Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Scuola di Dottorato in
Scienze Statistiche
Ciclo XX

Sampling from a variable dimension mixture model posterior

Antonio Parisi

Direttore: Prof.ssa A. Salvan

Supervisore: Prof. S. Coles

Co-supervisori: Prof. B. Liseo, Prof. C. Robert

31/01/2008

Summary

Goal of the thesis is the analysis of a real dataset concerning a biological problem that obtained an increasing interest in recent years. Commercial stocks of fish are not sufficient anymore to satisfy the global demand. Hence, fishermen are beginning to catch species living in the deep. As little is known about these species, there is an actual risk of extinction of these species.

As it is typically difficult and expensive to gather the ages of fish, in order to implement stock management policies, it is necessary to build up reliable growth models to infer ages from length data. The lengths, if we don't observe the ages, come from a mixture distribution, in which the components are the different cohorts.

As MCMC methods are not always satisfactory for the analysis of mixture models, to estimate the parameters of the model and the number of cohorts that form the sample, it is employed a Population Monte Carlo algorithm for mixtures generalized to the case of unknown number of components.

Riassunto

Scopo della tesi è l'analisi di un dataset reale che riguarda un problema biologico che sta ottenendo un crescente interesse negli ultimi anni. Gli stock commerciali di pesca non sono più sufficienti per soddisfare la domanda globale. Quindi la pesca si effettua sempre di più in profondità, a scapito di specie delle quali sappiamo poco, e questo comporta un grave rischio di estinzione di queste specie.

Dato che è generalmente difficile rilevare le età dei pesci, per implementare politiche di stock management, è necessario creare un modello di crescita tramite il quale si possano inferire le età a partire dai dati sulle lunghezze dei pesci. Tali lunghezze, se non osserviamo le età, provengono da una distribuzione mistura, in cui le componenti sono le singole coorti.

Dato che i metodi MCMC non sono sempre adeguati per l'analisi di modelli mistura, per stimare i parametri e il numero di coorti presenti nel campione viene utilizzato un algoritmo di tipo Population Monte Carlo per misture esteso al caso in cui il numero di componenti è incognito.

Acknowledgements

I would like to thank my supervisors, Professor Stuart Coles, Professor Brunero Liseo and Professor Christian P. Robert, for providing me their support, an interesting and challenging topic, references to relevant literature and for their infinite patience.

Further, I'd like to thank my colleagues and the staff of the Department of Statistical Science, who gave me the opportunity to undertake this interesting experience.

Most of all, I would like to thank my beloved Erika and my parents for bearing me even when I was too deep in thought regarding this work.

Finally, I would like to thank also all those persons that inspired me with their observations and constructive criticisms.

Contents

Chapter 1

Introduction

The thesis is motivated by the analysis of a real dataset concerning a biological problem. It turns out that data come from a particular mixture model. The analysis of mixture models are challenging, and MCMC methods are not always satisfactory. Hence we employ Population Monte Carlo (PMC) methods to estimate the model in a variable dimension setting. A simulated dataset will be used to make comparisons between the proposed method and the Reversible Jump, the most important and widely used method to estimate variable dimension models.

The remaining part of the chapter will illustrate the datasets that will be used; in the second chapter, the mixture model is introduced, reviewing its main features and methods that are commonly used to estimate it. The third chapter reviews the main concepts and results on the Population Monte Carlo (PMC) method. This method will be used in the fourth chapter to propose an algorithm that works in a variable dimension setting and can be useful for the problem at hand. The last chapter proposes another PMC algorithm which is tailored for mixture models in a fixed dimension setting.

Many statistical packages offer effective support to estimate this kind of models, even in the most complicated specifications. The following elabo-

rations are obtained using the package R (see R Development Core Team, 2006).

1.1 Datasets

1.1.1 Fish dataset

The sample contains the lengths and the ages of 1242 individuals of Atlantic Herring (Clupea harengus harengus, Linnaeus, 1758). In fishery, it is typically difficult and expensive to gather the ages of fish. In order to implement stock management policies, it is necessary to build up reliable growth models to infer ages from length data.

Many species of fish spawn in a particular period of the year and growth is very quick at little ages.Hence, younger cohorts are often clearly distringuishable from the rest of the data. Unfortunately, the growth process slows down at older ages, although fish continue to grow for all their lifetime. This cause a great difficulty in distinguish observations from the older cohorts.

Fig. 1.1.1 shows a typical situation: in the left panel, the distributions of lengths are separated by age. The curve represents the von Bertalanffy (1938) growth curve, that is the conditional mean length at age t:

$$\mathbb{E}(y \mid t, L_\infty, k, t_0) = L_\infty(1 - \exp\{-k(t - t_0)\}), \tag{1.1}$$

where t is the age. The other three parameters are biologically meaningful: L_∞ is an "asymptotic" length, that is the mean length of an infinitely old cohort, k controls the speed of growth and depends on the catabolysm and anabolysm of the organisms, finally t_0 can be seen as the time at conception. The right panel is the mixture density arising when age is not observed.

The VB growth equation is almost universally used in fisheries, as it is a good model for the mean length of the cohorts; however, it says nothing

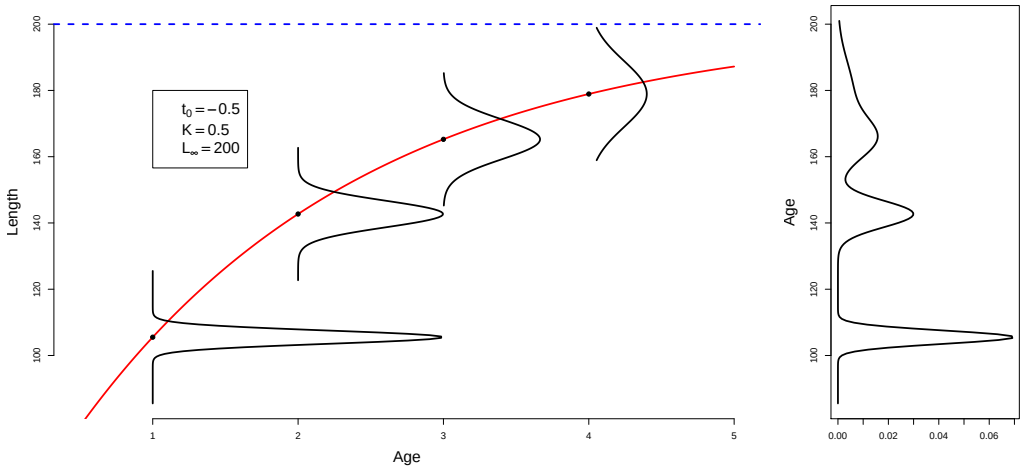

Figure 1.1: Distribution of the observations: separated by age (left panel) and the mixture deriving when age is a latent variable (right panel).

about the distribution of the lengths about their means. Actually, very little is known about these distributions: only recently it is in doubt that the variance of lengths is monotonically growing with age, and nothing is known about higher-order moments.

The most intuitive approach to this kind of problem is to impose a particular parametric specification to the cohorts, but often these specifications are not enough flexible to fit the features of the different component distributions. More recent works (Lv & Pitchford, 2007) try to build a individual model of growth to obtain the distribution of the different cohorts. This approach is promising, but still suffers from different problems. In particular, the authors simply transpose widely known financial models to fishery, even if the growth process behave differently from share pricing processes; besides the resulting models are quite difficult to estimate. Baldi et al. (to be submitted) try to overtake both problems.

Here we will simply assume that the different cohorts are normally distributed around their means; however, we can apply this method even with a more realistic and complex model. However, in this case the dataset is complete, allowing us to make comparisons with complete maximum likelihood estimates.

1.1.2 Simulated dataset

The simulated dataset is represented by $n = 150$ points from a mixture of Normals with the following parameters:

	Component		
	I	II	III
ω	0.3	0.5	0.2
μ	0	5	10
σ^2	1	1	1

The component means are well-separated in order to put in evidence the differences between the estimate of the distribution of the number of components using a PMC and a Reversible Jump.

1.1.3 Galaxy dataset

This dataset collects the velocities in km/sec of 82 galaxies from 6 well-separated conic sections of a survey of the Corona Borealis region. Multimodality in such surveys givess evidence that galaxies form superclusters, which are surrounded by voids. Many authors used this dataset, including Roeder (1990), which also gives a thorough description of the dataset, Venables & Ripley (1994) and Richardson & Green (1997). The different estimates of the number of components in this mixture are in sharp contrast, ranging from three (Roeder & Wasserman, 1997) up to seven (Escobar & West, 1995).

Chapter 2

Mixture models

2.1 Historical notes

From their beginnings, mixture models represent a flexible way to describe non-standard densities. In his seminal paper, Pearson (1893) analyzed a dataset on crabs. As it wasn't possible to use neither a normal distribution, neither the Pearson system of curves, he splitted the distribution mixture of two components. At the time there was no method to estimate that model, so he proposed to equate the first five moments to the respective sample counterparts.

Neyman (1939) discusses the very frequent situation in which usual distribution functions fail to describe real phenomena, making a particular mention to bacteriology and entomology. So he tried to deduce a family of distributions that could give a reasonably good fit of the kind of data he had in mind. Even if these distributions seemed to have a quite specialized field of application, Feller (1943) noticed that these distributions are intimately related with the results that various authors obtained working on totally different topics, as telephone traffic, risk theory, engineering problems and so on. He noticed that in all these cases, the population seems to be formed

by non-homogeneous groups, in the sense that the phenomenon of interest is homogeneous inside each group while is heterogeneous between the groups. Nowadays, mixture models find application in different areas such as cluster analysis, outlier detection, density estimation. Even considering only the most relevant works, it seems impossible to fulfill a complete list. However, the most comprehensive references are Titterington et al. (1985), McLachlan & Peel (2000), Lindsay (1995). A particular mention can be given to a special issue of the journal Computational Statistics and Data Analysis (2003).

2.2 Formal definitions

In what follows, the probability distribution of a random variable is characterized by its probability density function, which is defined wrt an appropriate measure which is, depending on the context, either the Lebesgue measure, a counting measure, or a combination of the two.

Mixture models can arise in different ways: following the missing data approach, we can consider a population made up of c subgroups mixed at random in proportions equal to $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_c\}$. Interest of the research is in the random feature Y, which has the same distribution for all the units that belong to any single group, but has different distributions for each group. Conditionally on the group indicator variable Z, the distribution of Y is $f_Z(y \mid \boldsymbol{\theta}_Z)$; it is generally assumed that the distributions of the different subgroups belong to the same parametric family, so that $f_Z(y \mid \boldsymbol{\theta}_Z) = f(y \mid \boldsymbol{\theta}_Z)$. The joint density of (Y, Z) would be given by $f(y, z) = f(z)f(y \mid z) = \omega_z f(y \mid \boldsymbol{\theta}_z)$. The latent structure arises as the allocation variable is not available; hence, the mixture density is given by

$$f(y \mid \boldsymbol{\vartheta}) = \sum_{j=1}^{c} \omega_j f(y \mid \boldsymbol{\theta}_j), \qquad \sum_{j=1}^{c} \omega_j = 1 \tag{2.1}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\omega}, \boldsymbol{\theta})$. One of the c component weights $\boldsymbol{\omega}$ is determined as

the other $c-1$ are known, so a generic $j^\star$th weight must be replaced with $1-\sum_{l\neq j^\star}\omega_l$.

It is also possible to go through the inverse argument: if Y follows the model (2.1), by a demarginalization argument, it is always possible to create a r.v. Z such that

$$(Y_i \mid Z_i = j) \sim f(y \mid \boldsymbol{\theta}_j) \qquad Z_i \sim \mathcal{M}(1; \omega_1 \ldots, \omega_c)$$

Hence, even if the missing data approach is not always inherent to the structure of data, we can always make use of it as a device to ease the estimation. The model admits many extensions as multivariate mixtures, presence of covariates and Markov switching models. In the thesis only univariate normal mixtures are examined.

