remaining 1183 data points is shown in Figure 4.20. Algorithm 4.2 was run for $10\,000$ iterations, with the variable-κ prior, and $\lambda = 3$. The first 1000 iterations were discarded as burn-in (the run was made rather short due to the computational complexity; we did not concern ourselves greatly with convergence for this example) and a predictive density estimate formed with the remaining 9000 sample points is shown in 4.20. Some of the circular structure has been preserved by fitting large numbers of normal components (up to 30), but the sharp edges of the boundary have been considerably smoothed out.

4.10 Discussion — Effect of prior on posterior for k

We have seen in our examples how the posterior distribution for the number of components k in a mixture 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 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 heavily favours a one component model for reasonable amounts of data, as the following proposition makes precise. (This point is also made by Jennison, 1997).

Proposition 4.3. *For the fixed-κ distribution described in Section 4.3.1 with $\boldsymbol{\pi} \sim \mathcal{D}(\delta, \ldots, \delta)$,*

$$\lim_{\kappa \to 0} \frac{p(x^n \mid k)}{p(x^n \mid k = 1)} = k \frac{\Gamma(k\delta)\Gamma(n + \delta)}{\Gamma(\delta)\Gamma(n + k\delta)}. \tag{4.44}$$

In particular, for $\delta = 1$,

$$\lim_{\kappa \to 0} \frac{p(x^n \mid k)}{p(x^n \mid k = 1)} = \frac{k!n!}{(n + k - 1)!} \tag{4.45}$$

Proof. Proof is deferred to the appendix to this chapter (Section 4.11.2). □

We note that a corollary of Proposition 4.3 is that for $\delta = 1$, the Bayes factor for comparing the two component model with the one component model tends to $2/(n + 1)$, and so a one component model is heavily favoured provided there is a reasonable amount of data. 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.

Further progress might also be made by seeking priors which express the idea that the components of the mixture are different enough for discrimination to be a reasonable aim, and thus avoid fitting several similar components where one will suffice. 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. Groups might be created by clustering together similar components in the mixture by post-processing the output of the MCMC sampler. Alternatively, a more principled approach would be to use a prior structure which expresses the idea that components representing aspects of the same group are similar when compared with components representing

other groups. For example, the 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.

Finally, we close with the remark that despite the manifest problems with specifying suitable priors, our examples show that it is possible to obtain *some* useful results with Bayesian methods for mixture models in a clustering context. In particular, the fixed-κ prior with $\lambda = 1$ appears to give a sensible posterior distribution for k in all our examples (provided we use t_4 distributions to model the galaxy data.)

4.11 Appendix

4.11.1 Proof of Proposition 4.1

Proof. 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. For notational convenience we drop the dependence on (ω, η) for $p(k \,|\, \omega, \eta)$ and $\widetilde{p}(\phi \,|\, \omega, \eta)$. The prior distribution is then of the form

$$p(k, \boldsymbol{\pi}, \boldsymbol{\phi}) = p(k) p(\boldsymbol{\pi}, \boldsymbol{\phi} \,|\, k) \tag{4.46}$$

where $p(\boldsymbol{\pi}, \boldsymbol{\phi} \,|\, k)$ has density

$$p(\boldsymbol{\pi}, \boldsymbol{\phi} \,|\, k) = (k-1)! I(\pi_1 + \cdots + \pi_{k-1} \leq 1) \widetilde{p}(\phi_1) \ldots \widetilde{p}(\phi_k) \tag{4.47}$$

when considered as a distribution on $[0,1]^{k-1} \times \Phi^k$.

Thus for $(\boldsymbol{\pi}, \boldsymbol{\phi}) \in \Lambda_k$ we have

$$\begin{aligned} dP_k\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} = p(k)(k-1)! I(\pi_1 + \cdots + \pi_{k-1} \leq 1) \\ \widetilde{p}(\phi_1) \ldots \widetilde{p}(\phi_k)\, d\pi_1 \ldots d\pi_{k-1}\, d\phi_1 \ldots\, d\phi_k. \end{aligned} \tag{4.48}$$

Also, by Bayes theorem we have

$$d\widetilde{P}\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} \propto L([\boldsymbol{\pi}, \boldsymbol{\phi}]) dP\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} \tag{4.49}$$

and so we will write

$$d\widetilde{P}\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} = f([\boldsymbol{\pi}, \boldsymbol{\phi}]) dP\{(\boldsymbol{\pi}, \boldsymbol{\phi})\} \tag{4.50}$$

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:

$$\int_{\Omega_k} g(y)\, d\mu_k(y) = \int_{\Lambda_k} g([\boldsymbol{\pi}, \boldsymbol{\phi}])\, dP_k\big\{(\boldsymbol{\pi}, \boldsymbol{\phi})\big\} \tag{4.51}$$

and

$$\begin{aligned}\int_{\Omega_k} g(y)\, d\widetilde{\mu}(y) &= \int_{\Lambda_k} g([\boldsymbol{\pi}, \boldsymbol{\phi}])\, d\widetilde{P}_k\big\{(\boldsymbol{\pi}, \boldsymbol{\phi})\big\} \\ &= \int g([\boldsymbol{\pi}, \boldsymbol{\phi}]) f([\boldsymbol{\pi}, \boldsymbol{\phi}])\, dP_k\big\{(\boldsymbol{\pi}, \boldsymbol{\phi})\big\} \\ &= \int_{\Omega_k} g(y) f(y)\, d\mu_k(y). \end{aligned} \tag{4.52}$$

We define births on Λ by

$$(\boldsymbol{\pi}, \boldsymbol{\phi}) \cup (\pi, \phi) := \Big(\big(\pi_1(1-\pi), \phi_1\big), \ldots, \big(\pi_k(1-\pi), \phi_k\big), \big(\pi, \phi\big) \Big) \tag{4.53}$$

and will require the following Lemma:

Lemma 4.4. *If* $(\boldsymbol{\pi}, \boldsymbol{\phi}) \in \Lambda_k$ *and* $(\pi, \phi) \in [0,1] \times \Phi$ *then*

$$dP_{k+1}\big\{(\boldsymbol{\pi}, \boldsymbol{\phi}) \cup (\pi, \phi)\big\} = \tfrac{p(k+1)}{p(k)} k (1-\pi)^{k-1} \widetilde{p}(\phi)\, d\pi\, d\phi\, dP_k\big\{(\boldsymbol{\pi}, \boldsymbol{\phi})\big\}$$

Proof.

$$\begin{aligned}
LHS &= dP_{k+1}\big\{(\boldsymbol{\pi}, \boldsymbol{\phi}) \cup (\pi, \phi)\big\} \\
&= dP_{k+1}\Big\{ \big((\pi_1(1-\pi), \phi_1), \ldots, (\pi_k(1-\pi), \phi_k), (\pi, \phi)\big) \Big\} \quad \text{[equation (4.53)]} \\
&= dP_{k+1}\Big\{ \big((\pi, \phi), (\pi_1(1-\pi), \phi_1), \ldots, (\pi_k(1-\pi), \phi_k)\big) \Big\} \quad \text{[symmetry]} \\
&= p(k+1)\widetilde{p}(\phi)\widetilde{p}(\phi_1)\ldots\widetilde{p}(\phi_k) k! I(\pi + \pi_1(1-\pi) + \cdots + \pi_{k-1}(1-\pi) \le 1) \\
&\qquad d\phi\, d\phi_1 \ldots d\phi_k\, d(\pi_1(1-\pi)) \ldots\, d(\pi_{k-1}(1-\pi))\, d\pi \quad \text{[equation (4.48)]} \\
&= p(k+1)\widetilde{p}(\phi)\widetilde{p}(\phi_1)\ldots\widetilde{p}(\phi_k) k(k-1)! I(\pi_1 + \cdots + \pi_{k-1} \le 1) \\
&\qquad d\phi\, d\phi_1 \ldots d\phi_k\, (1-\pi)^{k-1} d\pi_1 \ldots\, d\pi_{k-1}\, d\pi \quad \text{[changing variables in integral]} \\
&= \tfrac{p(k+1)}{p(k)} k (1-\pi)^{k-1} \widetilde{p}(\phi)\, d\pi\, d\phi\, dP_k\big\{(\boldsymbol{\pi}, \boldsymbol{\phi})\big\} \quad \text{[equation (4.48)]} \\
&= RHS.
\end{aligned}$$

□

Preston (1976) shows that for a process to possess stationary distribution $\widetilde{\mu}$ it is sufficient that the following detailed balance conditions hold:

Definition 2 (Detailed Balance Conditions). $\widetilde{\mu}$ is said to satisfy detailed balance conditions if

$$\int_F \beta(y)\, d\widetilde{\mu}_k(y) = \int_{\Omega_{k+1}} \delta(z) K_\delta^{(k+1)}(z; F)\, d\widetilde{\mu}_{k+1}(z) \quad \text{for } k \geq 0,\ F \subset \Omega_k \tag{4.54}$$

and

$$\int_G \delta(z)\, d\widetilde{\mu}_{k+1}(z) = \int_{\Omega_k} \beta(y) K_\beta^{(k)}(y; G)\, d\widetilde{\mu}_k(y) \quad \text{for } k \geq 0,\ G \subset \Omega_{k+1}. \tag{4.55}$$

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 $\widetilde{\mu}$ satisfies the detailed balance conditions for our process. 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 as follows:

$$\begin{aligned}
LHS &= \int_F \beta(y)\, d\widetilde{\mu}_k(y) \\
&= \int_{\Omega_k} I(y \in F)\beta(y) f(y)\, d\mu_k(y) \quad \text{[equation (4.52)]} \\
&= \int_{\Omega_k} I(y \in F) \int_{[0,1]} \int_\Phi b\big(y; (\pi, \phi)\big) k (1-\pi)^{k-1} \widetilde{p}(\phi) d\pi\, d\phi\, f(y)\, d\mu_k(y) \\
&\qquad\qquad \text{[equation (4.17)]} \\
&= \int_{\Omega_k} \int_{[0,1]} \int_\Phi I(y \in F) b\big(y; (\pi, \phi)\big) f(y) k (1-\pi)^{k-1} \widetilde{p}(\phi)\, d\pi\, d\phi\, d\mu_k(y) \\
&\qquad\qquad \text{[rearranging.]} \\
RHS &= \int_{\Omega_{k+1}} \delta(z) K_\delta^{(k+1)}(z; F)\, d\widetilde{\mu}_{k+1}(z) \\
&= \int_{\Omega_{k+1}} \delta(z) K_\delta^{(k+1)}(z; F)\, f(z)\, d\mu_{k+1}(z) \quad \text{[equation (4.52)]} \\
&= \int_{\Omega_{k+1}} \sum_{(\pi,\phi) \in z : z \backslash (\pi,\phi) \in F} d\big(z \backslash (\pi, \phi); (\pi, \phi)\big)\, f(z)\, d\mu_{k+1}(z) \quad \text{[equation (4.18)]} \\
&= \int_{\Lambda_{k+1}} \sum_{i=1}^{k+1} I([\boldsymbol{\pi}, \boldsymbol{\phi}] \backslash (\pi_i, \phi_i) \in F)\, d\big([\boldsymbol{\pi}, \boldsymbol{\phi}] \backslash (\pi_i, \phi_i); (\pi_i, \phi_i)\big) \cdot \\
&\qquad \cdot f([\boldsymbol{\pi}, \boldsymbol{\phi}])\, dP_{k+1}\big\{(\boldsymbol{\pi}, \boldsymbol{\phi})\big\} \quad \text{[equation (4.51)]} \\
&= \int_{\Lambda_{k+1}} (k+1) I([\boldsymbol{\pi}, \boldsymbol{\phi}] \backslash (\pi_{k+1}, \phi_{k+1}) \in F)\, d\big([\boldsymbol{\pi}, \boldsymbol{\phi}] \backslash (\pi_{k+1}, \phi_{k+1}); (\pi_{k+1}, \phi_{k+1})\big) \cdot \\
&\qquad \cdot f([\boldsymbol{\pi}, \boldsymbol{\phi}])\, dP_{k+1}\big\{(\boldsymbol{\pi}, \boldsymbol{\phi})\big\} \quad \text{[by symmetry of } P_{k+1}(\cdot)\text{]}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\Lambda_{k+1}} (k+1) I\big([\boldsymbol{\pi}', \boldsymbol{\phi}'] \in F\big)\, d\big([\boldsymbol{\pi}', \boldsymbol{\phi}']; (\pi_{k+1}, \phi_{k+1})\big)\, f\big([\boldsymbol{\pi}', \boldsymbol{\phi}'] \cup (\pi_{k+1}, \phi_{k+1})\big) \cdot \\
&\qquad \cdot\, dP_{k+1}\big\{(\boldsymbol{\pi}', \boldsymbol{\phi}') \cup (\pi_{k+1}, \phi_{k+1})\big\} \quad [(\boldsymbol{\pi}', \boldsymbol{\phi}') \cup (\pi_{k+1}, \phi_{k+1}) = (\boldsymbol{\pi}, \boldsymbol{\phi})] \\
&= \int_{\Lambda_{k+1}} (k+1) I\big([\boldsymbol{\pi}', \boldsymbol{\phi}'] \in F\big)\, d\big([\boldsymbol{\pi}', \boldsymbol{\phi}']; (\pi, \phi)\big) f\big([\boldsymbol{\pi}', \boldsymbol{\phi}'] \cup (\pi, \phi)\big) \cdot \\
&\qquad \cdot\, dP_{k+1}\big\{(\boldsymbol{\pi}', \boldsymbol{\phi}') \cup (\pi, \phi)\big\} \quad [\text{writing } (\pi, \phi) \text{ for } (\pi_{k+1}, \phi_{k+1})] \\
&= \int_{\Lambda_k} \int_{[0,1]} \int_{\Phi} I\big([\boldsymbol{\pi}', \boldsymbol{\phi}'] \in F\big)(k+1)\, d\big([\boldsymbol{\pi}', \boldsymbol{\phi}']; (\pi, \phi)\big)\, f\big([\boldsymbol{\pi}', \boldsymbol{\phi}'] \cup (\pi, \phi)\big) \cdot \\
&\qquad \cdot\, \tfrac{p(k+1)}{p(k)} k (1-\pi)^{k-1} \widetilde{p}(\phi)\, d\pi\, d\phi\, dP_k\big\{(\boldsymbol{\pi}', \boldsymbol{\phi}')\big\} \quad [\text{Lemma 4.4}] \\
&= \int_{\Omega_k} \int_{[0,1]} \int_{\Phi} I(y \in F)(k+1)\, d\big(y; (\pi, \phi)\big)\, f\big(y \cup (\pi, \phi)\big) \tfrac{p(k+1)}{p(k)} \cdot \\
&\qquad \cdot\, k (1-\pi)^{k-1} \widetilde{p}(\phi)\, d\pi\, d\phi\, d\mu_k(y) \quad [\text{equation (4.51)}]
\end{aligned}
$$

and so $LHS = RHS$ provided

$$(k+1) d\big(y; (\pi, \phi)\big) f\big(y \cup (\pi, \phi)\big) \tfrac{p(k+1)}{p(k)} = b\big(y; (\pi, \phi)\big) f(y)$$

which is equivalent to the conditions stated in the proposition as $f(y) \propto L(y)$.

The remaining detailed balance conditions can be shown to hold in a similar way. □

4.11.2 Proof of Proposition 4.3

The proof of Proposition 4.3 is given as a sequence of Lemmas and Corollaries.

Lemma 4.5.

$$\int \pi_1^{n_1} \dots \pi_k^{n_k} \mathcal{D}(\boldsymbol{\pi}; \delta, \dots, \delta)\, d\pi_1 \dots\, d\pi_{k-1} = \frac{\Gamma(k\delta)\Gamma(n_1+\delta)\dots\Gamma(n_k+\delta)}{\Gamma(\delta)^k \Gamma(\sum_i n_i + k\delta)}$$

where $\pi_k = 1 - \pi_1 + \dots + \pi_{k-1}$ *and the integral is over* $\pi_1 + \dots + \pi_{k-1} \le 1$.

Proof.

$$
\begin{aligned}
LHS &= \int \pi_1^{n_1} \dots \pi_k^{n_k} \mathcal{D}(\boldsymbol{\pi}; \delta, \dots, \delta)\, d\pi_1 \dots\, d\pi_{k-1} \\
&= \int \pi_1^{n_1} \dots \pi_k^{n_k} \pi_1^{\delta-1} \dots \pi_k^{\delta-1} \frac{\Gamma(k\delta)}{\Gamma(\delta)^k}\, d\pi_1 \dots\, d\pi_{k-1} \\
&= \frac{\Gamma(k\delta)}{\Gamma(\delta)^k} \int \pi_1^{n_1+\delta-1} \dots \pi_k^{n_k+\delta-1}\, d\pi_1 \dots d\pi_{k-1} \\
&= \frac{\Gamma(k\delta)\Gamma(n_1+\delta)\dots\Gamma(n_k+\delta)}{\Gamma(\delta)^k \Gamma(\sum_i n_i + k\delta)} \\
&= RHS.
\end{aligned}
$$

□

Lemma 4.6.

$$\int \exp\Big\{-\frac{v}{2}\sum_{j=1}^{n}(x_j-\mu)^2\Big\}\mathcal{N}(\mu;\xi,\kappa^{-1})\,d\mu = \sqrt{\frac{\kappa}{nv+\kappa}}\exp\Big\{-\frac{v}{2}\Big(SS+\frac{n\kappa(\xi-\bar{x})^2}{nv+\kappa}\Big)\Big\}$$

where

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

and

$$SS = \sum_{j=1}^{n}(x_j-\bar{x})^2.$$

Proof.

$$\begin{aligned}
LHS &= \int \exp\Big\{-\frac{v}{2}\sum_{j=1}^{n}(x_j-\mu)^2\Big\}\mathcal{N}(\mu;\xi,\kappa^{-1})\,d\mu \\
&= \int \sqrt{\frac{\kappa}{2\pi}}\exp\Big\{-\frac{v}{2}(SS+n(\bar{x}-\mu)^2)\Big\}\exp\Big\{-\frac{\kappa}{2}(\mu-\xi)^2\Big\}\,d\mu \\
&= \sqrt{\frac{\kappa}{2\pi}}\exp\Big\{-\frac{v}{2}SS\Big\}\int \exp\Big\{-\frac{nv}{2}(\bar{x}-\mu)^2-\frac{\kappa}{2}(\mu-\xi)^2\Big\}\,d\mu \\
&= \sqrt{\frac{\kappa}{2\pi}}\exp\Big\{-\frac{v}{2}SS\Big\}\int \exp\Big\{-\frac{nv+\kappa}{2}\Big(\mu-\frac{nv\bar{x}+\xi\kappa}{nv+\kappa}\Big)^2-\frac{nv\kappa}{2}\frac{(\xi-\bar{x})^2}{nv+\kappa}\Big\}\,d\mu \\
&= \sqrt{\frac{\kappa}{nv+\kappa}}\exp\Big\{-\frac{v}{2}\Big(SS+\frac{n\kappa(\xi-\bar{x})^2}{nv+\kappa}\Big)\Big\} \\
&= RHS.
\end{aligned}$$

□

Corollary 4.7.

$$\int \exp\Big\{-\frac{v}{2}\sum_{j=1}^{n}(x_j-\mu)^2\Big\}\mathcal{N}(\mu;\xi,\kappa^{-1})\,d\mu \begin{cases} = 1 & (n=0), \\ \le \sqrt{\frac{\kappa}{v}}\exp\Big\{-\frac{v}{2}SS\Big\} & (n\ge 1), \end{cases}$$

where $SS \ge 0$*, and* $SS > 0$ *if* $n > 1$ *(assuming the* x_j *are distinct).*

Proof. Follows directly from Lemma 4.6. □

Lemma 4.8.

$$I^*(\kappa;\beta) \le \begin{cases} 1 & (n=0), \\ \text{const}\sqrt{\kappa} & (n=1), \\ \text{const}\sqrt{\kappa}\beta^\alpha & (n>1), \end{cases}$$

where

$$I^*(\kappa;\beta) = \int \Gamma(v;\alpha,\beta)\Big(\frac{v}{2\pi}\Big)^{\frac{n}{2}} \int \exp\Big\{-\frac{v}{2}\sum_{j=1}^{n}(x_j-\mu)^2\Big\}\mathcal{N}\big(\mu;\xi,\kappa^{-1}\big)\,d\mu\,dv$$

and const *is a constant which does not depend on* κ *or* β.

Proof. Case $n=0$ is trivial. Otherwise

$$\begin{aligned} I^*(\kappa;\beta) &= \int \Gamma(v;\alpha,\beta)\Big(\frac{v}{2\pi}\Big)^{\frac{n}{2}} \int \exp\Big\{-\frac{v}{2}\sum_{j=1}^{n}(x_j-\mu)^2\Big\}\mathcal{N}\big(\mu;\xi,\kappa^{-1}\big)\,d\mu\,dv \\ &\le \int \frac{v^{\alpha-1}\beta^\alpha e^{-\beta v}}{\Gamma(\alpha)}\Big(\frac{v}{2\pi}\Big)^{\frac{n}{2}}\sqrt{\frac{\kappa}{v}}\exp\Big\{-\frac{v}{2}SS\Big\}\,dv \quad \text{[Corollary 4.7]} \\ &= \text{const}\sqrt{\kappa}\beta^\alpha \int v^{\alpha-1+\frac{n-1}{2}}\exp\big[-v\big(\beta+SS/2\big)\big]\,dv \\ &= \text{const}\sqrt{\kappa}\beta^\alpha \frac{\Gamma(\alpha+\frac{n-1}{2})}{\big(\beta+SS/2\big)^{\alpha+\frac{n-1}{2}}} \\ &= \text{const}\frac{\sqrt{\kappa}\beta^\alpha}{\big(\beta+SS/2\big)^{\alpha+\frac{n-1}{2}}}. \end{aligned}$$

If $n=1$ then $SS=0$, and

$$I^*(\kappa;\beta) \le \text{const}\sqrt{\kappa}.$$

If $n=2$ then $SS>0$, and

$$I^*(\kappa;\beta) \le \text{const}\frac{\sqrt{\kappa}\beta^\alpha}{\big(SS/2\big)^{\alpha+\frac{n-1}{2}}} = \text{const}\sqrt{\kappa}\beta^\alpha.$$

□

Lemma 4.9.

$$\begin{aligned} \int \Gamma(\beta;g,h)\prod_{i=1}^{k}\Bigg[\int \Gamma(v_i;\alpha,\beta)\Big(\frac{v_i}{2\pi}\Big)^{\frac{n_i}{2}} \cdot \\ \cdot \int \exp\Big\{-\frac{v_i}{2}\sum_{j:z_j=i}(x_j-\mu_i)^2\Big\}\mathcal{N}\big(\mu_i;\xi,\kappa^{-1}\big)\,d\mu_i\,dv_i\Bigg]\,d\beta \\ \le \text{const}\sqrt{\kappa}^{\#\{n_i>0\}} \end{aligned}$$

where $n_i = \#\{j : z_j = i\}$ *and* const *is a constant which does not depend on* κ.

Proof.

$$\begin{aligned}
LHS &\le \int \Gamma(\beta; g, h) \prod_{i:n_i=0} 1 \prod_{i:n_i=1} \left[\text{const}_i \sqrt{\kappa}\right] \prod_{i:n_i>1} \left[\text{const}_i \sqrt{\kappa}\beta^{\alpha}\right] d\beta \quad \text{[Lemma 4.8]} \\
&= \text{const} \sqrt{\kappa}^{\#\{n_i>0\}} \int \Gamma(\beta; g, h)(\beta^{\alpha})^{\#\{n_i>1\}} \, d\beta \\
&= \text{const} \sqrt{\kappa}^{\#\{n_i>0\}} \\
&= RHS.
\end{aligned}$$

□

Lemma 4.10.

$$p(x^n, z^n \mid k) \le \text{const} \sqrt{\kappa}^{\#\{n_i>0\}}$$

where $n_i = \#\{j : z_j = i\}$ *and* const *is a constant which does not depend on* κ.

Proof.

$$\begin{aligned}
p(x^n, z^n \mid k) &= \int p(x^n, z^n \mid \theta, k) p(\theta \mid k) \, d\theta \\
&= \int \prod_{j=1}^{n} \Big[\pi_{z_j} \mathcal{N}\big(x_j; \mu_{z_j}, v_{z_j}^{-1}\big)\Big] \mathcal{D}(\boldsymbol{\pi}; \delta, \ldots, \delta) \cdot \\
&\qquad \cdot \prod_{i=1}^{k} \Big[\mathcal{N}\big(\mu_i; \xi, \kappa^{-1}\big) \Gamma(v_i; \alpha, \beta)\Big] \Gamma(\beta; g, h) \, d\boldsymbol{\pi} \, d\boldsymbol{v} \, d\boldsymbol{\mu} \, d\beta \\
&= \int \pi_1^{n_1} \ldots \pi_k^{n_k} \mathcal{D}(\boldsymbol{\pi}; \delta, \ldots, \delta) \, d\pi_1 \ldots d\pi_{k-1} \cdot \\
&\qquad \cdot \int \Gamma(\beta; g, h) \prod_{i=1}^{k} \Bigg[\int \Gamma(v_i; \alpha, \beta) \Big(\frac{v_i}{2\pi}\Big)^{\frac{n_i}{2}} \cdot \\
&\qquad \cdot \int \exp\Big\{-\frac{v_i}{2} \sum_{j:z_j=i} (x_j - \mu_i)^2\Big\} \mathcal{N}\big(\mu_i; \xi, \kappa^{-1}\big) \, d\mu_i \, dv_i \Bigg] \, d\beta \\
&\le \text{const} \sqrt{\kappa}^{\#\{n_i>0\}} \quad \text{[Lemmas 4.5 and 4.9].}
\end{aligned}$$

□

Lemma 4.11. *If* $I^*(\kappa; \beta)$ *is as defined in Lemma 4.8 then*

$$\lim_{\kappa \to 0} \frac{I^*(\kappa; \beta)}{\sqrt{\kappa}} = \int \Gamma(v; \alpha, \beta) \Big(\frac{v}{2\pi}\Big)^{\frac{n}{2}} \sqrt{\frac{1}{nv}} \exp\Big\{-\frac{vSS}{2}\Big\} \, dv.$$

Proof. By definition of $I^*(\kappa;\beta)$ and Lemma 4.6 we have

$$\lim_{\kappa\to 0}\frac{I^*(\kappa;\beta)}{\sqrt{\kappa}}$$
$$= \lim_{\kappa\to 0}\int \Gamma(v;\alpha,\beta)\Big(\frac{v}{2\pi}\Big)^{\frac{n}{2}}\sqrt{\frac{1}{nv+\kappa}}\exp\Big\{-\frac{v}{2}\Big(SS+\frac{n\kappa(\xi-\bar{x})^2}{nv+\kappa}\Big)\Big\}\,dv.$$

The integrand on the RHS is bounded above by

$$\Gamma(v;\alpha,\beta)\Big(\frac{v}{2\pi}\Big)^{\frac{n}{2}}\sqrt{\frac{1}{v}}\exp\Big\{-\frac{vSS}{2}\Big\}$$

which is integrable by straightforward integration for $n > 0$. The integrand also clearly possesses a limit as $\kappa \to 0$, and so by the dominated convergence theorem (see for example Billingsley, 1986, page 213) the limit of the integral is equal to the integral of the limit, which gives the stated result. □

Lemma 4.12.

The limit $\lim_{\kappa\to 0}\dfrac{p(x^n \mid k=1)}{\sqrt{\kappa}}$ *exists and is non-zero and finite.*

Proof. We note that

$$\frac{p(x^n \mid k=1)}{\sqrt{\kappa}} = \int \frac{I^*(\kappa;\beta)}{\sqrt{\kappa}}\Gamma(\beta;g,h)\,d\beta$$

where $I^*(\kappa;\beta)$ is as defined in Lemma 4.8. By Lemma 4.8

$$\frac{I^*(\kappa;\beta)}{\sqrt{\kappa}}\Gamma(\beta;g,h) \le \begin{cases}\text{const}\,\Gamma(\beta;g,h) & (n=1),\\ \text{const}\,\beta^{\alpha}\Gamma(\beta;g,h) & (n>1),\end{cases} \tag{4.56}$$

which are both integrable (by straightforward integration). By Lemma 4.11

$$\frac{I^*(\kappa;\beta)}{\sqrt{\kappa}}\Gamma(\beta;g,h)$$

possesses a finite limit as $\kappa \to 0$. Thus we have

$$\begin{aligned}\lim_{\kappa\to 0}\frac{p(x^n \mid k=1)}{\sqrt{\kappa}} &= \lim_{\kappa\to 0}\int \frac{I^*(\kappa;\beta)}{\sqrt{\kappa}}\Gamma(\beta;g,h)\,d\beta\\ &= \int \lim_{\kappa\to 0}\frac{I^*(\kappa;\beta)}{\sqrt{\kappa}}\Gamma(\beta;g,h)\,d\beta \quad \text{[dominated convergence]}\end{aligned}$$

which is finite (by equation (4.56) above, and straightforward integration) and non-zero since the integrand is strictly positive for all β. □

Lemma 4.13.

$$\lim_{\kappa\to 0}\frac{p(x^n,z^n\,|\,k)}{p(x^n\,|\,k=1)}=\begin{cases}0 & (\#\{n_i>0\}>1),\\ \frac{\Gamma(k\delta)\Gamma(n+\delta)}{\Gamma(\delta)\Gamma(n+k\delta)} & (\#\{n_i>0\}=1).\end{cases}$$

Proof. If $\#\{n_i>0\}>1$ the result follows from Lemma 4.10 and Lemma 4.12.