Fig. 2.1 shows some different densities obtained varying the parameters of a normal mixture model. It is evident that the model is very flexible, and this flexibility, given a suitable number of components, can be exploited to approximate any unknown distribution, regardless of its features like skewness, kurtosis or even multimodality. The cost of this flexibility is a particular complexity, that can be easily understood at the first glimpse to the likelihood function.

2.3 Likelihood

Having n i.i.d. observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ from the mixture distribution (2.1), if we knew the allocation vector $\boldsymbol{z} = (z_1, \ldots, z_n)$, inference would be based on the complete-data likelihood

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \boldsymbol{y}, \boldsymbol{z}) \propto \prod_{i=1}^{n} f(y_i \mid z_i, \boldsymbol{\vartheta}) f(z_i \mid \boldsymbol{\vartheta}) = \prod_{j=1}^{c} \left(\omega_j^{n_j} \prod_{i:z_i=j} f(y_i \mid \theta_j) \right), \quad (2.2)$$

where $n_j = \sum_{i=1}^{n} \mathbb{I}(z_i = j)$. Hence, it is possible to estimate each component parameter vector separately both in a ML and, using independent priors,

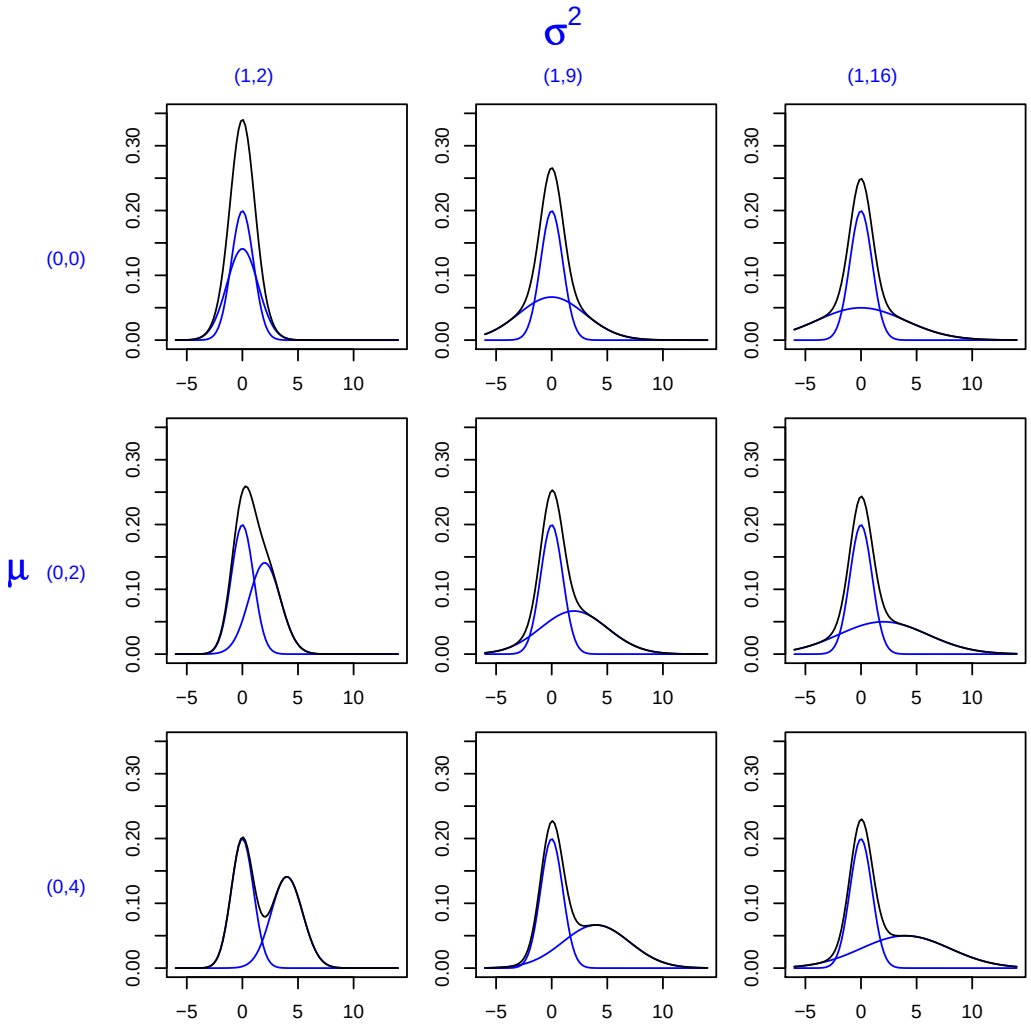

Figure 2.1: Densities of a two component normal mixture obtained varying the component parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. Blue curves are component densities, curves in black are mixture densities.

also in a Bayesian perspective. However, also in this simple case problems could arise: mixture models belong to ill-posed problems, in the sense that small changes in the data can induce large changes in the results. In fact, even for large n, there is a positive probability that no observation comes

from a generic $j^\star$-eth component. In this case, if $\omega_{j^\star}$ is among the free parameters, then the mode of the likelihood function will lie on the boundary of the parametric space, and its ML estimate will be $\hat{\omega}_{j^\star} = 0$. If $\omega_{j^\star}$ is not among the free parameters, the ML estimator will lie in a non-identifiability set corresponding to a mixture with $c-1$ components. The problem of identifiability of mixture density functions has been extensively studied, among others, by Teicher (1960; 1961; 1963). In both cases, the likelihood function is nonregular, and standard asymptotic theory is not valid.

Taking in account that the allocations are not observed, the likelihood function takes the form

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \boldsymbol{y}) \propto \prod_{i=1}^{n} \left(\sum_{j=1}^{c} \omega_j f(y_i \mid \theta_j) \right). \tag{2.3}$$

The pointwise computation of this function requires the evaluation of c^n terms. This highligths the first difficulty that one has to tackle when using a mixture models: as the factorization (2.2) is not valid in this case, any analytical solution via ML or Bayes estimators is precluded, and the computational burden required to estimate the model is generally very high.

2.4 Features of mixture models

Mixture models are characterised by many interesting features ; here we will review only the ones that will be useful for the following elaborations, referring to the cited books for more detailed studies.

2.4.1 Moments

Moments of mixture distributions can be easily found. If we consider a function $h(Y)$ of Y, the expectation $\mathbb{E}(h(Y) \mid \boldsymbol{\vartheta})$ is given by

$$\begin{aligned}\mathbb{E}(h(Y) \mid \boldsymbol{\vartheta}) &= \int_{\mathcal{Y}} h(y) f(y \mid \boldsymbol{\vartheta}) dy = \\ &= \sum_{j=1}^{c} \omega_j \int_{\mathcal{Y}} h(y) f(y \mid \boldsymbol{\theta}_j) dy = \sum_{j=1}^{c} \omega_j \mathbb{E}(h(Y) \mid \boldsymbol{\theta}_j)\end{aligned}$$

provided that $\mathbb{E}(h(Y) \mid \boldsymbol{\theta}_j)$ exists for all j. Substituting $h(Y) = Y$ or $h(Y) = (Y - \mu)^2$ we can simply find the expectation and the variance:

$$\mathbb{E}(Y \mid \boldsymbol{\vartheta}) = \mu = \sum_{j=1}^{c} \omega_j \mu_j \qquad \mathbb{V}\mathrm{ar}(Y \mid \boldsymbol{\vartheta}) = \sum_{j=1}^{c} \omega_j (\mu_j^2 + \sigma_j^2) - \mu^2.$$

For the higher-order moments we have

$$\mathbb{E}(Y^m \mid \boldsymbol{\vartheta}) = \sum_{j=1}^{c} \omega_j \mathbb{E}(Y^m \mid \boldsymbol{\theta}).$$

In particular, for higher order moments around the mean we can use $h(Y) = (Y - \mu)^m$ and the binomial formula:

$$\begin{aligned}\mathbb{E}((Y - \mu)^m \mid \boldsymbol{\vartheta}) &= \sum_{j=1}^{c} \omega_j \mathbb{E}((Y - \mu_j + \mu_j - \mu)^m) \mid \boldsymbol{\theta}) = \\ &= \sum_{j=1}^{c} \sum_{n=0}^{m} \binom{m}{n} \omega_j (\mu_j - \mu)^{m-n} \, \mathbb{E}((Y - \mu_j)^n \mid \boldsymbol{\theta})\end{aligned}$$

2.4.2 Identifiability and label switching

The estimation theory relies for several aspects on the concept of identifiability. A parametric family of distributions $\mathcal{F}$ indexed by a parameter $\psi \in \Psi$ over a sample space $\mathcal{Y}$ is said to be **identifiable** if

$$\psi_1, \psi_2 \in \Psi \text{ and } f(\boldsymbol{y} \mid \psi_1) = f(\boldsymbol{y} \mid \psi_2) \text{ for almost all } \boldsymbol{y} \in \mathcal{Y} \quad \Rightarrow \quad \psi_1 = \psi_2$$

If the family of distributions is not identifiable, any subset $\mathcal{U}(\psi)$ of Ψ defined as

$$\mathcal{U}(\psi) = \{\psi^\star \in \Psi : f(\boldsymbol{y} \mid \psi^\star) = f(\boldsymbol{y} \mid \psi), \text{ for almost all } \boldsymbol{y} \in \mathcal{Y}\}$$

and which contains more than one point is called a *non-identifiability set* (see, for example, Rothenberg, 1971).

It can be noted that the mixture likelihood (2.3) is invariant wrt permutations of the component labels: if $\boldsymbol{\vartheta}^\star$ can be obtained by permuting the indices of the component labels of $\boldsymbol{\vartheta}$

$$f(\boldsymbol{y} \mid \boldsymbol{\vartheta}) = \sum_{j=1}^{c} \omega_j f(\boldsymbol{y} \mid \boldsymbol{\theta}_j) = \sum_{j=1}^{c} \omega_j^\star f(\boldsymbol{y} \mid \boldsymbol{\theta}_j^\star) = f(\boldsymbol{y} \mid \boldsymbol{\vartheta}^\star)$$

for almost all $\boldsymbol{y} \in \mathcal{Y}$ and for every permutation of the labels $\{1, \ldots, c\}$. Although identifiability is generally not of particular concern in Bayesian statistics, this particular feature endowes that the likelihood has up to $c!$ equivalent and symmetric modes, each corresponding to one of the possible ways to label the components. If exchangeable priors are used, the posterior will inherit this feature, which represents a further hamper to the exploration of the posterior distribution by means of the usual strategies. In fact, while ML methods search for one of the local modes of the likelihood, posterior samplers should be able to explore all the different and eventually well-separated modes.

Label switching is another feature which is related with the identifiability. Given the structure of the posterior, all the posterior marginal distributions (and therefore the posterior expectations) will be identical for each group of parameters (weights, means and variances in the normal mixture case). In particular, it can happen that posterior expectations are in a region of low probability. Hence, alternative estimators to posterior expectations or a different prior modeling is required to solve this problem.

These methods are reviewed in Jasra et al. (2005). They can be divided in three groups: identifiability constraints, reordering constraints and loss

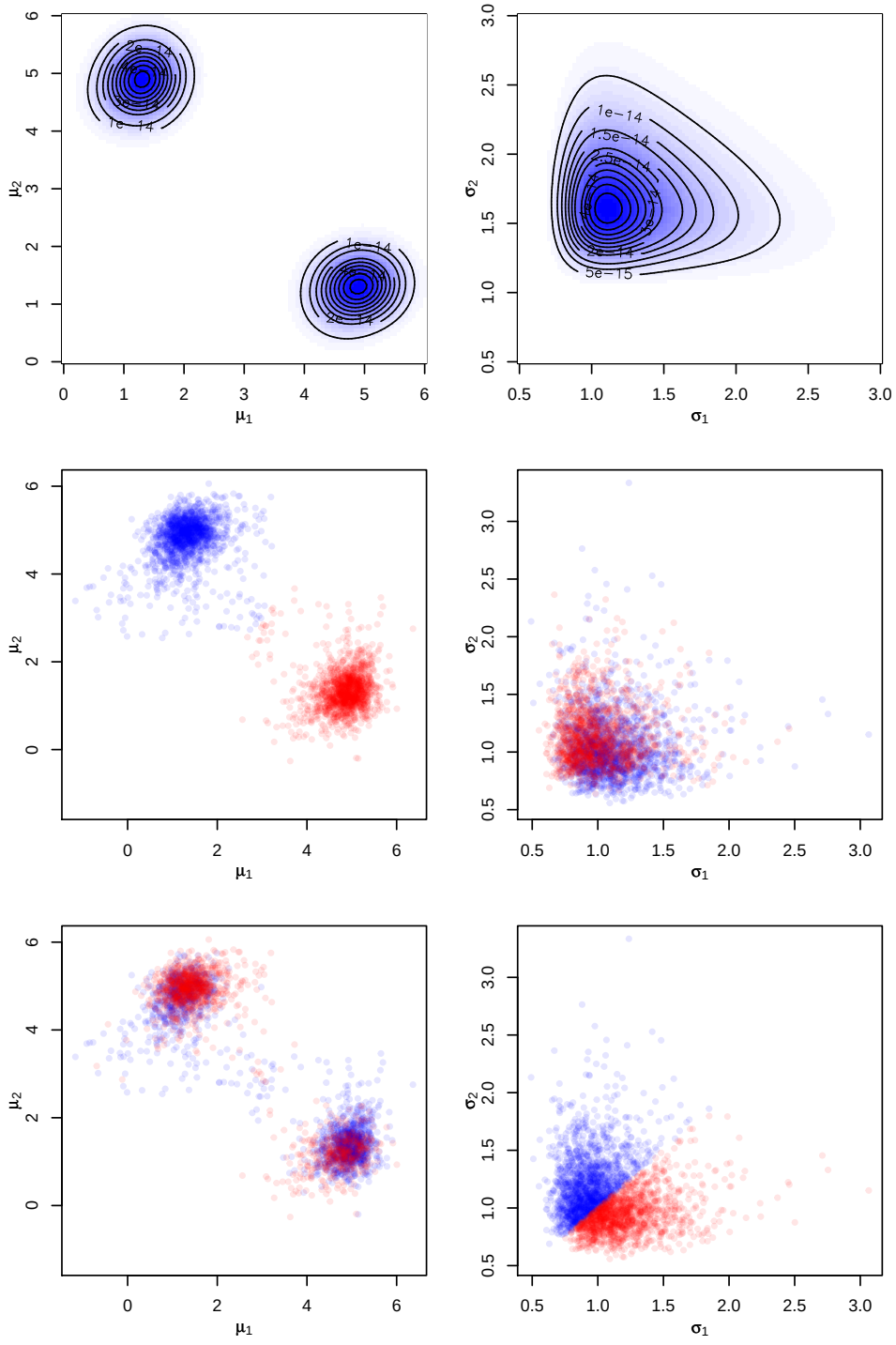

Figure 2.2: Upper panels: posterior density of the means (at the right) and variances (at the left) for the model (2.4), both conditional to the real value of the other parameters. Lower panels: sampling representation of the posterior means (left panels) and variances (right panels). In the second row, points are coloured according to the identifiability constraint $\mu_1 < \mu_2$; in the third row, points are coloured according to the constraint on variances: $\sigma_1 < \sigma_2$.

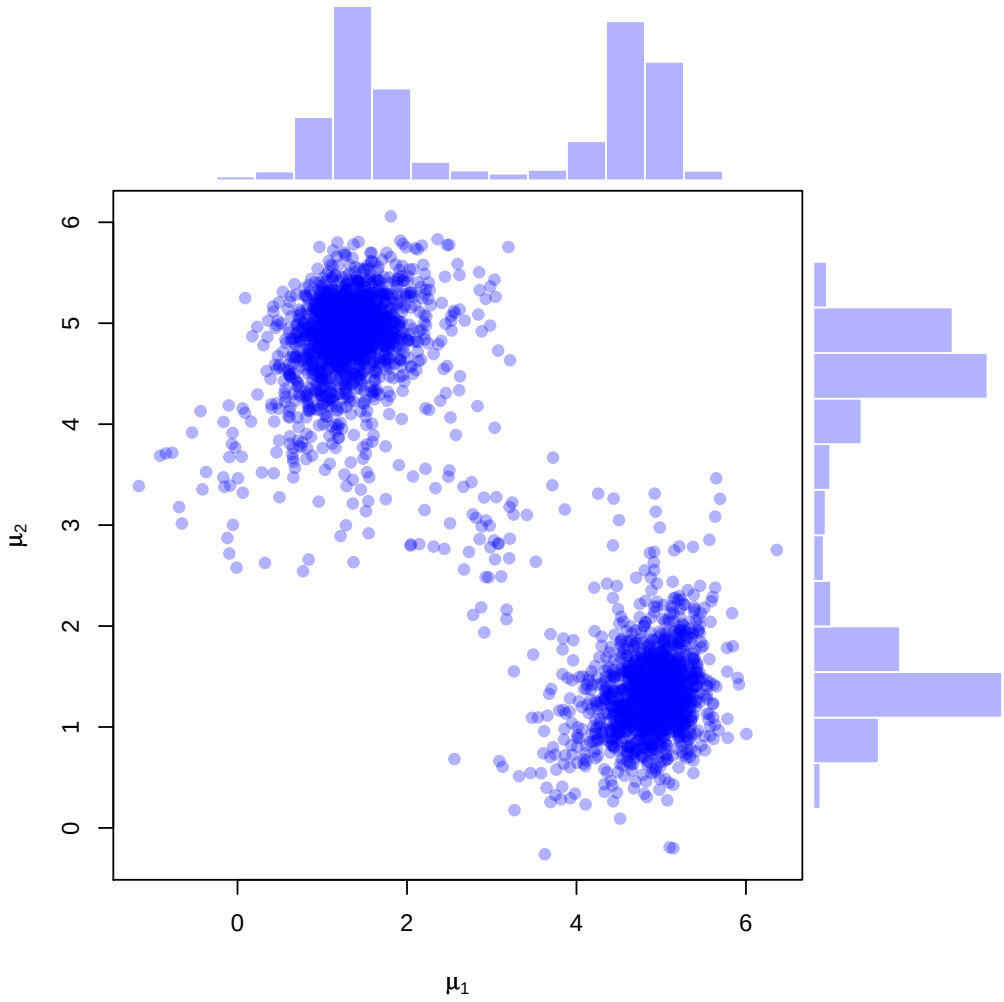

Figure 2.3: Sampling representation of the posterior means of the model (2.4). It can be noticed that the marginal distributions are identical.

functions.

The first method has been commonly used: for example, Richardson & Green (1997) and Mengersen & Robert (1996) use an ordering on the means such that $\mu_1 < \mu_2 < \ldots < \mu_c$, with the aim to single out only one mode, avoiding the problems of identifiability and exploration. The ordering can be imposed

even on weights or variances. The identifiability constraints are equivalent, in a Bayesian framework, to a truncation of the priors over the region that don't satisfy the constraint.
Unfortunately, this constraint not only can fail in its goal to separate one mode from all the others, but can make even more difficult the exploration of the parameter space and the following inference. The truncation of the space don't necessarily respect the topology of the prior and of the likelihood: it is possible that the truncated space contains more than one mode, eventually relegating them at the boundary of the constrained space.

As an example, the upper panels of Fig. 2.2 shows the posterior density of the means and variances for a sample from a two components normal mixture model in which the priors are exchangeable:

$$\begin{aligned} f(y \mid \boldsymbol{\vartheta}) &= \sum_{j=1}^{2} \omega_j \phi(y \mid \boldsymbol{\theta}_j) \\ \pi(\boldsymbol{\vartheta}) &= \pi(\boldsymbol{\omega}) \prod_{j=1}^{2} \pi(\boldsymbol{\theta}_j) \end{aligned} \tag{2.4}$$

Here, $\omega_1 = 0.5$, $\boldsymbol{\mu} = (1, 3)$ and $\boldsymbol{\sigma}^2 = (1, 1)$. In this model, the components have different means but equal weights and variances. Lower panels show some draws from the posterior, obtained with an unconstrained sampler, but coloured according to two different identifiability constraint. In the second row, an identifiability constraint is applied on the means: in this case, the constraint successfully separates the two modes. The third row shows that the constraint on variances cannot separate the modes and no unique labelling is induced by this constraint. In real applications, these constraints must be selected carefully if components have no physical meaning.

It is possible to impose constraints ex-post, only after that the simulation is terminated. Obiously, these constraints doesn't ease the exploration, but their goal is to obtain meaningful estimates. There exist many different

relabelling schemes, but they lead to widely different estimates (see Celeux et al., 2000), mostly because they impose an innatural ordering based on one of the parameters. A schemes that is based on all the parameters can be found in Marin et al. (2005). Denoting with $\boldsymbol{\vartheta}^\star$ the MAP estimate obtained with a posterior sample of size N and with

$$\zeta(\boldsymbol{\vartheta}) = \{\omega_{\zeta(1)}, \ldots, \omega_{\zeta(1)}, \boldsymbol{\theta}_{\zeta(1)}, \ldots, \boldsymbol{\theta}_{\zeta(1)}\}$$

a permutation of the parameter vector $\boldsymbol{\vartheta}$, for any permutation $\zeta \in \mathcal{S}$ of the labels $\{1, \ldots, c\}$, the simulated vectors of parameters are relabelled following the permutation that minimizes their distance from the pivot $\boldsymbol{\vartheta}^\star$.

An alternative can be found on estimators based in loss function which are insensitive to the particular labelling. Defining a permutation invariant loss function $L : \mathcal{A} \times \Theta \to [0, \infty)$ such that $L(a, \boldsymbol{\vartheta}) = L(a, \zeta(\boldsymbol{\vartheta}))$, the idea is to minimize the posterior expected loss

$$\begin{aligned} \mathbb{E}(L(a, \boldsymbol{\vartheta}) \mid y) &= \int_\Theta L(a, \boldsymbol{\vartheta}) \pi(\boldsymbol{\vartheta} \mid y) d\boldsymbol{\vartheta} \\ &\approx \frac{1}{T} \sum_{t=1}^{T} L(a, \boldsymbol{\vartheta}^{(t)}). \end{aligned}$$

where $\boldsymbol{\vartheta}^{(t)}$, $t = 1, \ldots, T$ represent an MCMC sample. This minimization must be performed by means of stochastic optimization methods. The drawbacks of this method are given by the computational cost and by the restricted class of loss functions that can be used.