If $\#\{n_i>0\}=1$ then z^n allocates all observations to the same component, c say, and

$$\begin{aligned}p(x^n,z^n\,|\,k)&=\int p(x^n,z^n\,|\,\theta,k)p(\theta\,|\,k)\,d\theta\\&=\int\prod_{j=1}^{n}\Big[\pi_c\mathcal{N}\big(x_j;\mu_c,v_c^{-1}\big)\Big]\mathcal{D}(\boldsymbol{\pi};\delta,\ldots,\delta)\cdot\\&\qquad\cdot\prod_{i=1}^{k}\Big[\mathcal{N}\big(\mu_i;\xi,\kappa^{-1}\big)\Gamma(v_i;\alpha,\beta)\Big]\Gamma(\beta;g,h)\,d\boldsymbol{\pi}\,d\boldsymbol{v}\,d\boldsymbol{\mu}\,d\beta\\&=p(x^n\,|\,k=1)\int\pi_c^n\mathcal{D}(\boldsymbol{\pi};\delta,\ldots,\delta)\,d\boldsymbol{\pi}\\&=p(x^n\,|\,k=1)\frac{\Gamma(k\delta)\Gamma(n+\delta)}{\Gamma(\delta)\Gamma(n+k\delta)}\quad\text{[by Lemma 4.5]}\end{aligned}$$

and so

$$\lim_{\kappa\to 0}\frac{p(x^n,z^n\,|\,k)}{p(x^n\,|\,k=1)}=\frac{\Gamma(k\delta)\Gamma(n+\delta)}{\Gamma(\delta)\Gamma(n+k\delta)}.$$

□

Proposition 4.3 then follows as a Corollary:

Corollary 4.14.

$$\lim_{\kappa\to 0}\frac{p(x^n\,|\,k)}{p(x^n\,|\,k=1)}=k\frac{\Gamma(k\delta)\Gamma(n+\delta)}{\Gamma(\delta)\Gamma(n+k\delta)}$$

Proof.

$$\begin{aligned}LHS&=\lim\frac{p(x^n\,|\,k)}{p(x^n\,|\,k=1)}\\&=\lim\sum_{z^n}\frac{p(x^n,z^n\,|\,k)}{p(x^n\,|\,k=1)}\\&=\sum_{z^n}\lim\frac{p(x^n,z^n\,|\,k)}{p(x^n\,|\,k=1)}\quad\text{[since these limits are finite]}\\&=\sum_{z^n:\#\{n_i>0\}=1}\lim\frac{p(x^n,z^n\,|\,k)}{p(x^n\,|\,k=1)}+\sum_{z^n:\#\{n_i>0\}>1}\lim\frac{p(x^n,z^n\,|\,k)}{p(x^n\,|\,k=1)}\\&=k\frac{\Gamma(k\delta)\Gamma(n+\delta)}{\Gamma(\delta)\Gamma(n+k\delta)}\quad\text{[Lemma 4.13].}\end{aligned}$$

□

Chapter 5

Sequential methods for mixture models

In this chapter we compare some methods of analysing mixture distributions in a Bayesian manner when the observations $x_1, x_2, \ldots$ from the model arrive sequentially; each observation is processed when it arrives, and then discarded before the next observation arrives. This situation is often referred to as *on-line learning* in the engineering literature, and contrasts with the *batch* learning methods we have studied in previous chapters, which require all the data to be available at the same time.

According to Huo and Lee (1997)

> "The advantage of a sequential algorithm over a batch algorithm is not necessarily in the final result, but in computational efficiency, reduced storage requirements, and the fact that an outcome may be provided without having to wait for all the data to be processed."

Sequential algorithms are often used in an *adaptive* context, where we have information about the mixture model parameters from a sample which comes from a similar (but different) source to that which is providing the current observations, and we wish to adapt the model to more accurately represent the current source. For example, a speech recognition system might have been previously trained to recognize words from one speaker, and we may wish to adapt the system to recognize words from another speaker. See for example Huo *et al.* (1995, 1996); Huo and Lee (1997).

Early work on sequential methods for mixture distributions was done in the field of signal processing, at a time when computers were significantly less powerful than they are today. As a result, many of the methods were rather simplistic and performed rather poorly even in very simple situations where perhaps only one parameter of the mixture was unknown. For details of the history see Titterington *et al.* (1985). A significant improvement in performance for simple problems (where the densities of the mixture model are assumed to be known, and only the mixture proportions $\boldsymbol{\pi}$ are unknown) came with the work of Makov and

Smith (1977) who introduced the *quasi-Bayes (QB)* method which is still in use today (Huo *et al.*, 1995). However, since then little progress has been made, despite a large increase in computer power. In searching for a method which would take advantage of this increased computer power we developed a method which we will refer to as the *Kullback–Leibler* (KL) method for reasons which will become apparent. This method was developed independently of Bernardo and Girón (1988), who claim that its performance is superior to QB. In this chapter we examine the performance of QB and KL on some simulated data, and find that this claim is justified. We also introduce some simple extensions to these methods and apply them to the more difficult case of observations from a mixture of univariate normal distributions where the mixture proportions, means and variances are all unknown. In such cases the asymptotic behaviour of these methods (as the number of observations $\to \infty$) is unclear, but we found we could obtain some useful information from samples of size 1000 and 2000.

5.1 Some simple sequential approximation methods

Assume that $x_1, x_2, \ldots$ are sequentially arriving independent observations from a mixture of k components (k assumed known and finite) with density

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = \pi_1 f(x; \phi_1, \eta) + \cdots + \pi_k f(x; \phi_k, \eta). \tag{5.1}$$

Let θ denote the unknown parameters of the model, and assume that we have a prior distribution for θ

$$p(\theta) = g(\theta; a^{(0)}) \tag{5.2}$$

where $g(\cdot; a)$ is the conjugate family of prior distributions, parameterised by a, for the mixture family f. By this we mean, as in Section 1.1.5, that the posterior distribution of θ given the completed data (x^n, z^n) will also be a distribution from the family $g(\cdot; a)$, and in particular we have

$$p(\theta \mid X_1 = x_1, Z_1 = i) = g(\theta; a(i)) \quad \text{for some } a(i). \tag{5.3}$$

Then the posterior distribution for θ given a single observation x_1 from the model will be

$$\begin{aligned} p(\theta \mid X_1 = x_1) &= \sum_{i=1}^{k} \Pr(Z_1 = i \mid X_1 = x_1) p(\theta \mid X_1 = x_1, Z_1 = i) \\ &= \sum_{i=1}^{k} p(i) g(\theta; a(i)) \quad \text{say,} \end{aligned} \tag{5.4}$$

where both $p(i)$ and $a(i)$ may depend on x_1 and $a^{(0)}$. In the situations we consider $p(i)$ and $a(i)$ will be straightforward to calculate.

Suppose now we are able approximate (5.4) by choosing $a^{(1)}$ such that

$$\sum_{i=1}^{k} p(i)g(\theta; a(i)) \approx g(\theta; a^{(1)}). \tag{5.5}$$

Then starting with a prior $g(\theta; a^{(0)})$ and an observation $X_1 = x_1$ we have a method of obtaining an approximate posterior $g(\theta; a^{(1)})$. Let us view this as a method of obtaining $a^{(1)}$ from $a^{(0)}$ and x_1, and write

$$a^{(1)} = \text{SU}(a^{(0)}, x_1)$$

to define a *Sequential Update* function SU. On receiving a further observation $X_2 = x_2$ we can update $a^{(1)}$ to $a^{(2)}$ in exactly the same way:

$$a^{(2)} = \text{SU}(a^{(1)}, x_2)$$

and in general we have

$$a^{(n+1)} = \text{SU}(a^{(n)}, x_{n+1}) \tag{5.6}$$

giving *sequential approximations*

$$\widehat{p}(\theta \mid x^n) = g(\theta; a^{(n)}) \tag{5.7}$$

to the posterior distributions $p(\theta \mid x^n)$ $(n = 1, 2, \dots)$.

Some previously suggested methods for making the approximation (5.5) include

1. *Decision Directed* (DD): choose $a^{(1)}$ to be the $a(i)$ corresponding to the largest $p(i)$.

2. *Probabilistic Teacher* (PT): select $a^{(1)}$ at random from $a(1), \dots, a(k)$, choosing $a(i)$ with probability $p(i)$.

3. *quasi-Bayes* (QB): set $a^{(1)} = \sum_{i=1}^{k} p(i)a(i)$.

4. *Kullback–Leibler* (KL): choose $a^{(1)}$ to minimise the Kullback–Leibler divergence from $\sum_{i=1}^{k} p(i)g(\theta; a(i))$ to $g(\theta; a^{(1)})$

   $$KL\left[\sum_{i=1}^{k} p(i)g(\theta; a(i)) \,\|\, g(\theta; a^{(1)})\right]$$

   where the Kullback–Leibler divergence from a density $g(\cdot)$ to a density $f(\cdot)$ is defined by

   $$KL[g(\cdot) \,\|\, f(\cdot)] = \int g(\theta) \log(g(\theta)/f(\theta))\, d\theta. \tag{5.8}$$

DD corresponds intuitively to classifying each incoming observation to the class with highest probability, and then updating the density for θ assuming this classification is correct. PT is based on a similar idea, with the classification being done by simulating from the class probabilities. Both these approaches have been shown to be seriously deficient in their long-term behaviour, even in very simple problems, leading to estimates of the posterior distribution $p(\theta \mid x^n)$ which are wildly inconsistent (in that their modes do not tend to the "true" value of θ as $n \to \infty$). We therefore do not study these methods here, but refer the reader to Titterington *et al.* (1985) and references therein for further information.

Makov and Smith (1977) (see also Smith and Makov, 1978) introduced the QB method in the context of observations from mixture distributions in which the component densities are assumed known, and the only unknowns are the mixture proportions $\boldsymbol{\pi}$. They show that the QB method is consistent in this context, in that the mode of $\widehat{p}(\boldsymbol{\pi} \mid x^n)$ tends to the "true" value of $\boldsymbol{\pi}$ as $n \to \infty$ (assuming the model for the data is correct). A later paper (Smith and Makov, 1981) extends the method and convergence result to the case of a mixture of two normal distributions with known mixture proportions and variances, but unknown means.

The KL method was suggested by Bernardo and Girón (1988), also in the context where the densities of the mixture components are assumed known. They claim that its performance is superior to QB in that it leads to a more accurate approximation of $p(\boldsymbol{\pi} \mid x^n)$. Despite this, examples of its use are hard to find while QB still enjoys a certain popularity (see for example Huo *et al.*, 1995).

In this chapter we compare the performance of the QB and KL sequential methods on some simulated data, and show that in general the KL method provides a more accurate approximation to the required posterior distribution, at the expense of greater computational complexity. We consider the following cases:

Case 1: data $x_1, x_2, \ldots$ are assumed to be independent observations from a mixture with density

$$p(x \mid \boldsymbol{\pi}) = \pi_1 f_1(x) + \cdots + \pi_k f_k(x) \tag{5.9}$$

where the component densities $f_1, \ldots, f_k$ are assumed known.

Case 2: data $x_1, x_2, \ldots$ are assumed to be independent observations from a mixture of k univariate normal distributions with density

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \pi_1 \mathcal{N}(x; \mu_1, \sigma_1^2) + \cdots + \pi_k \mathcal{N}(x; \mu_k, \sigma_k^2) \tag{5.10}$$

where $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ are all considered *unknown*.

In the literature, examination of sequential methods has been confined almost entirely to Case 1, which may arise if the component densities can be accurately estimated from classified data which is not informative for the mixture proportions. Tråvén (1991) considers an on-line algorithm for Case 2 from a non-Bayesian perspective.

Case 1 is simpler than Case 2, not just because of the smaller number of parameters to be estimated, but because the likelihood (5.9) is not symmetric in $(\pi_1, \ldots, \pi_k)$ and so the posterior distribution of $\boldsymbol{\pi}$ does not possess the multimodality due to symmetry which we have discussed in previous chapters (Section 1.1.5 for example). In contrast, if we use a symmetric prior for the parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$ in Case 2 then their posterior distribution may be highly multimodal due to symmetry. Multimodal distributions are generally difficult to approximate accurately analytically, and so for simplicity we will assume in Case 2 the existence of a *training set* $\mathcal{T}$ of observations whose correct classifications are known. This will remove the symmetry in the posterior distribution $p(\theta \mid \mathcal{T}, x^n)$, reducing its potential multimodality and thus making it easier to accurately approximate analytically. In assuming the existence of a set of classified observations, we are restricting ourselves to the situation where the components of the mixture have a physical interpretation. Further work might investigate sequential methods for use where a training set $\mathcal{T}$ is not available.

Before looking at some simulated examples (sections 5.4 and 5.5) we introduce a natural extension of the simple sequential methods described here, and also see how a Gibbs sampler (similar to the one used in Chapter 2, Section 2.2) can be used to approximate $p(\theta \mid x^n)$ in a non-sequential way, providing a benchmark against which to compare the sequential methods.

5.2 A natural extension of the simple sequential methods

The sequential approximations described in the previous section approximate the posterior distribution $p(\theta \mid x^n)$ by a *single* distribution from the conjugate family:

$$\widehat{p}(\theta \mid x^n) = g(\theta; a^{(n)}). \tag{5.11}$$

Updating $\widehat{p}(\theta \mid x^n)$ to $\widehat{p}(\theta \mid x^{n+1})$ on receiving an observation x_{n+1} involves updating $a^{(n)}$ to $a^{(n+1)}$ using x_{n+1}

$$a^{(n+1)} = \mathrm{SU}(a^{(n)}, x_{n+1}). \tag{5.12}$$

A natural extension would be to approximate the posterior distribution by a *mixture* of R distributions from the conjugate family

$$\widehat{p}(\theta \mid x^n) = \sum_{r=1}^{R} p_r^{(n)} g(\theta; a_r^{(n)}) \tag{5.13}$$

for some fixed integer R. Updating $\widehat{p}(\theta \mid x^n)$ to $\widehat{p}(\theta \mid x^{n+1})$ on receiving an observation x_{n+1} then involves updating

$$(p_1^{(n)}, \ldots, p_r^{(n)}, a_1^{(n)}, \ldots, a_r^{(n)})$$

to

$$(p_1^{(n+1)}, \ldots, p_r^{(n+1)}, a_1^{(n+1)}, \ldots, a_r^{(n+1)})$$

using x_{n+1}.

A computationally simple method of performing this update is to use

$$p_r^{(n+1)} = p_r^{(n)} \tag{5.14}$$

$$a_r^{(n+1)} = \text{SU}(a_r^{(n)}, x_{n+1}) \quad (r = 1, \ldots, R). \tag{5.15}$$

We will refer to the procedure defined by this updating method as QB(R) or KL(R) according to the sequential update function SU employed. Although this method of performing the updates is rather unprincipled, we found that these modified procedures resulted in a significant improvement in performance when applying the methods to data from Case 2 (see Section 5.5). Brief investigation of more principled methods, which involved updating the $p_r^{(n)}$ at each step, did not detect any further improvement in performance in the examples we looked at.

5.3 Analytic approximation of $p(\theta \mid x^n)$ with the Gibbs sampler

The sequential methods described in previous sections give an analytic approximation $\widehat{p}(\theta \mid x^n)$ to the posterior distribution $p(\theta \mid x^n)$. An analytic approximation can also be obtained by batch methods using a Gibbs sampler to construct an irreducible Markov chain with stationary distribution $p(\theta, z^n \mid x^n)$. If $(\theta^{(t)}, (z^n)^{(t)})$ $(t = 1, \ldots, N)$ is a sampled path of such a chain then an appropriate estimate for $p(\theta \mid x^n)$ is given by the sample path average:

$$\begin{aligned} p(\theta \mid x^n) &= \sum_{z^n} p(\theta \mid z^n, x^n) p(z^n \mid x^n) \\ &\approx \frac{1}{N} \sum_{t=1}^{N} p(\theta \mid (z^n)^{(t)}, x^n) \\ &= \frac{1}{N} \sum_{t=1}^{N} g(\theta; a_t) \quad \text{say.} \end{aligned} \tag{5.16}$$

For an accurate approximation we wish N to be large. However, (5.16) then becomes difficult to work with, and for some tasks we may need to simplify this approximation, for example by choosing $\widetilde{p}_1, \ldots, \widetilde{p}_R$ and $\widetilde{a}_1, \ldots, \widetilde{a}_R$ such that

$$\frac{1}{N} \sum_{t=1}^{N} g(\theta; a_t) \approx \sum_{r=1}^{R} \widetilde{p}_r g(\theta; \widetilde{a}_r) \tag{5.17}$$

where R is "small", and the $\widetilde{p}_r$ $(r = 1, \ldots, R)$ are constrained to be non-negative and sum to unity.