2.4.3 Prior distributions

The choice of prior distributions is a paramount in modelling mixtures. The main problem is due to the fact that using improper priors will lead to an improper posterior. Intuitively, improper priors bring no information in the

model; if no one of the allocation variables $\boldsymbol{z}$ assume the value j, neither the likelihood will say anything about the parameters of the j-eth component. More formally, if we use independent priors

$$\pi(\boldsymbol{\vartheta}) = \pi(\boldsymbol{\omega}) \prod_{j=1}^{c} \pi(\boldsymbol{\mu}_j)\pi(\boldsymbol{\sigma}_j^2)$$

and they are improper

$$\int_{\Theta_j} \pi(\boldsymbol{\vartheta}_j) d\boldsymbol{\vartheta}_j = \infty.$$

then the posterior will be improper, in fact the integral

$$\int_{\Theta} \pi(\boldsymbol{\vartheta} \mid \boldsymbol{y}) d\boldsymbol{\vartheta} \propto \int_{\Theta} \sum_{\mathcal{Z}} \mathcal{L}(\boldsymbol{\vartheta} \mid \boldsymbol{y}, \boldsymbol{z}) \prod_{j=1}^{c} \pi(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$$

contains $(c-1)!$ elements in which no observation is allocated to the j-eth component. Only imposing a structure between components it is possible to use improper priors (see Mengersen & Robert, 1996).

Many kinds of prior distributions have been proposed in literature: besides the independent priors, the most important are:

- conjugate priors (Diebolt & Robert, 1994):

$$p(\boldsymbol{\vartheta}) = \pi(\boldsymbol{\omega}) \prod_{j=1}^{c} \pi(\sigma_j^2)\pi(\mu_j \mid \sigma_j^2) \tag{2.5}$$

  where $\boldsymbol{\omega} \sim \mathcal{D}(\delta, \ldots, \delta)$, $\mu_j \mid \sigma_j^2 \sim \mathcal{N}(\xi, \sigma_j^2/\lambda)$, $\sigma_j^{-2} \sim \Gamma(\alpha, \beta)$, with α, β, δ, ξ and λ are fixed hyperparameters;

- Hierarchical priors (Richardson & Green, 1997):

$$\pi(\boldsymbol{\vartheta}, \beta) = \pi(\boldsymbol{\omega})\pi(\beta) \prod_{j=1}^{c} \pi(\mu_j)\pi(\sigma_j^2 \mid \beta)$$

  where $\boldsymbol{\omega} \sim \mathcal{D}(\delta, \ldots, \delta)$, $\mu_j \sim \mathcal{N}(\xi, \kappa^{-1})$, $\sigma_j^{-2} \sim \Gamma(\alpha, \beta)$ and $\beta \sim \Gamma(g, h)$, with δ, ξ, κ, α, g and h are fixed hyperparameters.

2.5 Estimating a mixture model

Despite its simple specification, estimating a mixture model is quite an hard task. In the following, only methods based on the likelihood are reviewed, even if many other method have been employed. For example, Pearson (1893) used the method of moments, while in fisheries graphical methods such as that of Bhattacharya (1967) are widely used. Maximum likelihood estimation of mixtures is usually performed using an EM or SEM algorithm, while there is a plethora of different algorithms working in a Bayesian context. Celeux et al. (2000) collects many of these algorithms.

2.5.1 EM algorithm

It is difficult to maximize the likelihood (2.3). In fact, $\partial\mathcal{L}(\boldsymbol{\vartheta} \mid \boldsymbol{y})/\partial\boldsymbol{\omega}$ is a polynomial of degree $(n-1)$ and, even if there is at most one real root, the estimate $\hat{\boldsymbol{\omega}}_{ML}$ may not satisfy the constraint $\omega_j > 0$, $\forall j = 1, \ldots, c$. Even numerical methods such as that of Newton–Raphson or the gradient method can find difficulties, especially when the components are not well-separated and the sample size is small. Besides, the likelihood of a mixture of location-scale distributions is generally unbounded, hence a global maximizer doesn't exists. In fact, if we suppose a normal mixture and we set $\mu_{j^\star}$ equal to any observation y_l, then $\mathcal{L}(\boldsymbol{\vartheta} \mid \boldsymbol{y}) \to \infty$ as $\sigma_{j^\star} \to 0$ (see Kiefer & Wolfowitz, 1956). As a consequence, ML estimators are local, and not global, optimizers.

The most important and widely used method, the EM algorithm, has been introduced by Dempster et al. in 1977. In this paper, the method is used for general latent variable models, but a reference to mixture models was already given.

The EM (Expectation - Maximization) algorithm is a deterministic optimisation procedure based on the missing data representation of the mixture.

The log of the complete-data likelihood (2.2) can be written as

$$\log \mathcal{L}(\boldsymbol{\vartheta} \mid \boldsymbol{y}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{j=1}^{c} D_{ij} \log(\omega_j \, f(y_i \mid \boldsymbol{\theta}_j))$$

where $D_{ij} = 1$ iff $z_i = j$, else $D_{ij} = 0$. Starting from an arbitrary value $\boldsymbol{\vartheta}^{(0)}$, the EM algorithm iterates between two steps:

- the E-step, where the expectation of $\log \mathcal{L}(\boldsymbol{\vartheta} \mid \boldsymbol{y}, \boldsymbol{z})$, conditional on the current vector of parameters, is computed
- the M-step, in which parameters that maximize the expected complete-data log likelihood function are determined to obtain an update $\boldsymbol{\vartheta}^{(t)}$.

Under mild regularity conditions, the EM converges to a local maximum of the likelihood function; it may, however, fail to converge to the "right" mode of the likelihood remaining trapped in a spurious one. For any mixture model, at the m-th iteration, the E-step leads to the estimate for D_{ij}

$$D_{ij}^{(m)} = \frac{\hat{\omega}_j^{(m-1)} p(y_i \mid \hat{\theta}_j^{(m-1)})}{\sum_{j=1}^{c} \hat{\omega}_j^{(m-1)} p(y_i \mid \hat{\theta}_j^{(m-1)})}$$

and the M-step to the estimates

$$\omega_j^{(m)} = \frac{n_j}{n} \qquad\qquad n_j = \sum_{i=1}^{n} \hat{D}_{ij}^{(m)}.$$

The estimator of the component parameters θ_j will depend on the distribution family underlying the mixture; for normal mixtures

$$\begin{aligned} \mu_j^{(m)} &= \frac{1}{n_j} \sum_{i=1}^{n} \hat{D}_{ij}^{(m)} y_i \\ (\sigma_j^2)^{(m)} &= \frac{1}{n_j} \sum_{i=1}^{n} \hat{D}_{ij}^{(m)} (y_i - \mu_j^{(m)})^2. \end{aligned}$$

2.5.2 SEM algorithm

The SEM (Stochastic EM, Celeux & Diebolt, 1985) algorithm is a stochastic version of EM incorporating an S step between the E and M steps a restoration of the allocation variables $\boldsymbol{z}$, by drawing them at random from their current predictive distribution $f(\boldsymbol{z} \mid \boldsymbol{\vartheta}^{(t-1)}, \boldsymbol{y})$.

SEM algorithm doesn't converge pointwise. It generates a Markov chain whose stationary distribution is more or less concentrated around the ML parameter estimator. Also in this case, the natural parameter estimate is the parameter vector leading to the maximum value of the likelihood.

2.6 The Bayesian approach

In this section, the most important algorithms used to explore the posterior density of mixture models will be reviewed. More details can be found in Robert & Casella (2004) and in Celeux et al. (2000).

2.6.1 Gibbs sampler with data augmentation

In 1987, Tanner & Wong proposed an iterative algorithm that can be useful where the parameter space can be augmented in such a way that it is easy to generate values from the augmented space given the parameters. Hence, also in the Bayesian approach, it turns out that the missing data approach can be a useful device to simplify the estimation algorithms. In its most standard form, used for example by Diebolt & Robert (1994), starting from an arbitrary value of $(\boldsymbol{\vartheta}^{(0)}, \boldsymbol{z}^{(0)})$, the method consists in iterating the simulations of $\boldsymbol{\vartheta}$ and $\boldsymbol{z}$ from the respective full conditional distributions. A pseudo code for normal mixture model is as follows:

At $t = 0$

choose an arbitrary point of $(\boldsymbol{\vartheta}^{(0)}, \boldsymbol{z}^{(0)})$

For $t = 1, \ldots, T$

(a) update $\boldsymbol{\omega}^{(t)} \sim \mathcal{D}(\delta^\star, \ldots, \delta^\star)$

(b) update $(\boldsymbol{\sigma}_j^{-2})^{(t)} \sim \Gamma(\alpha^\star, \beta^\star), \qquad j = 1, \ldots, c$

(c) update $\boldsymbol{\mu}_j^{(t)} \sim \mathcal{N}(\xi^\star, (\kappa^{-1})^\star), \qquad j = 1, \ldots, c$

For $i = 1, \ldots, n$

(d) update $z_i^{(t)}$, where $\mathbb{P}(z_i^{(t)} = j) \propto \frac{1}{\sqrt{2\pi(\sigma_j^2)^{(t)}}} \exp\left\{-\frac{(y_i - \mu_j^{(t)})^2}{2(\sigma_j^2)^{(t)}}\right\}$

The specification of the updated hyperparameters $\alpha^\star$, $\beta^\star$, $\delta^\star$, $\kappa^\star$ and $\xi^\star$ depend on the prior distributions.

This sampler can have difficulties in reaching convergence: the local modes can represent almost trapping states for the chain, and it could take a great amount of iterations to escape from each of them. If only few observations are allocated in the $j^\star$-eth component, the probabilities that an observation will enter in that component, as the probability that an observation allocated in $j^\star$ can escape, become small. In Robert & Casella (2004) and Guihenneuc-Jouyaux et al. (1998) can be found the descriptions of various convergence diagnostics.

2.6.2 Permutation sampler

Not only the complete exploration of the posterior is required for the convergence of the MCMC method, poor estimates can be obtained in case of an unbalanced label switching. It is in fact required to the Markov chain to

stay an approximately equal amount of iteration in each mode. In this sense, the label switching is essential for the convergence of the chain. Frühwirth-Schnatter (2001) propose to force the Gibbs sampler chain to jump from mode to mode. Essentially, at the end of each iteration, a random permutation of the labels is drawn. This permutation is applied to the component parameters and to the allocation vector. A pseudo-code is as follows:

At $t = 0$

choose an arbitrary point of $(\boldsymbol{\vartheta}^{(0)}, \boldsymbol{z}^{(0)})$

For $t = 1, \ldots, T$

perform steps (a) to (d) of the Gibbs sampler

sample a permutation $\zeta \sim U(1, \ldots, c!)$ and apply it to $\boldsymbol{\vartheta}^{(t)}$ and $\boldsymbol{z}^{(t)}$

2.7 Model selection

The most relevant source of uncertainty in the mixture model specification is in assuming a particular number of components. If there is no reason to assume a particular number c of components, we have to estimate it from the data. In this chapter the model choice will be treated as the choice of the number of components between the alternative models $\mathcal{M}_1, \ldots, \mathcal{M}_{c_{\max}}$.

In a Bayesian perspective, model uncertainty is usually dealt with Bayes factors. For istance, the evidence of model 1 against model 2 is given by the logarithm of the quantity

$$B_{12} = \frac{f(\boldsymbol{y} \mid \mathcal{M}_1)}{f(\boldsymbol{y} \mid \mathcal{M}_1)} = \frac{\int_{\Theta_1} f(\boldsymbol{\vartheta} \mid \mathcal{M}_1) f(\boldsymbol{y} \mid \boldsymbol{\vartheta}, \mathcal{M}_1) d\boldsymbol{\vartheta}}{\int_{\Theta_2} f(\boldsymbol{\vartheta} \mid \mathcal{M}_2) f(\boldsymbol{y} \mid \boldsymbol{\vartheta}, \mathcal{M}_2) d\boldsymbol{\vartheta}}.$$

Many tecniques have been proposed in order to estimate the marginal likelihoods appearing in this ratio, and in particular the bridge sampling (Meng & Wong, 1996) and the Chib's (1995) approximation to the posterior density ratio. However, all these methods find difficulties with mixture models due to the features of the posterior distribution.

It is also possible to tackle this problem in a trans-dimensional setting, considering simultaneously all the different competing models. If we denote with $\boldsymbol{\vartheta}_{(c)}$ a parameter belonging to Θ_c, the parameter space of the model with c components, the model is specified by a sampling distribution $f(\boldsymbol{y} \mid \boldsymbol{\vartheta}_{(c)}, \mathcal{M}_c)$, a prior distribution on the parameters $\pi(\boldsymbol{\vartheta}_{(c)} \mid \mathcal{M}_c)$ and a prior distribution for the number of components $\pi(\mathcal{M}_c)$. The posterior densities that arise are obviously quite difficult to explore, and the most popular algorithm used for this purpose is the Reversible Jump (RJ, Green, 1995), a trans-dimensional MCMC method.

Theorically, the two approaches are equivalent, as both of them try to estimate (at least implicitly) the posterior distribution of the indices of the competing models.

2.7.1 Reversible Jump

Richardson & Green (1997) describe the Reversible Jump for mixture models. RJ is an extension of the Metropolis–Hastings algorithm, in the sense that it allows the chain to make moves between couples of models $(\mathcal{M}_i, \boldsymbol{\vartheta}_{(i)})$ and $(\mathcal{M}_j, \boldsymbol{\vartheta}_{(j)})$. It creates a chain that moves around $\Theta = \bigcup_{c=1}^{c_{\max}} (\mathcal{M}_c \times \Theta_c)$, the whole parameter space formed by the union of the parameter spaces of the single submodels. The RJ sweeps around these moves:

At $t = 0$

choose an arbitrary point $(m^{(0)}, \boldsymbol{\vartheta}^{(0)}, \boldsymbol{z}^{(0)})$

For $t = 1, \ldots, T$

(1) update $\boldsymbol{\vartheta}^{(t)}, \boldsymbol{z}^{(t)}$ and the eventual random hyperparameters as in the steps (a) - (d) of the Gibbs sampler

(2) split / merge move

(3) birth / death move

The dimension preserving moves are performed using full conditional distributions. For dimension chaging moves, to mantain the detailed balance condition, a bijection is created between the two spaces $(\boldsymbol{\vartheta}_{(i)}, \boldsymbol{u}_i)$ and $(\boldsymbol{\vartheta}_{(j)}, \boldsymbol{u}_j)$, where $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ are sets of artificial variables created to match the dimension of the two spaces.

Split and birth moves require the choice of a matching function $\boldsymbol{\vartheta}_{(c+1)} = g_{c,c+1}(\boldsymbol{\vartheta}_{(c)}, \boldsymbol{u}_c)$ and of a proposal density $q_{c,c+1}(\boldsymbol{u})$ to propose moves from the model $\mathcal{M}_c$ to $\mathcal{M}_{c+1}$. This proposal must form a reversible pair with the proposal for a merge (or a death, respectively) move $(\boldsymbol{\vartheta}_{(c)}, u) = g^{-1}_{c,c+1}\boldsymbol{\vartheta}_{(c+1)}$.

The main advantage of the RJ is that it has less "local" moves than a MCMC running in a fixed dimension setting, leading to better estimates of the single sub-models. In fact, allowing the chain to jump between models, it can visit all the modes of the model with $j^\star$ components by jumping in adjacent models and then returning back to a different region of the $j^\star$-eth parameter space. The main disadvantage is the difficulty in choosing the the proposal densities q and of the matching function g, and these choices are fundamental for the efficiency of the algorithm. Besides, in the particular case of mixture models it is questionable if a RJ, as any MCMC method, can really find convergence (see Celeux et al., 2000).

Chapter 3

Population Monte Carlo algorithms

In recent years, we assist to a growing interest in adaptive sampling schemes. This is due to the fact that many MCMC algorithms often require a fine tuning of several parameters of the proposal distribution, as we have seen with the RJ. Hence, one possible solution is to construct algorithms that automatically learn about the optimal values of these parameters. It is not a simple matter to design an adaptive MCMC algorithm, as adaptivity is in contrast with the Markovianity of the chain. There are many solutions to this problem, but the simpler is to stop adapting while the chain is stil in the burn-in period[1].

Population Monte Carlo (PMC) algorithms (Cappé et al., 2004) are essentially iterated sampling importance resampling algorithms (Rubin, 1987). They are not based on convergence arguments, hence adaptivity is easily obtained by changing the proposal distributions over iterations on the basis of

[1] A simple but effective example that illustrates the perils of naïve use of adaptive MCMC algorithms can be found in the website of Prof. Rosenthal: http://www.probability.ca/jeff/java/adapt.html.

past performances of the sampler. This will not jeopardize the (approximate) unbiasedness of the method, as PMC is based on the importance sampling identity. Moreover, there is no need of a burn-in period or of a stopping rule. This kind of algorithms offer a greater freedom in choosing the proposal distributions, being also possible to draw from the experience gained in the MCMC setting.

3.1 Importance Sampling algorithms

Suppose we want to estimate the quantity $\mathfrak{I}$:

$$\mathfrak{I} = \int_{\Theta} h(\theta)\pi(\theta)d\theta.$$

where π is a normalized posterior distribution. The quantity $\mathfrak{I}$ can represent any feature of the posterior distribution, as its mean, its variance or a quantile. Let's also suppose that the integral is analytically intractable; if we could obtain a sample $\theta^{(1)}, \ldots, \theta^{(N)}$ from π, we would simply estimate $\mathfrak{I}$ as

$$\hat{\mathfrak{I}} = \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)})$$

but generally π is complicated and difficult to sample from. We can however rewrite this quantity, using the *importance sampling identity*, as

$$\mathfrak{I} = \int_{\Theta} h(\theta)\frac{\pi(\theta)}{q(\theta)}q(\theta)d\theta \tag{3.1}$$

as long as $\pi \ll q$. The distribution q will be called "proposal density". Its specification is somehow arbitrary, in the sense that one can choose it between those distribution which satisfy the requirements that will be specified below and that are simple to sample from, even if the properties of the sampler will strongly depend on this choice. The idea is to obtain a sample from q and

estimate $\mathfrak{I}$ as

$$\hat{\mathfrak{I}}_N = \frac{1}{N}\sum_{i=1}^{N} \rho(\theta^{(i)})h(\theta^{(i)}), \tag{3.2}$$

where $\rho_i = \rho(\theta^{(i)}) = \pi(\theta^{(i)})/q(\theta^{(i)})$ is called importance weight. In this way, importance sampling assigns more weight to those particles $\theta^{(i)}$ for which $\pi(\theta^{(i)}) > q(\theta^{(i)})$ and less weight to those particles for which $\pi(\theta^{(i)}) < q(\theta^{(i)})$ in order to estimate $\mathfrak{I}$ correctly. Under weak assumptions, by the strong law of large numbers,

$$\hat{\mathfrak{I}}_N \to \pi(h).$$

Besides, if we suppose that the variances $\mathbb{V}\text{ar}(\rho(\theta^{(i)})h(\theta^{(i)}))$ exist, we have

$$\mathbb{V}\text{ar}(\hat{\mathfrak{I}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{V}\text{ar}(\rho(\theta^{(i)})h(\theta^{(i)}))$$

In most cases, π is not normalized; in this case, we will estimate $\mathfrak{I}$ as

$$\hat{\mathfrak{I}}_N^{IS} = \frac{\sum_{i=1}^{N} \rho(\theta^{(i)})h(\theta^{(i)})}{\sum_{i=1}^{N} \rho(\theta^{(i)})},$$

Under the following assumptions (see Geweke, 1989), the law of large numbers still holds:

1: π is proportional to a proper probability density function $\bar{\pi}$ defined on Θ;

2: $\{\theta^{(i)}\}_{i=1}^{\infty}$ is a sequence of i.i.d. random particles, the common distribution having a probability density function q;

3: The support of q includes Θ;

4: $\pi(h)$ exists and is finite.

but the decomposition of the variance holds only approximately. In this case, if we suppose that the quantities

$$\mathbb{E}(\rho(\theta)) = c^{-1}\int_{\Theta} \frac{\pi(\theta)^2}{q(\theta)} d\theta$$

and

$$\mathbb{E}(\mathfrak{I}^2\rho(\theta)) = c^{-1}\int_\Theta \mathfrak{I}^2\pi(\theta)^2/q(\theta)d\theta$$

are finite and denoting with

$$\begin{aligned}\sigma^2 &= \mathbb{E}((\hat{\mathfrak{I}}_N^{IS}-\mathfrak{I})^2\rho(\theta)) = c^{-1}\int_\Theta (\hat{\mathfrak{I}}_N^{IS}-\mathfrak{I})^2\rho(\theta)\pi(\theta)d\theta \\ \hat{\sigma}_N^2 &= \frac{\sum_{i=1}^N((h(\theta^{(i)})-\hat{\mathfrak{I}}_N^{IS})^2\rho(\theta^{(i)})^2)}{\left(\sum_{i=1}^N \rho(\theta^{(i)})\right)^2}\end{aligned}$$

then

$$N^{1/2}(\hat{\mathfrak{I}}_N^{IS}-\mathfrak{I}) \rightarrow \mathcal{N}(0,\sigma^2) \tag{3.3}$$

$$N\hat{\sigma}_N^2 \rightarrow \sigma^2 \tag{3.4}$$

where c is the normalization constant of π.

IS-based algorithms are simple and effective, but they are strongly affected by the choice of the proposal density. A good proposal leads to excellent results in a reasonable time. Unfortunately we have only a little guidance in the choice of q, and even seemengly reasonable choices can result in a bad behaviour of the sampler and the failure of the convergence of $\hat{\mathfrak{I}}$ to its correct value. Rubinstein (1981) stated that the minimum of σ^2 can be reached choosing the importance function

$$q(\theta) \propto |h(\theta)|\pi(\theta).$$

This theorem gives an optimal proposal distribution, in the sense that the estimator has zero variance. Unfortunately, this result is of little practical interest, as its normalization constant is similar to the quantity we are trying to estimate. It, however, highlights the need to tailor the proposal distribution on the features of the target distribution.

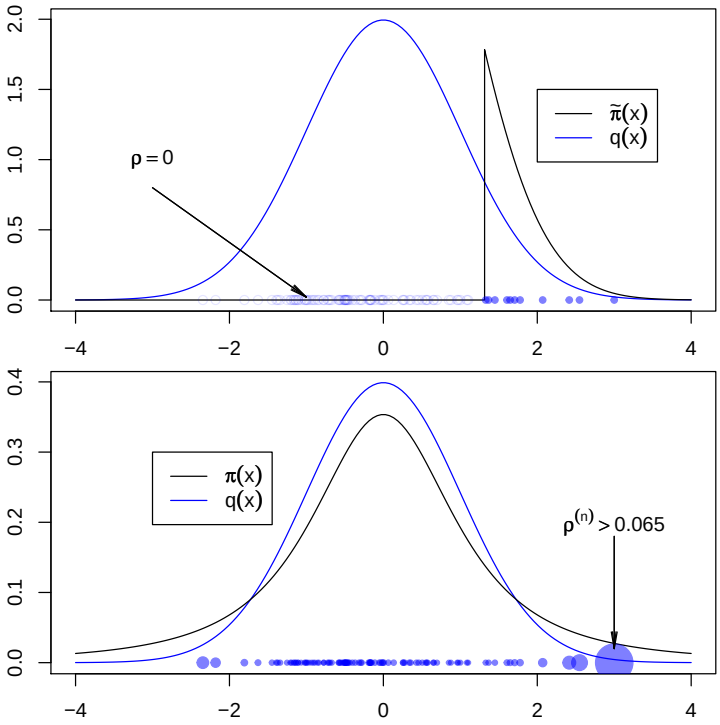

Figure 3.1: Requirements of the proposal density: the upper panel shows a proposal density (in blue) which doesn't adequately overlap with the posterior distribution (in black), the lower panel shows a proposal with tails lighter than those of the posterior. The diameter of the cirles are proportional to the importance weights.