We make the approximation (5.17) by grouping the component densities $g(\theta; a_t)$ $(t = 1, \ldots, N)$ into R groups, the rth group consisting of N_r "similar" densities $(r = 1, \ldots, R)$ using the following k-means type clustering algorithm:

Algorithm 5.1. Start with an initial grouping of the densities $g(\theta; a_t)$ $(t = 1, \ldots, N)$ into R groups, by putting the first N/R in the first group, the next N/R in the second group, and so on. Let w_t be the group of the tth density, and N_r be the number of densities in the rth group. Iterate the following steps until a fixed point is reached:

1. Choose $\widetilde{a}_r$ $(r = 1, \ldots, R)$ to minimise
$$KL\Big[\frac{1}{N_r}\sum_{t:w_t=r} g(\theta; a_t) \,\|\, g(\theta; \widetilde{a}_r)\Big].$$
2. Reallocate the N densities into R groups by choosing w_t $(t = 1, \ldots, N)$ to minimise
$$KL\Big[g(\theta; a_t) \,\|\, g(\theta; \widetilde{a}_{w_t})\Big].$$

When a fixed point is reached, approximate the posterior distribution of the parameters by

$$\widehat{p}(\theta \mid x^n) = \sum_{r=1}^{R} \frac{N_r}{N} g(\theta; \widetilde{a}_r). \tag{5.18}$$

We will use the notation GS(N, R) to denote this method of forming an approximation to $p(\theta \mid x^n)$.

5.3.1 Notes on Algorithm 5.1

The iterated steps of Algorithm 5.1 are guaranteed to reach a fixed point, as (by simple application of the definition of $KL[\cdot \,\|\, \cdot]$) each step increases

$$\sum_{t=1}^{N} \int g(\theta; a_t) \log g(\theta; \widetilde{a}_{w_t})\, d\theta = \sum_{r=1}^{R} \sum_{t:w_t=r} \int g(\theta; a_t) \log g(\theta; \widetilde{a}_r)\, d\theta$$

and there are only a finite number of possible values for $w_1, \ldots, w_n$.

Step 1 of the algorithm ensures that we have

$$\frac{1}{N_r}\sum_{t:w_t=r} g(\theta; a_t) \approx g(\theta; \widetilde{a}_r) \tag{5.19}$$

and so

$$\begin{aligned} p(\theta \mid x^n) &\approx \frac{1}{N}\sum_{t=1}^{N} g(\theta; a_t) \quad \text{[from (5.16)]} \\ &= \frac{1}{N}\sum_{r=1}^{R}\sum_{t:w_t=r} g(\theta; a_t) \\ &\approx \sum_{r=1}^{R} \frac{N_r}{N} g(\theta; \widetilde{a}_r) \quad \text{[from (5.19)]} \end{aligned} \tag{5.20}$$

which gives us the approximation (5.18).

We will make use of this approximation as a benchmark against which to compare the performance of the sequential methods in some of the simulated examples which follow.

5.4 Case 1: Component densities assumed known

We now compare the performance of the KL and QB sequential methods on some simulated data from a mixture with density

$$p(x \mid \boldsymbol{\pi}) = \pi_1 f_1(x) + \cdots + \pi_k f_k(x) \tag{5.21}$$

where the component densities $f_1, \ldots, f_k$ are assumed *known*. The form of the conjugate prior for $\boldsymbol{\pi}$ is

$$p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; \mathrm{d}) \tag{5.22}$$

and the sequential update steps for QB and KL are given in the appendix to this chapter (Section 5.6.4).

5.4.1 Mixtures of 2 univariate normal distributions

Our first examples assume that we have sequentially arriving observations from a mixture of two known distributions, mixed in unknown proportions. Since (π_1, π_2) are constrained by $\pi_1 + \pi_2 = 1$ there is effectively only one unknown π_1 in the model. In this particularly simple case the full Bayesian posterior distribution $p(\boldsymbol{\pi} \mid x^n)$ can actually be calculated for reasonable sample sizes n (see the appendix to this chapter, Section 5.6.1), and this provides a benchmark against which we can compare the performance of the sequential approximation methods QB and KL.

We simulated 1000 independent observations from the following densities, which have an increasing amount of overlap in the scaled component densities (Figure 5.1):

a) $p(x) = 0.3\mathcal{N}(x; 0, 1) + 0.7\mathcal{N}(x; 2, 1)$

b) $p(x) = 0.3\mathcal{N}(x; 0, 2) + 0.7\mathcal{N}(x; 2, 2)$

c) $p(x) = 0.3\mathcal{N}(x; 0, 4) + 0.7\mathcal{N}(x; 2, 4)$.

In all three cases we assumed a $\mathcal{D}(\boldsymbol{\pi}; 1, 1)$ prior on $\boldsymbol{\pi}$ (that is, a uniform prior on $[0, 1]$ for π_1).

We applied both the QB and KL sequential methods (the details of which are given in Section 5.6.4) to the 1000 sampled values from the models a)–c), and thus obtained approximations to the posterior distribution for π_1. As expected,

the QB method was faster than KL. For example, for the 1000 observations from model a) QB took 0.95 seconds while KL took 1.64 seconds[1].

Figure 5.2 compares the true posterior distribution for π_1 with the approximations obtained using the QB and KL sequential methods. It can be seen that the QB approximation is quite different from both the true posterior distribution and the KL approximation, which are indistinguishable by eye on this scale.

We note that a large amount of overlap in the scaled component densities makes it harder to guess which component any particular observation arose from, and so $p(\pi_1 \mid x^n)$ will tend to be more variable for observations x^n from c) than for observations from b) or a). It is clear from Figure 5.2 that the QB approximation is not able to take this into account, as the uncertainty in the QB approximation to the posterior is the same for a), b) and c).

5.4.2 A mixture of 4 univariate normal distributions

We now compare the performance of the QB and KL methods on a sample of 2000 independent observations simulated from a mixture of four known normal distributions

$$p(x) = 0.2\mathcal{N}(x; -1, 1) + 0.1\mathcal{N}(x; 0, 2) + 0.3\mathcal{N}(x; 1, 3) + 0.4\mathcal{N}(x; 1.5, 0.5) \tag{5.23}$$

whose scaled components are illustrated in Figure 5.3. We used a $\mathcal{D}(\pi; (1, 1, 1, 1))$ prior on $\boldsymbol{\pi}$. QB took 2.3 seconds and KL took 10.4 seconds for the 2000 observations.

In this case the full posterior distribution is not computationally tractable. We therefore use the GS$(10\,000, 1)$ method described in Section 5.3 as a benchmark against which to compare the sequential approximations. Figure 5.4 compares the marginal posterior densities of π_1, π_2, π_3 and π_4 obtained from the QB and KL methods with those obtained using the GS$(10\,000, 1)$ method. The known "true" parameter values are shown as vertical dotted lines. The QB method gives densities which are far too peaked, and the KL method seems to give densities which are slightly too peaked. Further experience with more problems of this kind has lead us to conclude that there is a general tendency for KL to produce densities which are overly peaked about slightly incorrect parameter values. However, its performance is certainly superior to QB.

[1]CPU times on a Sun UltraSparc 200 workstation

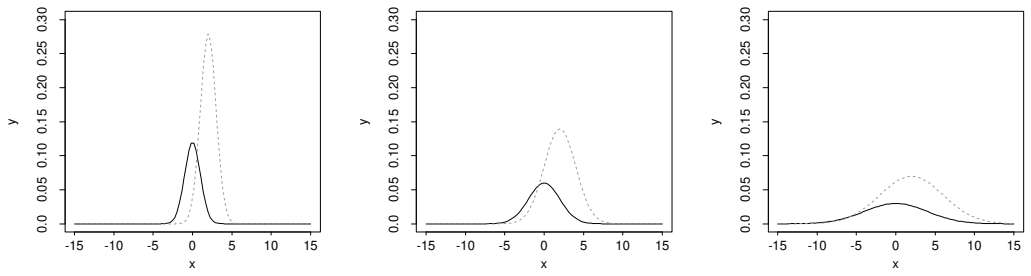


Figure 5.1: Plots showing the two scaled components of the mixtures used in Section 5.4.1. **Left:** a) $p(x \mid \pi_1) = 0.3\mathcal{N}(x; 0, 1) + 0.7\mathcal{N}(x; 2, 1)$; **Middle:** b) $p(x \mid \pi_1) = 0.3\mathcal{N}(x; 0, 2) + 0.7\mathcal{N}(x; 2, 2)$; **Right:** c) $p(x \mid \pi_1) = 0.3\mathcal{N}(x; 0, 4) + 0.7\mathcal{N}(x; 2, 4)$.

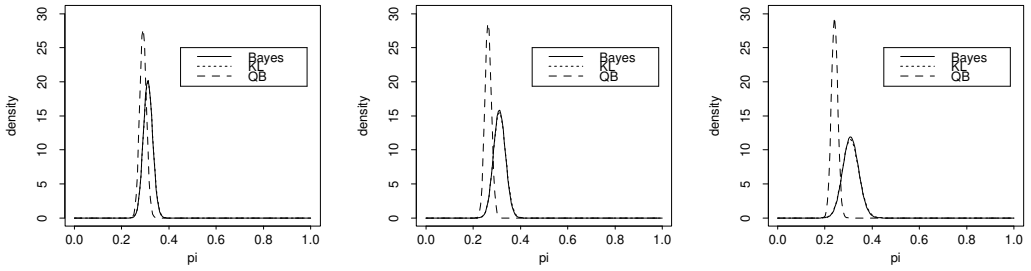


Figure 5.2: Comparison of approximations $\widehat{p}(\pi_1 \mid x^n)$ obtained using QB and KL sequential approximations with the true posterior distribution $p(\pi_1 \mid x^n)$. The KL approximation and the true posterior distribution are indistinguishable by eye on this scale. The three graphs show results for x^n being 1000 observations simulated from **Left:** a) $p(x) = 0.3\mathcal{N}(x; 0, 1) + 0.7\mathcal{N}(x; 2, 1)$; **Middle:** b) $p(x) = 0.3\mathcal{N}(x; 0, 2) + 0.7\mathcal{N}(x; 2, 2)$; **Right:** c) $p(x) = 0.3\mathcal{N}(x; 0, 4) + 0.7\mathcal{N}(x; 2, 4)$.

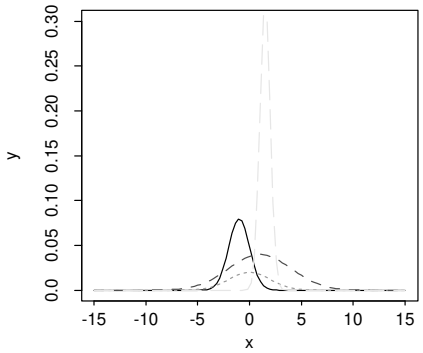
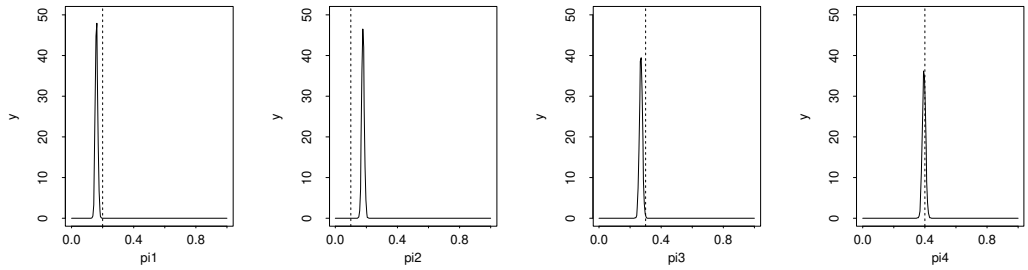
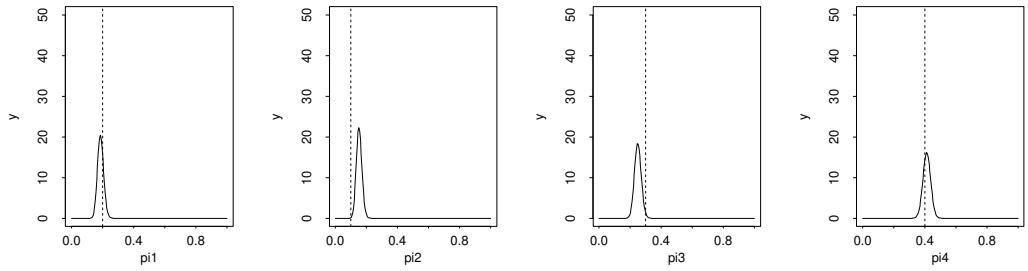


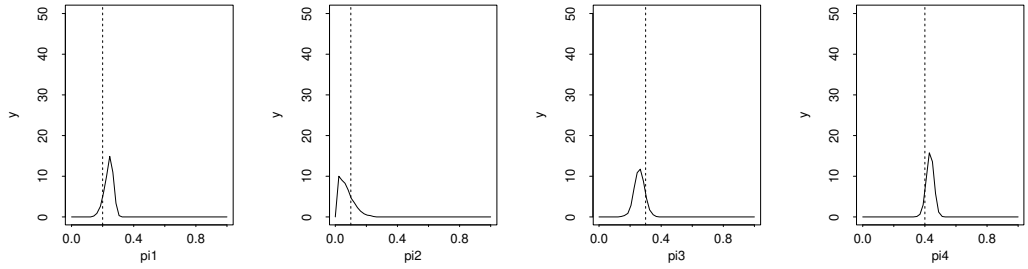
Figure 5.3: The four scaled components of the mixture (5.23).



(a) QB



(b) KL



(c) GS$(10\,000, 1)$

Figure 5.4: Estimates of the marginal posterior densities for (from **Left** to **Right**) π_1, π_2, π_3 and π_4, using the QB, KL, and GS$(10\,000, 1)$ approximations on data from (5.23). The true parameter values are indicated with a vertical dotted line

5.5 Case 2: Mixtures of univariate normal distributions

We now consider the case where $x_1, x_2, \ldots$ are assumed to be independent observations from a mixture of k univariate normal distributions with density

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \pi_1 \mathcal{N}(x; \mu_1, \sigma_1^2) + \cdots + \pi_k \mathcal{N}(x; \mu_k, \sigma_k^2) \tag{5.24}$$

where $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ are now all considered *unknown*. For notational convenience we will write v_i for σ_i^{-2} $(i = 1, \ldots, k)$. The form of the conjugate prior for $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{v})$ is then

$$p(\theta) = \mathcal{D}(\boldsymbol{\pi}; \mathbf{d}) \prod_{j=1}^{k} \Gamma(v_j; m_j/2, l_j/2) \mathcal{N}(\mu_j; u_j, 1/(s_j v_j)) \tag{5.25}$$

$$= g(\theta; \mathbf{d}, \mathbf{m}, \mathbf{l}, \mathbf{u}, \mathbf{s}) \text{ say.} \tag{5.26}$$

Details of the KL and QB sequential update steps for this case are given in the appendix to this chapter (Section 5.6.5).