However, the expression of σ^2 suggests that it is adversely affected by large values of $\mathbb{V}\mathrm{ar}(h(\theta))$ and of importance weights. While the first cannot be modified, the latter aspect can be controlled with a good choice of the proposal distribution: it should concentrate its mass on the important part(s) of the target density and it should have tails heavier than those of the target. These requirements are in contrast, as the heavier are the tails of a distribution, the less it can be concentrated. The first requirement can be illustrated by the upper panel of Fig. 3.1: while the proposal is a normal distribution, the posterior is a truncated normal; the support of q includes Θ, but only the tails of the proposal "cover" the the support of the poste-

rior, with negative consequences on the efficiency of the algorithm. Particles generated below the truncation of the posterior will get a null importance weight and will not enter in the calculation of $\mathfrak{I}$.

At the same time, unimportant parts of the posterior should not be neglected. First of all, the identity (3.1) requires that π must be absolutely continouous wrt q. In fact, if Θ could be expressed as the union of Θ_1 and Θ_2, where the first represents the support common to π and q and the second represents the support only of π, the estimate $\hat{\mathfrak{I}}_N^{IS}$ would converge to

$$\int_{\Theta_1} h(\theta)\pi(\theta)d\theta$$

which is different from $\mathfrak{I}$. Besides, results (3.3) and (3.4) hold only if $\mathfrak{I}$ and σ^2 are finite. In particular, $\sigma^2 < \infty$ implies that the proposal has tails that decay more slowly than π. In fact, if this requirement is not satisfied, as in the lower panel of fig. 3.1, the ratio π/q is unbounded, hence particles representing extreme values for q will obtain extremely high importance weigths. Whenever these extreme values are generated, the estimate of $\mathfrak{I}$ may oscillate rather than converge to the correct value.

As an example, in Yuan & Druzdzel (2005) it is shown the path of the variance of the estimator in the a case in which both the proposal and the posterior distributions are normals with different means (μ_q and μ_π) and variances (σ_q^2 and σ_π^2), for different choices of the parameters of the proposal. This pattern highlights that, in this case, the variance of the estimator depends crucially on the ratio σ_q/σ_π: in fact, as this ratio decreases under a certain threshold, which depends also on $\mu_q - \mu_\pi$, there is a sudden groeth of this variance.

3.2 Population Monte Carlo algorithms

In real applications it is often difficult to obtain good proposal distributions, and the difficulty increases with the number of parameters. Due to this problem, more automatical recipes, as MCMC algorithms, have been preferred to IS. Only recently, adaptive methods received an increasing interest.

While MCMC methods hardly can reach an adaptive perspective, PMC algorithms iterate the Importance Sampling step in order to exploit the past performance of the sampler to find increasingly better proposal distributions. It is important to stress that it is possible to use different proposals for each particle and for each iteration mantaining the approximative unbiasedness of the resulting estimators (see Cappé et al., 2004).

This kind of algorithms can be affected by the degeneracy phenomenon, that happens when a few particles have large importance weights, with negative consequences on the variance of the estimates. To eliminate irrelevant particles and alleviate the degeneracy phenomenon, as in the SIR algorithm of Rubin (1987), at the end of each iteration particles are multinomially resampled, with weights equal to $\boldsymbol{\rho}$. A PMC pseudo-code could be as follows:

At $t = 0$

choose arbitrary values of $\theta^{(1:N,0)}$

For $t = 1, \ldots, T$

for $i = 1 \ldots, N$

sample $\tilde{\theta}^{(i,t)} \sim q_{it}(\theta)$

compute $\bar{\rho}^{(i,t)} = \pi(\tilde{\theta}^{(i,t)})/q(\tilde{\theta}^{(i,t)})$

normalize $\rho^{(1:N,t)}$ to sum up to 1

generate $(J^{(i,t)})_{1 \leq i \leq N} \overset{iid}{\sim} \mathcal{M}(1, \bar{\rho}^{(1:N,t)})$

compute intermediate estimates $\mathfrak{I}_t^{PMC}$

set $\theta^{(1:N,t)} = \tilde{\theta}^{(J,t)}$

At the end of each iteration, we can compute intermediate estimates

$$\hat{\mathfrak{I}}_t^{PMC} = \sum_{i=1}^{N} h(\theta^{(i,t)})\rho^{(i,t)};$$

after resampling, irrelevant particles have been removed from the sample. At the next iteration, it is possible to modify proposal densities using the informations about the entire history of the sampler, and simulate the particles $\theta^{(i,t+1)}$ from different proposal distributions q_{it}. After T iterations, an asymptotically unbiased estimator for $\mathfrak{I}$ is given by

$$\hat{\mathfrak{I}}^{PMC} = \frac{1}{T}\sum_{t=1}^{T} \mathfrak{I}_t^{PMC}.$$

3.3 An algorithm for mixture models

In Celeux et al. (2006) the authors propose a PMC algorithm for missing data models. The steps can summarized as follows:

At $t = 0$

choose arbitrary values of $(\boldsymbol{\vartheta}^{(1:N,0)}, \boldsymbol{z}^{(1:N,0)})$

For $t = 1, \ldots, T$

for $i = 1 \ldots, N$

generate $\tilde{\boldsymbol{z}}^{(i,t)} \sim k(\boldsymbol{z} \mid \boldsymbol{\vartheta}^{(i,t-1)}, \boldsymbol{y})$

generate $\tilde{\boldsymbol{\vartheta}}^{(i,t)} \sim \pi(\boldsymbol{\vartheta} \mid \boldsymbol{z}^{(i,t)}, \boldsymbol{y})$

compute $\bar{\rho}^{(i,t)} = \pi(\tilde{\boldsymbol{\vartheta}}^{(i,t)}, \tilde{\boldsymbol{z}}^{i,t})/k(\tilde{\boldsymbol{z}} \mid \boldsymbol{\vartheta}^{(i,t-1)}, \boldsymbol{y})\pi(\tilde{\boldsymbol{\vartheta}} \mid \boldsymbol{z}^{(i,t)}, \boldsymbol{y})$

normalize $\rho^{(1:N,t)}$ to sum up to 1

compute intermediate estimates $\mathfrak{I}_t^{PMC}$

generate $(J^{(i,t)})_{1\leq i\leq N} \overset{iid}{\sim} \mathcal{M}(1, \bar{\rho}^{(1:N,t)})$

set $\theta^{(1:N,t)} = \tilde{\theta}^{(J,t)}$

Here the proposals are represented by the full conditional distributions. Even if the competition between particles guarantees a better performance than the Gibbs sampler, the algorithm can suffer from the degeneacy phenomenon due to the completion of the parameter space. In the same work it is proposed to use an idea similar to the Rao-Blackwellization (Gelfand & Smith, 1990): the additional randomness introduced by the generation of missing data can be eliminated by considering the marginal proposal density of $\boldsymbol{\vartheta}^{(i,t)}$ given $\boldsymbol{\vartheta}^{(i,t-1)}$:

$$\int_{\mathcal{Z}} \pi(\boldsymbol{\vartheta} \mid \boldsymbol{z}, \boldsymbol{y}) k(\boldsymbol{z} \mid \boldsymbol{\vartheta}^{(i,t-1)}) d\boldsymbol{z}.$$

Instead of approximate this integral simulating an entire vector of missing data for each iteration, it is possible to use a pre-simulated set of $\boldsymbol{z}$ and correct for their sampling distribution. This approximation can be used twice in the computation of the importance weights. This method reduces the degeneracy phenomenon, at the cost of an higher computational burden.

Chapter 4

A trans-dimensional algorithm

The algorithm is essentially an extension to the variable dimension case of the Alg. 1 described in Celeux et al. (2006).

Also in this case, the possibility to jump between models opens different problems. The main difficulty here is not to mantain the detailed balance condition, but the absolute continuity of the target distribution wrt the proposal distribution. This means that each particle, from one iteration to the next, should be able to jump from any other point of any parameter subspace Θ_j to any point of the whole parametric space Θ. Hence, we have to propose jumping moves that allow for jumps between any couple of models. Hence, we cannot restrict to consider only jumps between adjacent models, as in the MCMC case.

A particular issue is in determination of the probabilities $p(c_i \rightarrow c_j)$ to jump between models. When there are seveal competing models, it is not possible to exploit the adaptativity to learn about the entire matrix of probabilities to jump between each couple of models, as it would require a very large number of particles. The natural alternative is to apply a probability distribution which is centered on the current model and which has probabilities that decrease with the "distance" between the current and the proposed

model. These probabilities, however, cannot decrease in a very fast way (such as at a geometric rate), as the tails of the proposal distribution of c should be thicker than the tails of the posterior ditribution of the number of components. In the application, $p(c_i \to c_j)$ gives probability p to remain in the current model and divides the remaining probability mass over the other $(c_{\max} - 1)$ models.

4.1 A pseudo-code

At $t = 0$

generate N particles $(c^{(1:N,0)}, \boldsymbol{\vartheta}^{(1:N,0)}, \boldsymbol{z}^{(1:N,0)})$ from the prior distributions;

for $t = 1, \ldots, T$

for $i = 1 \ldots, N$

choose $\tilde{c}^{(i,t)} \sim \mathcal{M}(1, p(c^{(i,t-1)}, \cdot))$

if $\tilde{c}^{(i,t)} = c^{(i,t-1)}$ (1)

update $\boldsymbol{\vartheta}^{(i,t)}$ and $\boldsymbol{z}^{(i,t)}$ from the respective full conditional distributions

compute $\bar{\rho}^{(i,t)} = \pi(c^{(i,t)}, \boldsymbol{\vartheta}^{(i,t)}, \boldsymbol{z}^{(i,t)})/(p(c^{(i,t-1)} \to c^{(i,t)})q(\boldsymbol{\vartheta}^{(i,t)})k(\boldsymbol{z}^{(i,t)})$

normalize the importance weights: $(\rho^{(i,t)})_{1<i<N} = (\bar{\rho}^{(i,t)})_{1<i<N} / \sum_i \bar{\rho}^{(1:N,t)}$

compute intermediate estimates $\mathfrak{I}_t^{PMC}$

generate $(J^{i,t})_{1 \le i \le N} \overset{iid}{\sim} \mathcal{M}(1, (\rho^{(i,t)})_{1 \le i \le N})$

set $(c^{(i,t)}, \boldsymbol{\vartheta^{(i,t)}}, \boldsymbol{z}^{(i,t)}) = (\tilde{c}^{(J_i,t)}, \tilde{\boldsymbol{\vartheta}}^{(J_i,t)}, \tilde{\boldsymbol{z}}^{(J_i,t)})$

In these steps, q and k are given by the full conditional distributions. While in a Gibbs sampler the parameters could be updated in any sequence, even following a random sequence, in this algorithm the sequence is fixed in order to reduce the computational effort and to avoid the use of arbitrary constants.

When required, the dimension changing moves can be exploited before the step (1). The only move to reduce the number of components is similar to the merge move of the RJ. If $\tilde{c}^{(i,t)} < c^{(i,t-1)}$ we have to merge $d = \tilde{c}^{(i,t)} - c^{(i,t-1)}$ components. We can choose which components to merge generating an auxiliary variable $u \sim U(1, \tilde{c}^{(i,t)})$ distribution, and imposing to merge the components $\{u, \ldots, u+d\}$, that is a group of adjacent components. It is, in fact, important to note that the density of jumping moves is given by

$$p(c^{(i,t-1)} \rightarrow \tilde{c}^{(i,t)}) \sum_{\mathcal{U}} q(\tilde{\boldsymbol{\vartheta}}_{\star}^{(i,t)} \mid u) k(\boldsymbol{z}^{(i,t)} \mid u) f(u),$$

where $\mathcal{U}$ denotes all the possible groupings of d different components. The choice to merge only adjacent components reduces the number of elements in $\mathcal{U}$ to d.