Our examples consist of the following simulated data sets:

a) 1000 observations (10 of which form a training set of classified observations) from

$$p(x) = 0.3\mathcal{N}(x; 0, 1) + 0.7\mathcal{N}(x; 2, 1) \tag{5.27}$$

b) 2000 observations (30 of which form a training set of classified observations) from

$$p(x) = 0.2\mathcal{N}(x; -1, 1) + 0.1\mathcal{N}(x; 0, 2) + 0.3\mathcal{N}(x; 1, 3) + 0.4\mathcal{N}(x; 1.5, 0.5) \tag{5.28}$$

In each case we used a vague but proper prior for θ, namely $p(\theta) = g(\theta; \mathbf{d}, \mathbf{m}, \mathbf{l}, \mathbf{u}, \mathbf{s})$ with $d_j = 1, m_j = 0.1, l_j = 0.1, u_j = 0$, and $s_j = 0.01$ $(j = 1, \ldots, k)$. We then found the distribution of θ given the training set $\mathcal{T}$ of classified observations

$$p(\theta \mid \mathcal{T}) = g(\theta; \mathbf{d}^{(0)}, \mathbf{m}^{(0)}, \mathbf{l}^{(0)}, \mathbf{u}^{(0)}, \mathbf{s}^{(0)}) \tag{5.29}$$

using the conjugate analysis given in the appendix to this chapter (Section 5.6.2). Using (5.29) as our prior distribution, we obtained approximations to the posterior distribution $p(\theta \mid \mathcal{T}, x^n)$ by applying the following procedures to the remaining observations x^n:

1. The QB sequential method.

2. The KL sequential method.

3. The QB(5) sequential method. This was performed by applying GS$(10\,000, 5)$ to the first 10 unclassified observations to obtain an approximation

$$\widehat{p}(\theta \mid \mathcal{T}, x^{10}) = \sum_{r=1}^{5} p_r^{(10)} g(\theta; a_r^{(10)})$$

and then applying the QB(5) sequential update steps (described in Section 5.2) to the remaining unclassified observations, to obtain an approximation

$$\widehat{p}(\theta \mid \mathcal{T}, x^{n}) = \sum_{r=1}^{5} p_r^{(n)} g(\theta; a_r^{(n)}).$$

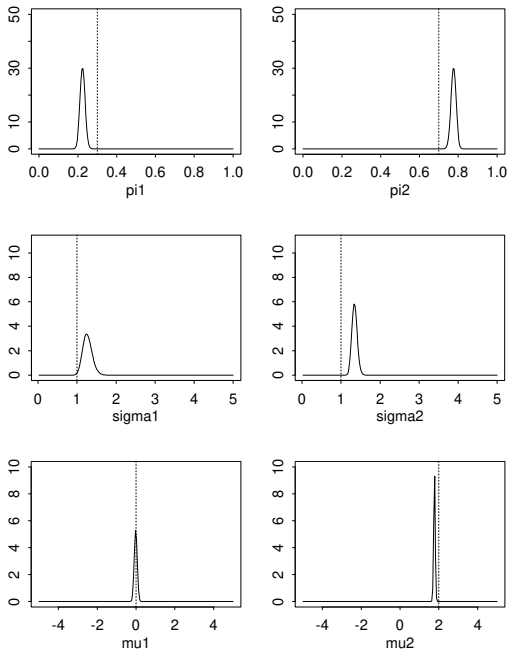
4. The KL(5) sequential method, performed as in 3 above, but replacing the QB(5) update steps with KL(5) update steps.

5. The GS$(10\,000, 1)$ batch method (see algorithm 5.1).

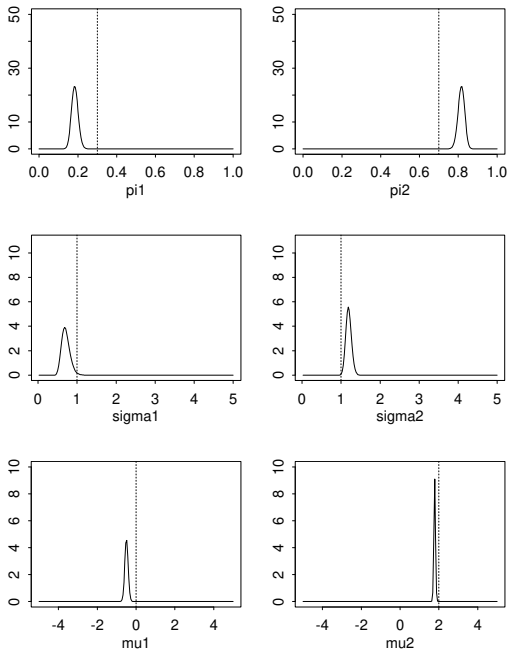
6. The GS$(10\,000, 5)$ batch method (see algorithm 5.1).

The full details of the Gibbs sampler used are given in the appendix to this chapter (Section 5.6.3); in each use of the Gibbs sampler the first 1000 iterations were discarded as burn-in.

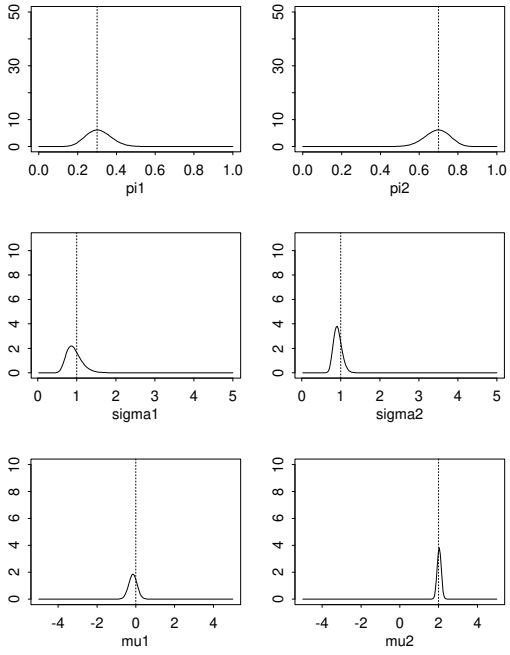
The approximations to the marginal posterior distributions for the individual components of $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ obtained using these six methods are shown in Figures 5.5 and 5.6 (for the sample from (5.27)) and Figures 5.7, 5.8 and 5.9 (for the sample from (5.28)). The batch methods GS$(10\,000, 1)$ and GS$(10\,000, 5)$ provide benchmarks against which the sequential methods may be judged. We see that QB tends to produce densities which are overly peaked about incorrect values, as does KL to a lesser extent. The modified procedures QB(5) and KL(5) are a definite improvement over QB and KL respectively, with KL being the better (though computationally more complex) of the two.



(a) QB

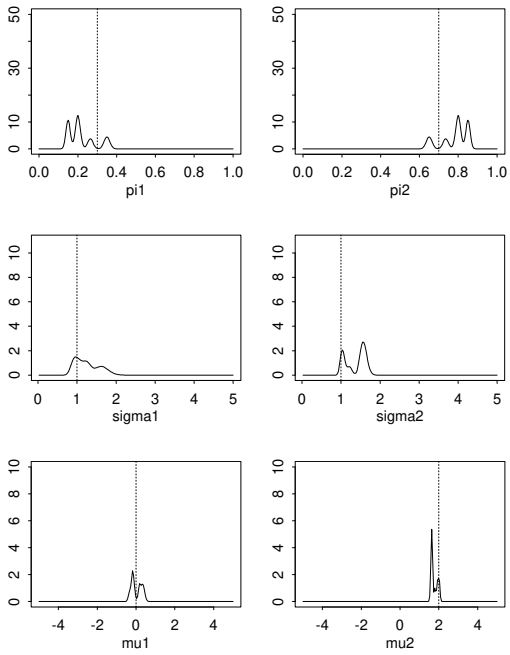


(b) KL

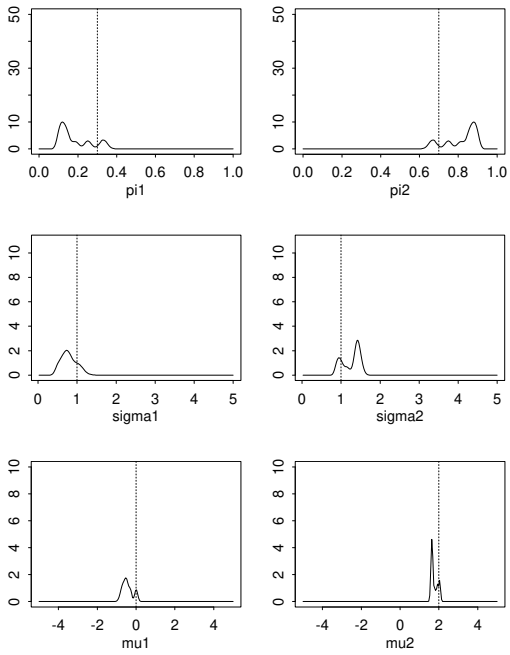


(c) GS$(10\,000, 1)$

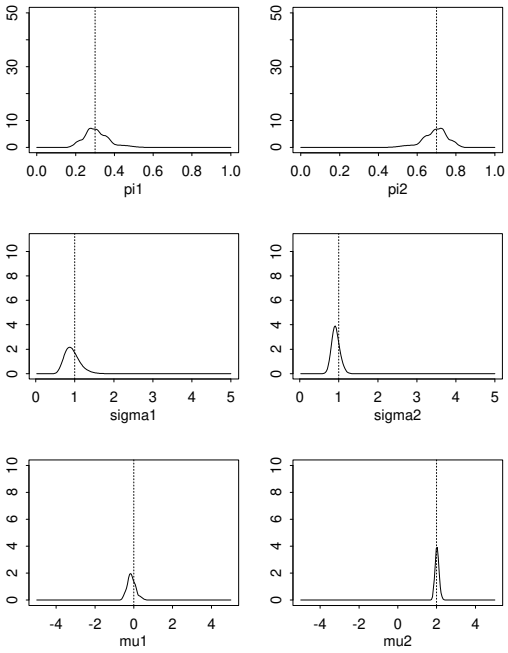
Figure 5.5: Estimates of the marginal posterior densities for the components of $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma^2}$ using QB, KL, and GS$(10\,000, 1)$ with observations from (5.27). The true parameter values are indicated with a vertical dotted line.



(a) QB(5)

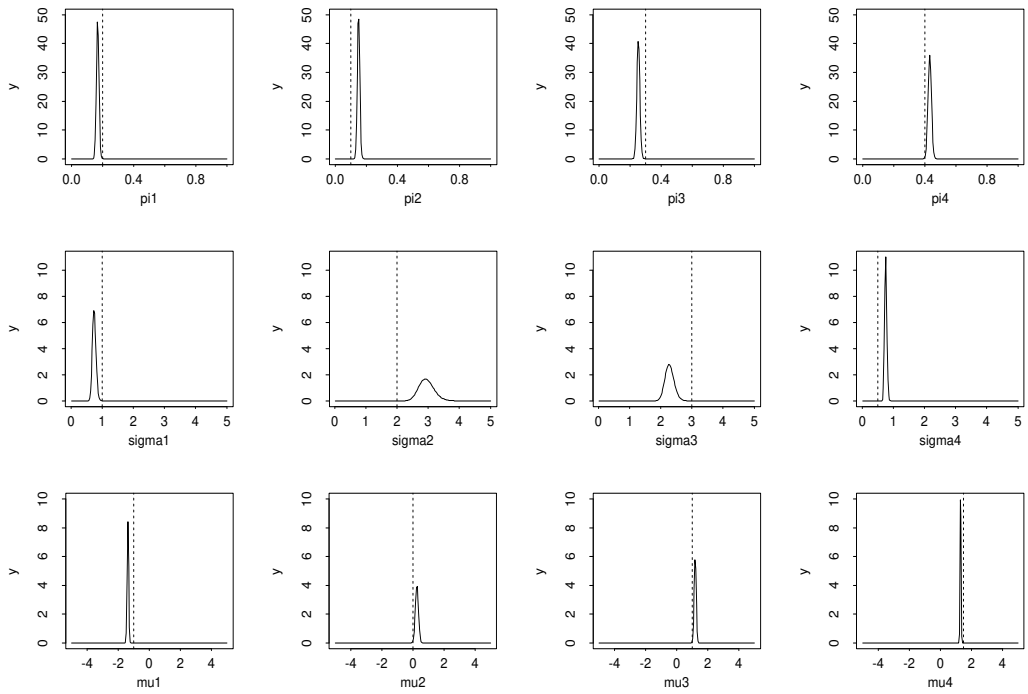


(b) KL(5)

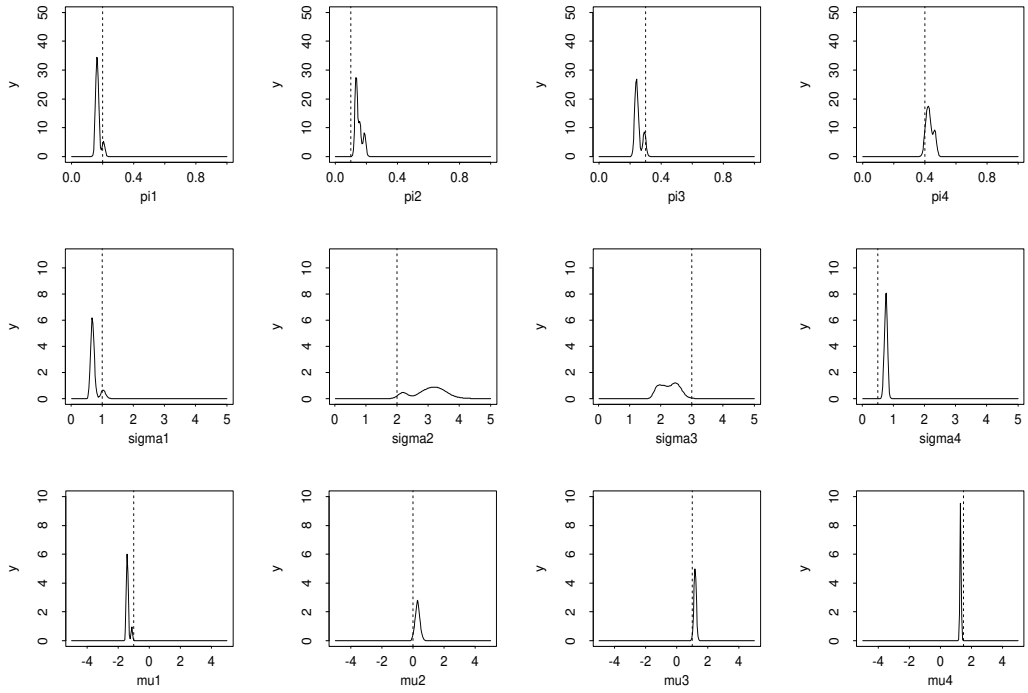


(c) GS$(10\,000, 5)$

Figure 5.6: Estimates of the marginal posterior densities for the components of π, μ, and σ^2 using QB(5), KL(5), and GS$(10\,000, 5)$ with observations from (5.27). The true parameter values are indicated with a vertical dotted line.