In a manner which is similar to the RJ merge move, the old component parameters will be replaced by

$$\omega_{\star} = \sum_{j=u}^{u+d} \omega_j^{(i,t-1)} \qquad\qquad w_{\star}\mu_{\star} = \sum_{j=u}^{u+d} \omega_j^{(i,t-1)} \mu_j^{(i,t-1)}$$

$$\omega_{\star}(\mu_{\star}^2 + \sigma_{\star}^2) \;\; = \sum_{j=u}^{u+d} \;\; \omega_j^{(i,t-1)} ((\mu_j^{(i,t-1)})^2 + (\sigma_j^{(i,t-1)})^2)$$

These parameters values can be updated as follows:

$$\begin{aligned} \mu_u &\sim \mathcal{N}(\mu_{\star}, \sigma_{\star}^2) \\ \sigma_u^2 &\sim \Gamma(\tilde{c}^{i,t}\sigma_{\star}^2, \tilde{c}^{i,t}). \end{aligned}$$

The remaining parameters will be updated in the following steps.

To increase the number of components we propose d further weight equal to $1/\tilde{c}^{(i,t)}$. So we generate d variates $\boldsymbol{u}$ from a multinomial distribution on $1, \ldots, c^{(i,t-1)}$ with probabilities equal to $\omega^{(i,t-1)}$; the new means and variances will be sampled as:

$$\begin{aligned} \tilde{\mu}_{1:d}^{(i,t)} &\sim \mathcal{N}(\mu_{u_1:u_d}^{(i,t-1)}), (\sigma^2)_{u_1:u_d}^{(i,t-1)}) \\ \tilde{\sigma}_{1:d}^{(i,t)} &\sim \Gamma(\tilde{c}^{(i,t)}\sigma_{u_1:u_d}^{(i,t-1)}, \tilde{c}^{(i,t)}). \end{aligned}$$

In particular, it turns out that the new means are generated from the sampling density $f(y \mid \omega^{(i,t-1)}, \mu^{(i,t-1)}, (\sigma^2)^{(i,t-1)})$ of the mixture represented by the particle $\boldsymbol{\vartheta}^{(i,t-1)}$.

4.1.1 An empirical comparison with the Reversible Jump

The simulated dataset will be used in the following section to make a comparison with the RJ. This dataset presents three distinct modes, as we are interested mainly in the estimation of the posterior distribution of the number of components rather than the estimates of the other parameters of the model. Comparisons with the Reversible Jump are made using *Nmix*, an executable written by prof. Green to carry out the same analysis[1]. Using the described PMC method with $c_{\max} = 5$, $N = 10000$, $T = 20$, $p = 0.3$, at the 20^{th} iteration the estimates, given $c = 3$, the estimated mixture parameters are:

	Component		
	I	II	III
$\hat{\omega}$	0.293	0.516	0.191
$\hat{\mu}$	0.062	5.016	9.945
$\hat{\sigma}^2$	0.845	0.890	1.409

[1] *Nmix* is available at the website http://www.stats.bris.ac.uk/~peter/Nmix/

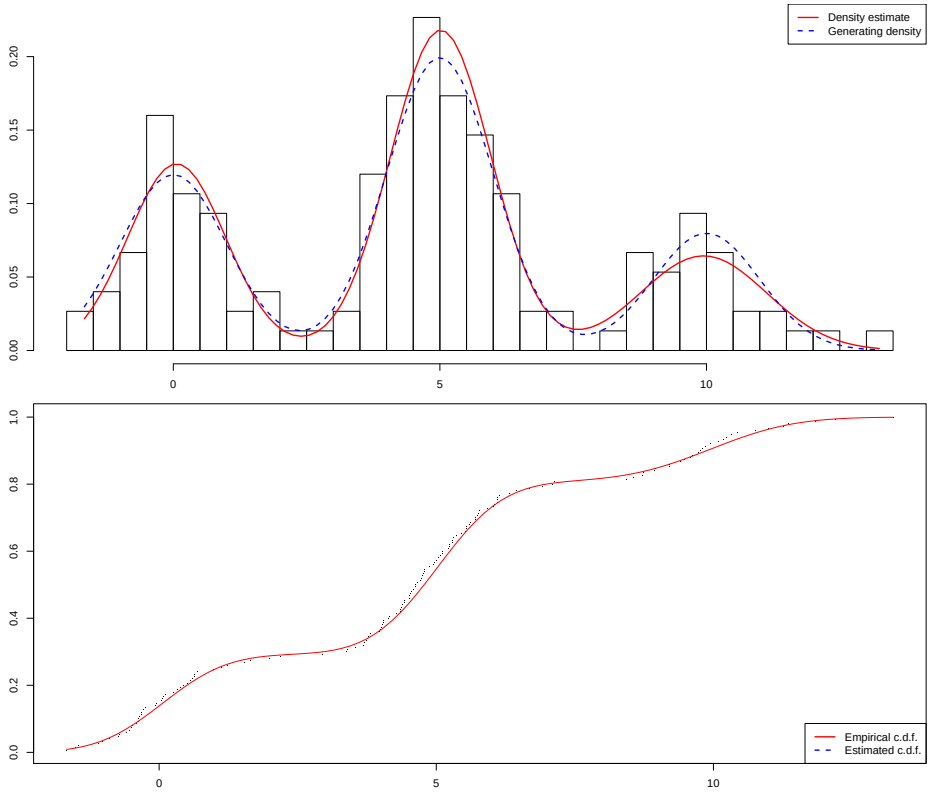

Figure 4.1: Upper panel: density estimate and generating density of the simulated dataset. Lower panel: empirical and estimated cdf.

As expected, given that the components are well-separated, these results are quite good. However, also the posterior on the number of components gives a relevant probability mass to the real model and decreases quickly in both senses:

c	1	2	3	4	5
$\pi(c \mid y)$	6e-13	2e-11	0.863	0.130	0.007

The model underlying the Reversible Jump is slightly different, so consequences must be taken with caution. In particular, *Nmix* uses an identifiability constraint on the means. Running 100000 iterations after a burn-in

period of other 100000 iterations, *Nmix* gives, conditional on $c = 3$, similar estimates for the parameters:

	Component		
	I	II	III
$\hat{\omega}$	0.284	0.513	0.203
$\hat{\mu}$	0.021	4.902	9.915
$\hat{\sigma}^2$	0.898	0.940	1.263

The most evident difference is in the estimate of the posterior distribution of c: while the MAP estimate is again the real model ($c = 3$), there is a long right tail that gives a significative probabilities to larger models.

c	1	2	3	4	5
$\pi(c \mid y)$	0.000000	0.000000	0.360435	0.293335	0.172575
c	6	7	8	9	10
$\pi(c \mid y)$	0.088910	0.043120	0.021045	0.009670	0.004865
c	11	12	13	14	15
$\pi(c \mid y)$	0.002855	0.001705	0.000820	0.000335	0.000245

The difference is mainly due to the fact that the chain visits often points on the space $\{\Theta, \boldsymbol{z}\}$ representing mixtures with empty components. The following table summarizes the empty components in the last 10000 iterations:

N. of empty components	0	1	2	3	4	5
N. of iterations	8626	1168	173	26	6	1

The large number of iterations passed in such points can be seen as a clue that the chain, after 200000 iteration, is still far from reaching convergence. It is worthwhile to note that, while PMC visits points with empty components, it never resamples such points.

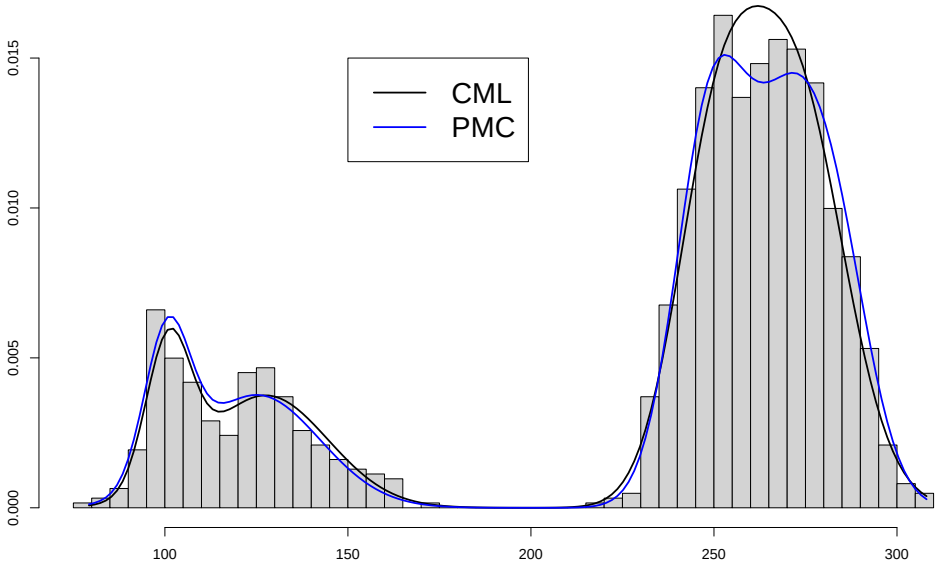

Figure 4.2: Histogram of the lengths, density estimation with PMC and CML conditional on $c = 6$, the real number of cohorts.

4.2 Results with the fish dataset

Fig. (4.2) shows the histogram of the lengths of fish: this is a typical situation in which, while younger cohorts are clearly distinguishable, the rest of the data form a single bunch. The estimation of a mixture model, even in a fixed dimension setting, would be really difficult without the help of the von Bertalanffy growth equation (1.1), as this equation allows to shrink information from the different cohorts. The superimposed density estimations are obtained using a slightly modified version of the algorithm described above and the complete maximum likelihood (CML) estimates. The latter are obtained using also the ages of fish to estimate the same model, condtional on $c = 6$, in a ML setting.

Even if more realistic models are available, here it is assumed that the lengths follow a normal mixture model:

$$\begin{cases} y_i \mid z_i, \boldsymbol{\vartheta}, c \sim \mathcal{N}(VB_{z_i}, \tau_{z_i}) \\ p(z_i = j \mid \boldsymbol{\vartheta}) = \omega_j \end{cases}$$

where VB_j is the mean length for a cohort of age j according to the Von Bertalanffy growth equation. The vector $\boldsymbol{\vartheta}$ contains all the parameters of the model, including L_∞, k and t_0.