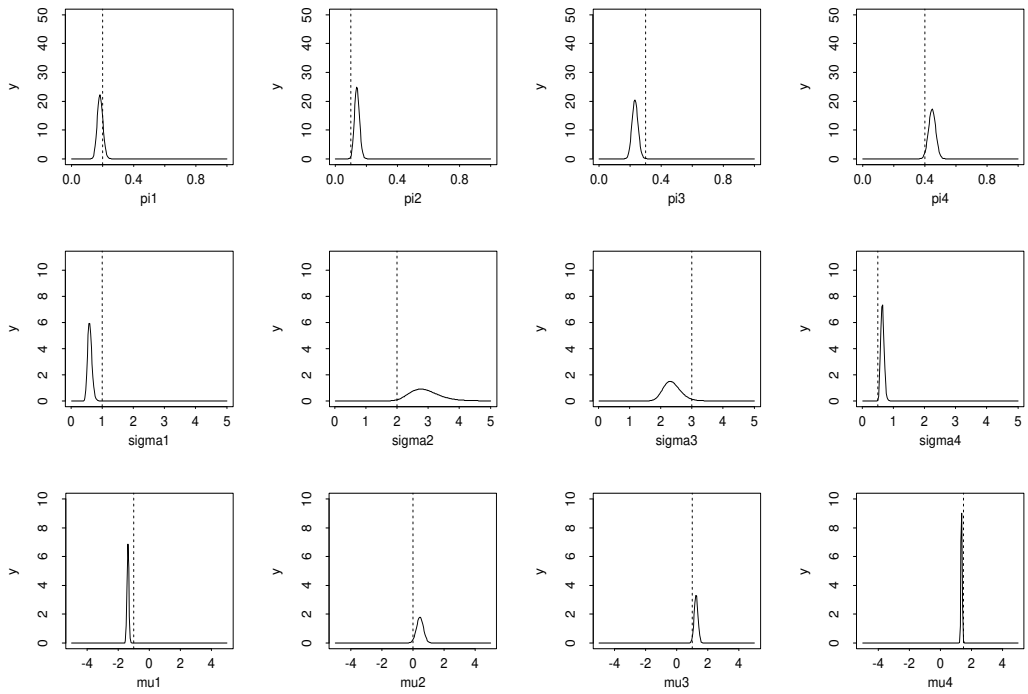


(a) QB

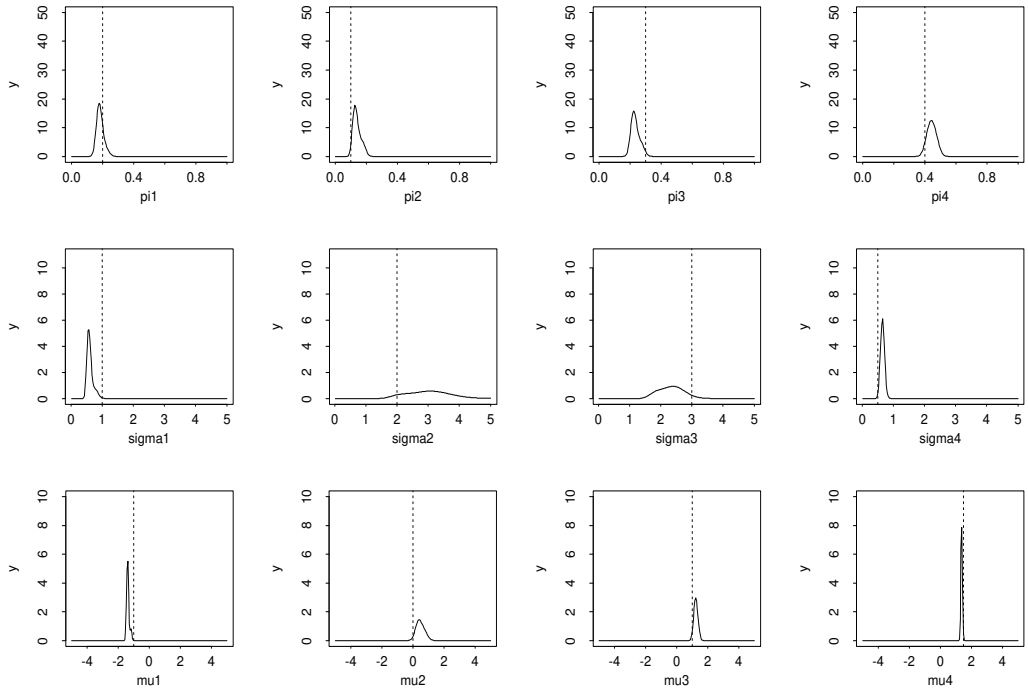


(b) QB(5)

Figure 5.7: Estimates of the marginal posterior densities for the components of π, μ, and σ^2 using QB and QB(5) approximations with observations from (5.28). The true parameter values are indicated with a vertical dotted line.

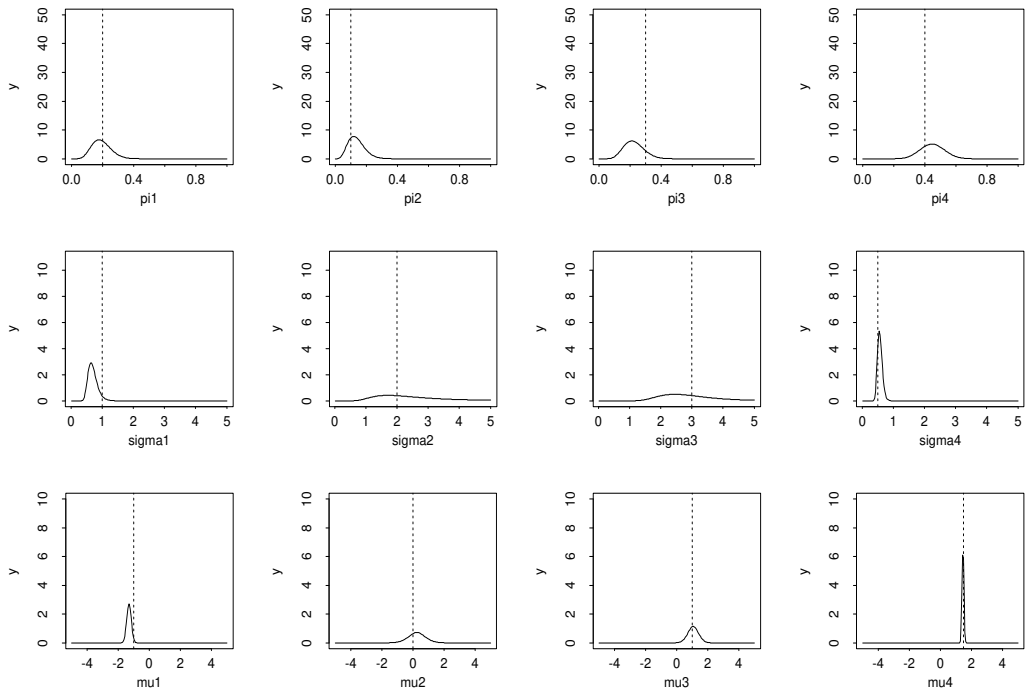


(a) KL

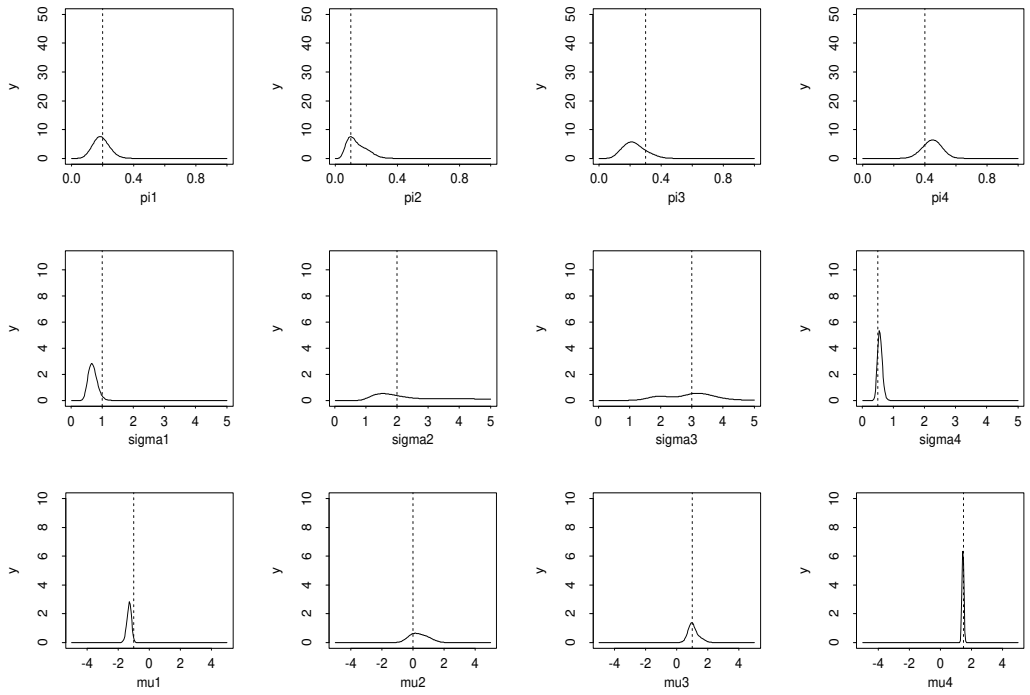


(b) KL(5)

Figure 5.8: Estimates of the marginal posterior densities for the components of π, μ, and σ^2 using KL and KL(5) approximations with observations from (5.28). The true parameter values are indicated with a vertical dotted line.



(a) GS$(10\,000, 1)$



(b) GS$(10\,000, 5)$

Figure 5.9: Estimates of the marginal posterior densities for the components of $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma^2}$ using GS$(10\,000, 1)$ and GS$(10\,000, 5)$ approximations with observations from (5.28). The true parameter values are indicated with a vertical dotted line.

5.6 Appendix

5.6.1 The full Bayesian solution for mixture components known, $k = 2$

Assuming we have a prior $p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; d_1, d_2)$ we show by induction that $p(\boldsymbol{\pi} \mid x^n)$ is of the form

$$p(\boldsymbol{\pi} \mid x^n) = \sum_{i=0}^{n} \lambda_i^{(n)} \mathcal{D}(\boldsymbol{\pi}; d_1 + i, d_2 + n - i) \tag{5.30}$$

and obtain a recurrence relation for the $\lambda_i^{(n)}$. Define for convenience $\lambda_{-1}^{(n)} = \lambda_{n+1}^{(n)} = 0$ for all n. Equation (5.30) clearly holds for $n = 0$ as $p(\boldsymbol{\pi} \mid x^0) = p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; d_1, d_2)$. Assume it holds for $n = r$ and consider $p(\boldsymbol{\pi} \mid x^{r+1})$:

$$\begin{aligned}
p(\boldsymbol{\pi} \mid x^{r+1}) &\propto \sum_{i=0}^{r} \lambda_i^{(r)} \mathcal{D}(\boldsymbol{\pi}; d_1 + i, d_2 + r - i)(\pi_1 f_1(x_{r+1}) + \pi_2 f_2(x_{r+1})) \\
&\propto \sum_{i=0}^{r} \frac{\left(\lambda_i^{(r)} f_1(x_{r+1}) \pi_1^{d_1+i} \pi_2^{d_2+r-i-1} + \lambda_i^{(r)} f_2(x_{r+1}) \pi_1^{d_1+i-1} \pi_2^{d_2+r-i}\right)}{\Gamma(d_1 + i)\Gamma(d_2 + r - i)} \\
&\propto \sum_{i=0}^{r+1} \Big(\frac{\lambda_{i-1}^{(r)} f_1(x_{r+1})}{\Gamma(d_1 + i - 1)\Gamma(d_2 + r - i + 1)} + \frac{\lambda_i^{(r)} f_2(x_{r+1})}{\Gamma(d_1 + i)\Gamma(d_2 + r - i)} \Big) \cdot \\
&\qquad \cdot \pi_1^{d_1+i-1} \pi_2^{d_2+r-i} \\
&\propto \sum_{i=0}^{r+1} (\lambda_{i-1}^{(r)} f_1(x_{r+1})(d_1 + i - 1) + \lambda_i^{(r)} f_2(x_{r+1})(d_2 + r - i)) \cdot \\
&\qquad \cdot \frac{\pi_1^{d_1+i-1} \pi_2^{d_2+r-i}}{\Gamma(d_1 + i)\Gamma(d_2 + r + 1 - i)} \\
&\propto \sum_{i=0}^{r+1} \lambda_i^{(r+1)} \mathcal{D}(\boldsymbol{\pi}; d_1 + i, d_2 + (r + 1) - i)
\end{aligned}$$

where

$$\lambda_i^{(r+1)} \propto \lambda_{i-1}^{(r)} f_1(x_{r+1})(d_1 + i - 1) + \lambda_i^{(r)} f_2(x_{r+1})(d_2 + r - i)$$

and $\sum_{i=0}^{r+1} \lambda_i^{(r+1)} = 1$ as $p(\boldsymbol{\pi} \mid x^{r+1})$ is a density.

5.6.2 The conjugate analysis for Case 2

For independent observations $x_1, x_2, \ldots$ from a mixture with density

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \pi_1 \mathcal{N}(x; \mu_1, v_1^{-1}) + \cdots + \pi_k \mathcal{N}(x; \mu_k, v_k^{-1}) \tag{5.31}$$

the form of the conjugate prior for $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{v})$ is

$$p(\theta) = \mathcal{D}(\boldsymbol{\pi}; \mathbf{d}) \prod_{j=1}^{k} \Gamma(v_j; m_j/2, l_j/2) \mathcal{N}(\mu_j; u_j, 1/(s_j v_j)) \tag{5.32}$$

$$= g(\theta; \mathbf{d}, \mathbf{m}, \mathbf{l}, \mathbf{u}, \mathbf{s}) \text{ say,} \tag{5.33}$$

and the corresponding posterior for θ given classified observations (x^n, z^n) is

$$p(\theta \mid x^n, z^n) = g(\theta; \mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}') \tag{5.34}$$

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

$$d'_j = d_j + n_j \tag{5.35}$$

$$m'_j = m_j + n_j \tag{5.36}$$

$$l'_j = l_j + SS_j + \Big(\frac{n_j s_j}{s_j + n_j} (\bar{x}_j - u_j)^2 \Big) \tag{5.37}$$

$$u'_j = u_j + n_j \Big(\frac{\bar{x}_j - u_j}{s_j + n_j} \Big) \tag{5.38}$$

$$s'_j = s_j + n_j \tag{5.39}$$

where

$$n_j = \#\{i : z_i = j\} \tag{5.40}$$

$$\bar{x}_j = \frac{1}{n_j} \sum_{i: z_i = j} x_i \tag{5.41}$$

$$SS_j = \sum_{i: z_i = j} (x_i - \bar{x}_j)^2. \tag{5.42}$$

See, for example, Robert (1994, pages 154-155).

5.6.3 Details of the Gibbs sampler for Case 2

The Gibbs sampler makes use of the conjugate analysis given in Section 5.6.2, and the theory outlined in Chapter 2. We assume a prior of the form (5.32), and given values $(\theta^{(t)}, (z^n)^{(t)})$ we simulated values for $(\theta^{(t+1)}, (z^n)^{(t+1)})$ as follows:

1. Simulate $\theta^{(t+1)}$ from $p(\theta \mid x^n, (z^n)^{(t)})$ which is given by (5.34).

2. Simulate $(z^n)^{(t+1)}$ by simulating $z_1, \ldots, z_n$ independently from

$$p(z_j = i \mid \theta^{(t+1)}) \propto \pi_i^{(t+1)} \mathcal{N}(x_j; \mu_i^{(t+1)}, v_i^{(t+1)}).$$

This simulates from a Markov chain with stationary distribution $p(\theta, z^n \mid x^n)$, as shown in Chapter 2. A starting point was chosen by randomly dividing the n observations x^n evenly between the k components, to give an initial value for $(z^n)^{(0)}$.

5.6.4 QB and KL update steps for Case 1

For independent observations $x_1, x_2, \ldots$ from a mixture with density

$$p(x \mid \boldsymbol{\pi}) = \pi_1 f_1(x) + \cdots + \pi_k f_k(x) \tag{5.43}$$

where the component densities $f_1, \ldots, f_k$ are assumed known, the form of the conjugate prior for $\boldsymbol{\pi}$ is

$$p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; \mathrm{d}) \tag{5.44}$$

and on receiving an observation x, the sequential update step is given by

$$\mathrm{SU}(\mathrm{d}, x) = \mathrm{d}'$$

where d' is given by

QB: $\mathrm{d}' = (d_1 + p(1), \ldots, d_k + p(k))$.

KL: $\mathrm{d}' = (d'_1, \ldots, d'_k)$ chosen by maximising (using numerical methods)

$$\sum_{j=1}^{k} (d'_j - 1) \sum_{i=1}^{k} p(i)[\psi\{d_j(i)\} - \psi\{d_1 + \cdots + d_k + 1\}] + \log \Gamma(\sum_{j=1}^{k} d'_j) - \sum_{j=1}^{k} \log \Gamma(d'_j) \tag{5.45}$$

where (for $i = 1, \ldots, k$)

$$p(i) = \frac{\widehat{\pi}_i f_i(x)}{\widehat{\pi}_1 f_1(x) + \cdots + \widehat{\pi}_k f_k(x)} \tag{5.46}$$

$$\widehat{\pi}_i = \frac{d_i}{d_1 + \cdots + d_k} \tag{5.47}$$

$$d_j(i) = d_j + \delta_{ij} \quad (j = 1, \ldots, k) \tag{5.48}$$

$$\delta_{ij} = 1 \text{ if } i = j, \text{ and 0 otherwise} \tag{5.49}$$

and $\psi(\cdot)$ is the *digamma function* (the first derivative of the log of the gamma function). See for example Bernardo and Girón (1988).

5.6.5 QB and KL update steps for Case 2

For independent observations $x_1, x_2, \ldots$ from a mixture with density

$$p(x \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \pi_1 \mathcal{N}(x; \mu_1, v_1^{-1}) + \cdots + \pi_k \mathcal{N}(x; \mu_k, v_k^{-1}). \tag{5.50}$$

the form of the conjugate prior for $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{v})$ is

$$p(\theta) = \mathcal{D}(\boldsymbol{\pi}; \mathbf{d}) \prod_{j=1}^{k} \Gamma(v_j; m_j/2, l_j/2) \mathcal{N}(\mu_j; u_j, 1/(s_j v_j)) \tag{5.51}$$

$$= g(\theta; \mathbf{d}, \mathbf{m}, \mathbf{l}, \mathbf{u}, \mathbf{s}) \text{ say,} \tag{5.52}$$

and on receiving an observation x, the sequential update step is given by

$$\mathrm{SU}\big((\mathbf{d}, \mathbf{m}, \mathbf{l}, \mathbf{u}, \mathbf{s}), x\big) = (\mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}')$$

where $(\mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}')$ is given by

QB: For $j = 1, \ldots, k$

$$d_j' = d_j + p(j) \tag{5.53}$$

$$m_j' = m_j + p(j) \tag{5.54}$$

$$l_j' = l_j + p(j)\Big(\frac{s_j}{s_j + 1}(x - u_j)^2\Big) \tag{5.55}$$

$$u_j' = u_j + p(j)\Big(\frac{x - u_j}{s_j + 1}\Big) \tag{5.56}$$

$$s_j' = s_j + p(j) \tag{5.57}$$

KL: $\mathbf{d}' = (d_1', \ldots, d_k')$ is chosen by maximising (using numerical methods)

$$\sum_j (d_j' - 1) \sum_i p(i)[\psi\{d_j(i)\} - \psi\{d_1 + \cdots + d_k + 1\}] + \log \Gamma\big(\sum_j d_j'\big) - \sum_j \log \Gamma(d_j'), \tag{5.58}$$

m_j' $(j = 1, \ldots, k)$ is chosen by maximising (using numerical methods)

$$\frac{m_j'}{2} \sum_i p(i) \Big[\psi\Big\{\frac{m_j(i)}{2}\Big\} - \log\Big\{\frac{l_j(i)}{2}\Big\}\Big] + \frac{m_j'}{2}\Big(\log \frac{m_j'}{2} - 1\Big) - \log \Gamma\Big(\frac{m_j'}{2}\Big) \tag{5.59}$$

and

$$l_j' = m_j' \Big/ \sum_i p(i) \frac{m_j(i)}{l_j(i)} \tag{5.60}$$

$$u_j' = \sum_i p(i) \frac{m_j(i)}{l_j(i)} u_j(i) \Big/ \sum_i p(i) \frac{m_j(i)}{l_j(i)} \tag{5.61}$$

$$s_j' = 1 \Big/ \sum_i p(i) \Big[\frac{1}{s_j(i)} + \frac{m_j(i)}{l_j(i)} (u_j' - u_j(i))^2\Big] \tag{5.62}$$

where for $i, j = 1, \ldots, k$

$$p(i) \propto d_i \left(\frac{s_i}{s_i + 1} \right)^{\frac{1}{2}} \frac{l_i^{m_i/2}}{\left[l_i + \frac{s_i}{s_i+1} \big(x - u_i\big)^2 \right]^{(m_i+1)/2}} \frac{\Gamma\big((m_i+1)/2\big)}{\Gamma\big(m_i/2\big)} \tag{5.63}$$

$$d_j(i) = d_j + \delta_{ij} \tag{5.64}$$

$$m_j(i) = m_j + \delta_{ij} \tag{5.65}$$

$$l_j(i) = l_j + \Big(\frac{s_i}{s_i + 1} (x - u_i)^2 \Big) \delta_{ij} \tag{5.66}$$

$$u_j(i) = u_j + \Big(\frac{x - u_i}{s_i + 1} \Big) \delta_{ij} \tag{5.67}$$

$$s_j(i) = s_j + \delta_{ij} \tag{5.68}$$

and $\psi(\cdot)$ is the *digamma function* (the first derivative of the log of the gamma function).

Proof. Given the prior distribution (5.51) for θ, the posterior distribution on receiving a single observation x is (by the conjugate analysis given in Section 5.6.2, and straightforward integration)

$$p(\theta \mid x) = \sum_{i=1}^{k} p(i) g(\theta; \mathbf{d}(i), \mathbf{m}(i), \mathbf{l}(i), \mathbf{u}(i), \mathbf{s}(i))$$

where $p(i)$ is given by (5.63), and $\mathbf{d}(i), \mathbf{m}(i), \mathbf{l}(i), \mathbf{u}(i), \mathbf{s}(i)$ are given by (5.64)–(5.68).

The QB sequential update rule then follows from

$$\mathbf{d}' = \sum_{i=1}^{k} p(i) \mathbf{d}(i)$$

and similar equations for $\mathbf{m}', \mathbf{l}', \mathbf{u}'$ and $\mathbf{s}'$.

The KL sequential update rule requires us to find $(\mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}')$ to minimise

$$KL\Big[\sum_{i=1}^{k} p(i) g\big(\theta; \mathbf{d}(i), \mathbf{m}(i), \mathbf{l}(i), \mathbf{u}(i), \mathbf{s}(i)\big) \,\|\, g(\theta; \mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}') \Big]$$

which, by the definition of $KL[\cdot \,\|\, \cdot]$ (given by (5.8), page 122) involves *maximising*

$$\begin{aligned} K(\mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}') &= \int \sum_i p(i) g(\theta; \mathbf{d}(i), \mathbf{m}(i), \mathbf{l}(i), \mathbf{u}(i), \mathbf{s}(i)) \cdot \\ &\qquad \cdot \log g(\theta; \mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}') \, d\theta \\ &= \sum_i p(i) K^{(i)}(\mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}') \end{aligned}$$

where

$$K^{(i)}(\mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}') = \int g(\theta; \mathbf{d}(i), \mathbf{m}(i), \mathbf{l}(i), \mathbf{u}(i), \mathbf{s}(i)) \log g(\theta; \mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}')\, d\theta. \tag{5.69}$$

Now

$$g(\theta; \mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}') = \mathcal{D}(\boldsymbol{\pi}; \mathbf{d}') \prod_j \Gamma(v_j; m_j'/2, l_j'/2) \mathcal{N}(\mu_j; u_j', 1/(s_j' v_j))$$

and so

$$\begin{aligned}
\log g = & \sum_{j=1}^{k} \{(d_j' - 1) \log \pi_j - \log \Gamma(d_j')\} + \log \Gamma(d_1' + \cdots + d_k') \\
& + \sum_{j=1}^{k} \{(\tfrac{m_j'}{2} - 1) \log v_j - \tfrac{l_j'}{2} v_j + \tfrac{m_j'}{2} \log \tfrac{l_j'}{2} - \log \Gamma(\tfrac{m_j'}{2})\} \\
& + \sum_{j=1}^{k} \{\tfrac{1}{2} \log(\tfrac{s_j' v_j}{2\pi}) - \tfrac{s_j'}{2} v_j (\mu_j - u_j')^2\}.
\end{aligned}$$

If we write $E_i(\cdot)$ for $\int \cdot g(\theta; \mathbf{d}(i), \mathbf{m}(i), \mathbf{l}(i), \mathbf{u}(i), \mathbf{s}(i))\, d\theta$ then the following results can be obtained by straightforward integration:

$$\begin{aligned}
E_i(\log \pi_j) &= [\psi(d_j(i)) - \psi(d_1(i) + \cdots + d_k(i))] \\
E_i(\log v_j) &= [\psi(\tfrac{m_j(i)}{2}) - \log(\tfrac{l_j(i)}{2})] \\
E_i(v_j) &= \tfrac{m_j(i)}{l_j(i)} \\
E_i(v_j(\mu_j - u_j')^2) &= E_i(v_j([\tfrac{1}{s_j(i) v_j} + (u_j' - u_j(i))^2])) \\
&= [\tfrac{1}{s_j(i)} + \tfrac{m_j(i)}{l_j(i)} (u_j' - u_j(i))^2]
\end{aligned}$$

Gathering the $\log v_j$ terms and performing the integration (5.69) we obtain

$$\begin{aligned}
& K^{(i)}(\mathbf{d}', \mathbf{m}', \mathbf{l}', \mathbf{u}', \mathbf{s}') \\
& = \sum_{j=1}^{k} \{(d_j' - 1)[\psi(d_j(i)) - \psi(d_1(i) + \cdots + d_k(i))] - \log \Gamma(d_j')\} + \log \Gamma(d_1' + \cdots + d_k') \\
& \quad + \sum_{j=1}^{k} \{(\tfrac{m_j'}{2} - \tfrac{1}{2})[\psi(\tfrac{m_j(i)}{2}) - \log(\tfrac{l_j(i)}{2})] - \tfrac{l_j'}{2} \tfrac{m_j(i)}{l_j(i)} + \tfrac{m_j'}{2} \log \tfrac{l_j'}{2} - \log \Gamma(\tfrac{m_j'}{2})\} \\
& \quad + \sum_{j=1}^{k} \{\tfrac{1}{2} \log(\tfrac{s_j'}{2\pi}) - \tfrac{s_j'}{2} [\tfrac{1}{s_j(i)} + \tfrac{m_j(i)}{l_j(i)} (u_j' - u_j(i))^2]\}
\end{aligned} \tag{5.70}$$

We wish to maximise $K = \sum_i p(i)K^{(i)}$. Noting that $d_1(i) + \cdots + d_k(i) = d_1 + \cdots + d_k + 1$ gives (5.58). Equations (5.60),(5.61) and (5.62) can be obtained by differentiating with respect to l'_j, u'_j and s'_j ($j = 1, \ldots, k$) respectively. (5.59) is obtained by substituting (5.60) for l'_j in (5.70).

□

Appendix A

Tables of data

A.1 Galaxy data

The galaxy data consists of the velocities (in 10^3 km/s) of distant galaxies diverging from our own. The version of the data given here is taken from Roeder (1990), which omits an observation (a velocity of 5.607×10^3 km/s) which is in the original dataset presented by Postman *et al.* (1986).

Obs	Velocity (10^3 km/s)	Obs	Velocity (10^3 km/s)	Obs	Velocity (10^3 km/s)	Obs	Velocity (10^3 km/s)
1	9.172	22	19.541	43	20.875	64	23.263
2	9.350	23	19.547	44	20.986	65	23.484
3	9.483	24	19.663	45	21.137	66	23.538
4	9.558	25	19.846	46	21.492	67	23.542
5	9.775	26	19.856	47	21.701	68	23.666
6	10.227	27	19.863	48	21.814	69	23.706
7	10.406	28	19.914	49	21.921	70	23.711
8	16.084	29	19.918	50	21.960	71	24.129
9	16.170	30	19.973	51	22.185	72	24.285
10	18.419	31	19.989	52	22.209	73	24.289
11	18.552	32	20.166	53	22.242	74	24.366
12	18.600	33	20.175	54	22.249	75	24.717
13	18.927	34	20.179	55	22.314	76	24.990
14	19.052	35	20.196	56	22.374	77	25.633
15	19.070	36	20.215	57	22.495	78	26.960
16	19.330	37	20.221	58	22.746	79	26.995
17	19.343	38	20.415	59	22.747	80	32.065
18	19.349	39	20.629	60	22.888	81	32.789
19	19.440	40	20.795	61	22.914	82	34.279
20	19.473	41	20.821	62	23.206		
21	19.529	42	20.846	63	23.241		

A.2 Old Faithful data

The Old Faithful data (Härdle, 1991, the version from) 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.

Obs	duration (min)	waiting (min)	Obs	duration (min)	waiting (min)	Obs	duration (min)	waiting (min)
1	3.600	79	36	2.017	52	71	4.033	82
2	1.800	54	37	1.867	48	72	1.967	56
3	3.333	74	38	4.833	80	73	4.500	79
4	2.283	62	39	1.833	59	74	4.000	71
5	4.533	85	40	4.783	90	75	1.983	62
6	2.883	55	41	4.350	80	76	5.067	76
7	4.700	88	42	1.883	58	77	2.017	60
8	3.600	85	43	4.567	84	78	4.567	78
9	1.950	51	44	1.750	58	79	3.883	76
10	4.350	85	45	4.533	73	80	3.600	83
11	1.833	54	46	3.317	83	81	4.133	75
12	3.917	84	47	3.833	64	82	4.333	82
13	4.200	78	48	2.100	53	83	4.100	70
14	1.750	47	49	4.633	82	84	2.633	65
15	4.700	83	50	2.000	59	85	4.067	73
16	2.167	52	51	4.800	75	86	4.933	88
17	1.750	62	52	4.716	90	87	3.950	76
18	4.800	84	53	1.833	54	88	4.517	80
19	1.600	52	54	4.833	80	89	2.167	48
20	4.250	79	55	1.733	54	90	4.000	86
21	1.800	51	56	4.883	83	91	2.200	60
22	1.750	47	57	3.717	71	92	4.333	90
23	3.450	78	58	1.667	64	93	1.867	50
24	3.067	69	59	4.567	77	94	4.817	78
25	4.533	74	60	4.317	81	95	1.833	63
26	3.600	83	61	2.233	59	96	4.300	72
27	1.967	55	62	4.500	84	97	4.667	84
28	4.083	76	63	1.750	48	98	3.750	75
29	3.850	78	64	4.800	82	99	1.867	51
30	4.433	79	65	1.817	60	100	4.900	82
31	4.300	73	66	4.400	92	101	2.483	62
32	4.467	77	67	4.167	78	102	4.367	88
33	3.367	66	68	4.700	78	103	2.100	49
34	4.033	80	69	2.067	65	104	4.500	83
35	3.833	74	70	4.700	73	105	4.050	81

Obs	duration (min)	waiting (min)	Obs	duration (min)	waiting (min)	Obs	duration (min)	waiting (min)
106	1.867	47	146	1.983	59	186	4.433	78
107	4.700	84	147	4.633	80	187	4.083	84
108	1.783	52	148	2.017	49	188	1.833	46
109	4.850	86	149	5.100	96	189	4.417	83
110	3.683	81	150	1.800	53	190	2.183	55
111	4.733	75	151	5.033	77	191	4.800	81
112	2.300	59	152	4.000	77	192	1.833	57
113	4.900	89	153	2.400	65	193	4.800	76
114	4.417	79	154	4.600	81	194	4.100	84
115	1.700	59	155	3.567	71	195	3.966	77
116	4.633	81	156	4.000	70	196	4.233	81
117	2.317	50	157	4.500	81	197	3.500	87
118	4.600	85	158	4.083	93	198	4.366	77
119	1.817	59	159	1.800	53	199	2.250	51
120	4.417	87	160	3.967	89	200	4.667	78
121	2.617	53	161	2.200	45	201	2.100	60
122	4.067	69	162	4.150	86	202	4.350	82
123	4.250	77	163	2.000	58	203	4.133	91
124	1.967	56	164	3.833	78	204	1.867	53
125	4.600	88	165	3.500	66	205	4.600	78
126	3.767	81	166	4.583	76	206	1.783	46
127	1.917	45	167	2.367	63	207	4.367	77
128	4.500	82	168	5.000	88	208	3.850	84
129	2.267	55	169	1.933	52	209	1.933	49
130	4.650	90	170	4.617	93	210	4.500	83
131	1.867	45	171	1.917	49	211	2.383	71
132	4.167	83	172	2.083	57	212	4.700	80
133	2.800	56	173	4.583	77	213	1.867	49
134	4.333	89	174	3.333	68	214	3.833	75
135	1.833	46	175	4.167	81	215	3.417	64
136	4.383	82	176	4.333	81	216	4.233	76
137	1.883	51	177	4.500	73	217	2.400	53
138	4.933	86	178	2.417	50	218	4.800	94
139	2.033	53	179	4.000	85	219	2.000	55
140	3.733	79	180	4.167	74	220	4.150	76
141	4.233	81	181	1.883	55	221	1.867	50
142	2.233	60	182	4.583	77	222	4.267	82
143	4.533	82	183	4.250	83	223	1.750	54
144	4.817	77	184	3.767	83	224	4.483	75
145	4.333	76	185	2.033	51	225	4.000	78

Obs	duration (min)	waiting (min)	Obs	duration (min)	waiting (min)	Obs	duration (min)	waiting (min)
226	4.117	79	242	2.350	47	258	4.450	83
227	4.083	78	243	4.933	86	259	2.000	56
228	4.267	78	244	2.900	63	260	4.283	79
229	3.917	70	245	4.583	85	261	4.767	78
230	4.550	79	246	3.833	82	262	4.533	84
231	4.083	70	247	2.083	57	263	1.850	58
232	2.417	54	248	4.367	82	264	4.250	83
233	4.183	86	249	2.133	67	265	1.983	43
234	2.217	50	250	4.350	74	266	2.250	60
235	4.450	90	251	2.200	54	267	4.750	75
236	1.883	54	252	4.450	83	268	4.117	81
237	1.850	54	253	3.567	73	269	2.150	46
238	4.283	77	254	4.500	73	270	4.417	90
239	3.950	79	255	4.150	88	271	1.817	46
240	2.333	64	256	3.817	80	272	4.467	74
241	4.150	75	257	3.917	71			

A.3 *Iris Virginica* data

The original *Iris* data was collected by Anderson (1935), and consists of four measurements (petal and sepal length and width) for 50 specimens of each of three species (*setosa, versicolor*, and *virginica*) of iris, giving 150 specimens in all. We consider only the measurements of sepal length and petal length for the 50 examples of the *virginica* species. The observations are numbered 1 to 50 in order of their listing in Table 1.1 in Andrews and Herzberg (1985), which is the numbering used by McLachlan (1992) and also the numbering used in the data supplied with S-PLUS.

Obs	sepal length (mm)	petal length (mm)	Obs	sepal length (mm)	petal length (mm)
1	6.3	6.0	26	7.2	6.0
2	5.8	5.1	27	6.2	4.8
3	7.1	5.9	28	6.1	4.9
4	6.3	5.6	29	6.4	5.6
5	6.5	5.8	30	7.2	5.8
6	7.6	6.6	31	7.4	6.1
7	4.9	4.5	32	7.9	6.4
8	7.3	6.3	33	6.4	5.6
9	6.7	5.8	34	6.3	5.1
10	7.2	6.1	35	6.1	5.6
11	6.5	5.1	36	7.7	6.1
12	6.4	5.3	37	6.3	5.6
13	6.8	5.5	38	6.4	5.5
14	5.7	5.0	39	6.0	4.8
15	5.8	5.1	40	6.9	5.4
16	6.4	5.3	41	6.7	5.6
17	6.5	5.5	42	6.9	5.1
18	7.7	6.7	43	5.8	5.1
19	7.7	6.9	44	6.8	5.9
20	6.0	5.0	45	6.7	5.7
21	6.9	5.7	46	6.7	5.2
22	5.6	4.9	47	6.3	5.0
23	7.7	6.7	48	6.5	5.2
24	6.3	4.9	49	6.2	5.4
25	6.7	5.7	50	5.9	5.1

A.4 Pima data

The Pima data is discussed by Ripley (1996), and the full data consists of the readings of 9 variables for 768 women of Pima Indian heritage, living near Phoenix, Arizona. These women were tested for diabetes according to World Health Organisation criteria, and we consider only those who tested positive. We consider only two of the measured variables: *plasma glucose concentration (glu)* and *diastolic blood pressure (bp)* (in mm Hg). We discarded records which had one or both of these variables missing, leaving 250 records for analysis.

Observation	glu	bp	Observation	glu	bp	Observation	glu	bp
1	148	72	34	134	72	67	139	80
2	183	64	35	122	90	68	159	66
3	137	40	36	163	72	69	158	84
4	78	50	37	95	85	70	107	62
5	197	70	38	171	72	71	109	64
6	125	96	39	155	62	72	148	60
7	168	74	40	160	54	73	196	76
8	189	60	41	146	92	74	162	104
9	166	72	42	124	74	75	184	84
10	118	84	43	162	76	76	140	65
11	107	74	44	113	76	77	112	82
12	115	70	45	88	30	78	151	70
13	196	90	46	117	88	79	109	62
14	119	80	47	105	84	80	85	74
15	143	94	48	173	70	81	112	66
16	125	70	49	122	56	82	177	60
17	147	76	50	170	64	83	158	90
18	158	76	51	108	66	84	162	52
19	102	76	52	156	86	85	142	86
20	90	68	53	188	78	86	134	80
21	111	72	54	152	88	87	171	72
22	171	110	55	163	72	88	181	84
23	180	66	56	131	88	89	179	90
24	103	66	57	104	74	90	164	84
25	176	90	58	102	82	91	139	54
26	187	68	59	134	70	92	119	50
27	133	72	60	179	72	93	184	85
28	114	66	61	129	110	94	92	62
29	109	88	62	130	82	95	113	64
30	100	66	63	194	68	96	155	76
31	126	90	64	181	68	97	123	62
32	137	108	65	128	98	98	101	86
33	136	70	66	109	76	99	106	60

Observation	glu	bp	Observation	glu	bp	Observation	glu	bp
100	146	70	141	131	66	182	198	66
101	161	86	142	193	70	183	121	66
102	108	80	143	95	64	184	118	80
103	119	86	144	136	84	185	197	70
104	107	62	145	168	64	186	151	90
105	128	78	146	115	72	187	124	76
106	128	48	147	197	74	188	143	66
107	146	70	148	172	68	189	176	86
108	100	78	149	138	60	190	111	84
109	144	58	150	173	84	191	132	80
110	115	98	151	144	82	192	188	82
111	161	68	152	129	64	193	173	74
112	128	68	153	151	78	194	150	78
113	124	68	154	184	78	195	181	78
114	155	74	155	181	64	196	174	58
115	109	80	156	95	82	197	168	88
116	182	74	157	189	104	198	138	74
117	194	78	158	108	70	199	112	82
118	112	74	159	117	62	200	114	64
119	124	70	160	180	78	201	104	72
120	152	90	161	104	64	202	97	76
121	122	64	162	134	70	203	147	80
122	102	86	163	175	62	204	167	74
123	115	76	164	148	84	205	179	50
124	152	78	165	105	80	206	136	84
125	178	84	166	158	70	207	155	52
126	165	88	167	135	68	208	80	82
127	125	50	168	125	70	209	199	76
128	196	76	169	195	70	210	167	106
129	189	64	170	180	90	211	145	80
130	146	78	171	84	72	212	115	60
131	124	72	172	163	70	213	145	82
132	133	102	173	145	88	214	111	70
133	173	82	174	130	70	215	195	70
134	140	82	175	129	92	216	156	86
135	156	75	176	100	74	217	121	52
136	116	74	177	128	72	218	162	76
137	105	100	178	90	85	219	125	80
138	144	82	179	186	90	220	144	82
139	166	76	180	187	76	221	158	114
140	158	78	181	125	76	222	129	68

Observation	glu	bp	Observation	glu	bp	Observation	glu	bp
223	142	90	233	149	68	243	136	70
224	169	74	234	130	78	244	181	88
225	125	78	235	120	86	245	154	78
226	168	88	236	174	88	246	128	88
227	164	78	237	102	74	247	123	72
228	93	64	238	120	80	248	190	92
229	129	62	239	140	94	249	170	74
230	187	50	240	147	94	250	126	60
231	173	78	241	187	70			
232	97	76	242	162	62			

Bibliography

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Eds B. N. Petrov and F. Cáski), pp. 267–281, Budapest. Akademiai Kaidó. Reprinted in *Breakthroughs in Statistics*, eds Kotz, S. & Johnson, N. L. (1992), volume I, pp. 599–624. New York: Springer.

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

Andrews, D. F. and Herzberg, A. M. (1985) *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. New York: Springer-Verlag.

Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Berger, J. O. and Bernardo, J. M. (1989) Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association*, **84**, 200–207.

Berger, J. O. and Bernardo, J. M. (1992) On the development of reference priors (with discussion). In *Bayesian Statistics 4* (Eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 35–60. Oxford University Press.

Bernardo, J. M. (1979) Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society, series B*, **41**, 113–147.

Bernardo, J. M. (1997) Noninformative priors do not exist: A discussion with Jose M. Bernardo. *Journal of Statistical Planning and Inference*. To appear.

Bernardo, J. M. and Girón, F. J. (1988) A Bayesian analysis of simple mixture problems. In *Bayesian Statistics 3* (Eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 67–78. OUP.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Billingsley, P. (1986) *Probability and Measure*. New York: John Wiley and Son, second edition.

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.

Celeux, G., Chauveau, D. and Diebolt, J. (1995) On stochastic versions of the EM algorithm. Technical Report 2514, INRIA Rhône-Alpes.

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.

de Lapparent, V., Geller, M. J. and Huchra, J. P. (1986) A slice of the universe. *Astrophysical Journal*, **302**, L1–L5.

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.

Eaton, M. L. and Olkin, I. (1987) Best equivariant estimators of a Cholesky decomposition. *Annals of Statistics*, **15**, 1639–1650.

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

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, **6**, 721–741.

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

Green, P. J. (1994) Contribution to the discussion of paper by Grenander and Miller (1994). *Journal of the Royal Statistical Society, series B*, **56**, 589–590.

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.

Hosmer, D. W. (1973) A comparison of interactive maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, **29**, 761–770.

Huo, Q. and Lee, C.-H. (1997) On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate. *IEEE Transactions on Speech and Audio Processing*, **5**, 161–172.

Huo, Q., Chan, C. and Lee, C. (1995) Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition. *IEEE Transactions on Speech and Audio Processing*, **3**, 334–345.

Huo, Q., Chan, C. and Lee, C.-H. (1996) On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition. *IEEE Transactions on Speech and Audio Processing*, **4**, 141–144.

Izenman, A. J. and Sommer, C. J. (1988) Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, **83**, 941–953.

Jeffreys, H. (1967) *Theory of Probability*. Oxford: Clarendon Press, third edition.

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.

Kooperberg, C. and Stone, C. J. (1992) Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, **1**, 301–328.

Lloyd, S. P. (1957) Least squares quantization in PCM. Technical note, Bell Laboratories. [Published as Lloyd (1982)].

Lloyd, S. P. (1982) Least squares quantization in PCM. *IEEE Transactions on Information Theory*, **28**, 128–137.

Makov, U. E. and Smith, A. F. M. (1977) A quasi-Bayes unsupervised learning procedure for priors. *IEEE Transactions in Information Theory*, **23**, 761–764.

Marriott, F. H. C. (1975) Separating mixtures of normal distributions. *Biometrics*, **31**, 767–769.

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.

Pearson, K. (1894) Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, **185**, 71–110.

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.

Posse, C. (1997) Fully Bayesian analysis of mixtures of multivariate Gaussian data. Preprint from Christian Posse, Department of Statistics, University of Minnesota.

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.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.

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.

Smith, A. F. M. and Makov, U. E. (1978) A quasi-Bayes sequential procedure for mixtures. *Journal of the Royal Statistical Society, series B*, **40**, 106–111.

Smith, A. F. M. and Makov, U. E. (1981) Unsupervised learning for signal verses noise. *IEEE Transactions in Information Theory*, **27**, 498–500.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R. and Inskip, H. (1996) Hepatitis B: a case study in MCMC techniques. In *Markov Chain Monte Carlo in Practice* (Eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 21–43. Chapman & Hall.

Stephens, M. (1996) Dealing with the multimodal distributions of mixture model parameters. Available from the MCMC Preprint Service at `http://www.stats.bris.ac.uk/MCMC/`.

Stephens, M. (1997) Contribution to the discussion of paper by Richardson and Green (1997). *Journal of the Royal Statistical Society, series B*, **59**, 768–769.

Taha, H. A. (1989) *Operations Research: An Introduction*. New York: Macmillan Publishing Company, fourth edition.

Tanner, M. and Wong, W. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–550.

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701–1762.

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. (1997) Mixture distributions (update). In *Encyclopedia of Statistical Sciences*, volume Update Volume 1, pp. 399–407. New York: Wiley.

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

Tråvén, H. G. C. (1991) A neural network approach to statistical pattern classification by "semiparametric" estimation of probability density functions. *IEEE Transactions on Neural Networks*, **2**, 366–377.

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

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

Venables, W. N. and Ripley, B. D. (1997b) Complements to Venables and Ripley (1997a). On-line update, available from `http://www.stats.ox.ac.uk/pub/MASS2/index.html`.

Wand, M. P. and Jones, M. C. (1995) *Kernel smoothing*. London: Chapman & Hall.

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

West, M. (1997) Contribution to the discussion of paper by Richardson and Green (1997). *Journal of the Royal Statistical Society, series B*, **59**, 783–784.

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.