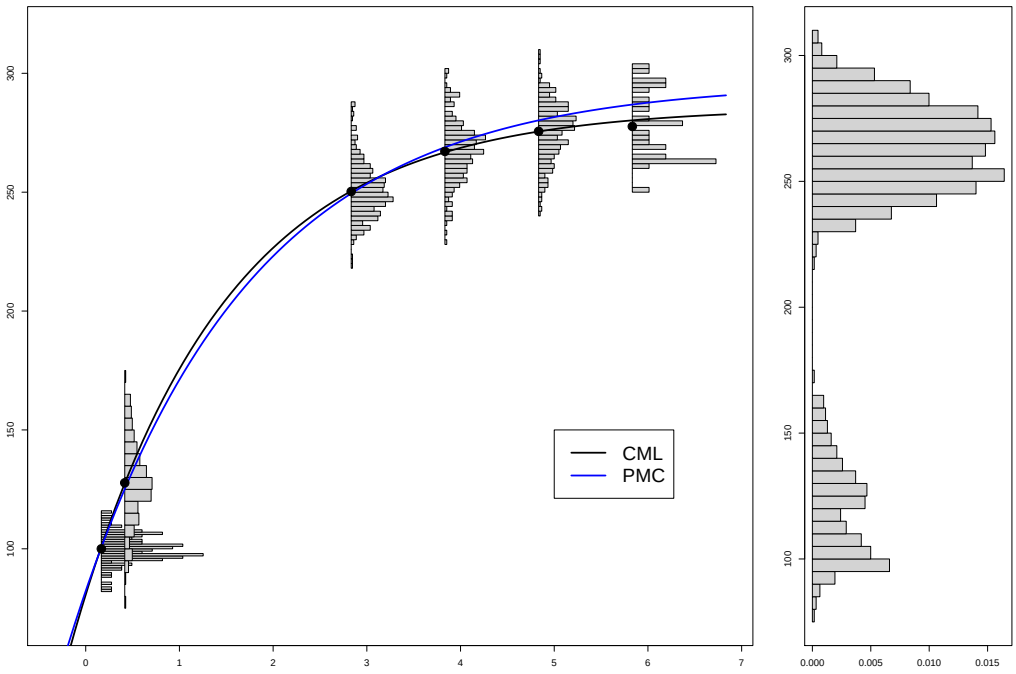

Figure 4.3: Left panel: histograms of lengths divided per age, with the PMC and the CML estimate of the Von Bertalanffy groth equation; dots represent sample means. Right panel: histogram of data.

Many studies have been carried on this species, and it is possible to make

use of the available information[2] to elicit subjective prior distributions:

$$\begin{array}{rclrcl} L_\infty & \sim & \mathcal{N}(300, 20^2) & \tau & \sim & \Gamma(2, 200) \\ k & \sim & \Gamma(2, 4) & \omega & \sim & \mathcal{D}(c, \ldots, 1) \\ t_0 & \sim & \mathcal{N}(-1, 4^2) & c & \sim & U(4, 6) \end{array}$$

The following results are obtained with a modification of the preceding algorithm: here we have three parameters that replace the vector of means. These parameters are updated using a D-kernel proposals (see Douc et al., 2007), while the others follow the preceding pseudo-code.

Conditional on $c_{\max} = 6$, the real number of cohorts in the dataset, we obtain the following results:

	CML	PMC
L_∞	285.660	292.805
k	0.622	0.568
t_0	-0.534	-0.583

	CML	PMC
ω_1	0.074	0.078
ω_2	0.160	0.162
ω_3	0.297	0.320
ω_4	0.205	0.238
ω_5	0.242	0.136
ω_6	0.023	0.067

	CML	PMC
σ_1	6.101	6.214
σ_2	17.042	17.150
σ_3	11.222	9.759
σ_4	12.844	10.102
σ_5	12.465	9.759
σ_6	14.408	9.017

Fig. 4.3 represents the lengths divided per age. Not surprisingly, the CML estimate of the VB curve is close to all the empirical means. Also the PMC estimate is satisfactory, even for older cohorts. The major drawback in using this model can be seen in the estimation of the number of components: the MAP estimate is five, while the real number of cohorts is six.

c	4	5	6
$\Pr(c \mid y)$	6.36462e-11	0.8208362	0.1791638

[2] The website http://www.fishbase.org/PopDyn/PopGrowthList.cfm?ID=24&GenusName=Clupea&SpeciesName=harengus+harengus&fc=43 collects the estimates of the von Bertalanffy parameters obtained in several studies.

Chapter 5

A tailored sampler

In the previous chapter, the von Bertalanffy growth curve represents a constraint on the means of the components of the mixture. In case no such constraints exist, one can make a better use of the possibilities of a PMC sampler. We indeed know that, in this case, the mixture posterior presents $c!$ symmetric and equivalent modes. A tailored sampler can therefore be characterized by a proposal distribution with the same feature: denoting with $\boldsymbol{\vartheta}^\star = (\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ a MAP estimate, it is possible to sample an auxiliary variable $\zeta \sim U(\zeta_1, \ldots, \zeta_{c!})$ representing one of the possible permutations of the labels; given ζ, a particle can be sampled, in the non-augmented space, from a distribution centered on $\zeta(\boldsymbol{\vartheta}^\star)$, the point obtained by permuting the labels of the MAP estimate.

All the parameters can be updated with a single multivariate distribution on $\boldsymbol{\psi} = (\boldsymbol{\omega}, \boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$. As suggested in Geweke (1989), a scale parameter V for this proposal can be obtained as the opposite of the inverse of the hessian of the likelihood calculated in $\boldsymbol{\psi}^\star$. It is also possible to calculate the hessian of the posterior distribution, but in this case the computational burden would be higher. Given V, the matrices relative to all the other permutations of $\boldsymbol{\psi}^\star$ can be obtained with a simple rearrangement of the indices of the lines

of V.
In order to compute the matrix V only once per iteration, all the particles, given ζ, can be sampled from a $t(\nu, \zeta(\boldsymbol{\psi}^\star), \zeta(V))$ distribution. In this case, the only arbitrary parameter is given by the number of degrees of freedom. At the end of each iteration, it is eventually possible to update the MAP estimate, improving the location and the scale of the proposal distribution.

The algorithm is particularly fast and simple to implement. Besides, the explicit sampling of the index of the permutation used for each particle, allows the use of these indexes to solve the label switching problem, easily obtaining meaningful estimates: taking as reference one permutation $\zeta^\star$, another permutation ζ_i to the labels of all the particles is performed, such that $\zeta_i(\zeta^{(i,t)}) = \zeta^\star$.

For $t = 1, \ldots, T$

for $i = 1 \ldots, N$

sample $\zeta^{(i,t)} \sim U(\zeta_1, \ldots, \zeta_{\mathrm{c}!})$

sample $\boldsymbol{\psi}^{(i,t)} \mid \zeta^{(i,t)} \sim t(\nu, \zeta^{(i,t)}(\boldsymbol{\psi}^\star), \zeta^{(i,t)}(V))$

compute $\bar{\rho} = \pi(\boldsymbol{\psi}^{(i,t)}) / \sum_{l=1}^{\mathrm{c}!} f_t(\boldsymbol{\psi}^{(i,t)} \mid \nu, \zeta^{(i,t)}(\boldsymbol{\psi}^\star), \zeta^{(i,t)}(V))$

normalize the importance weights ρ

if any $\pi(\boldsymbol{\psi}^{(i,t)}) > \pi(\boldsymbol{\psi}^\star)$ update $\boldsymbol{\psi}^\star$ and V.

Importance weights are computed only to obtain the estimates of interest: particles are never resampled, as each particle is drawn independently from their past values given the MAP estimate of the previous iteration. However, we can resample them in order to draw the marginal sampling representation of the posterior for couples of parameters as in Fig. 5.3. With this representation, we could perform the mode hunting, an informal method given by

Frühwirth-Schnatter (2001) for diagnosing purposes. If the mixture is not overfitting, then $c(c-1)$ clusters will be clearly visible.

5.1 Galaxy dataset

As an application of this sampler, we used the Galaxy dataset. We used data-dependent conjugate priors as in (2.5).

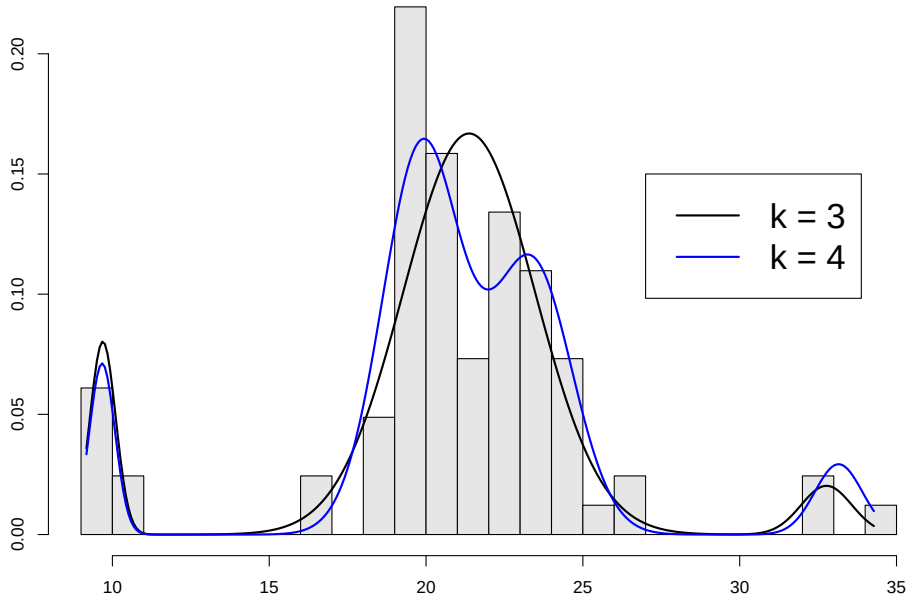

Figure 5.1: Histogram of the galaxy data and density estimates.

Fig. 5.1 shows the density estimation with $c = 3$ and $c = 4$. Fig. 5.2 shows the estimates of the marginal likelihood for models $\mathcal{M}_3$ and $\mathcal{M}_4$: the stability of this method is remarkable, as shown by the absence of the degeneracy phenomenon. However, increasing of the number of components, the situation gets quickly worse.

In Cappé et al. (2003), the authors find that a Gibbs sampler could not

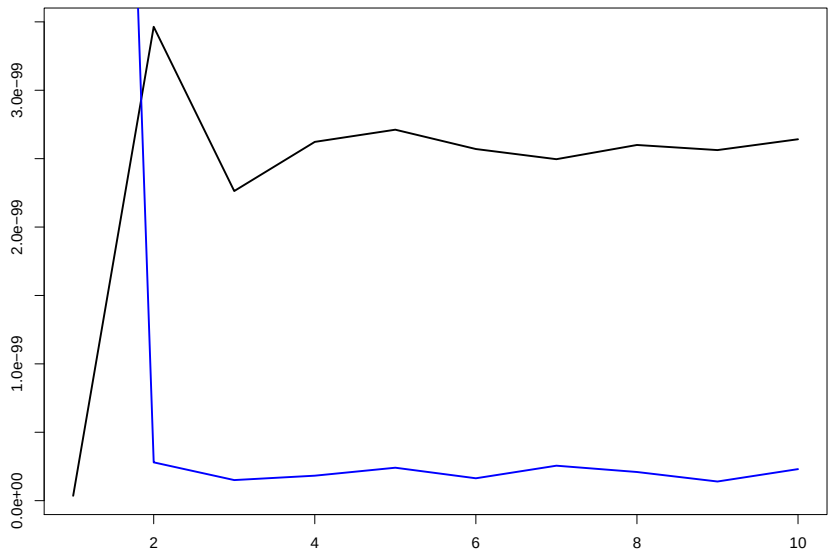

Figure 5.2: Estimates of the marginal likelihood over iterations.

visit all the modes of the posterior; hence, to accelerate the convergence of the sampler, they renounced to the completion of the space. Also in this case we don't use the augmentation: this choice makes harder the computation of the posterior density appearing in the numerator of the importance weights, but this disadvantage is offset by the efficient generation of particles and the computation of the relative proposal densities. As a result, the computational burden is not very high, at least for models with a moderate number of components.

In Fig. (5.3) we see the means generated by the proposals for the model $c = 4$ and the resampled means. The sampling representation denotes that all the modes have been adequately visited: in the right panel there are clearly $c(c-1) = 12$ modes, higlighting that this model is not overfitting.

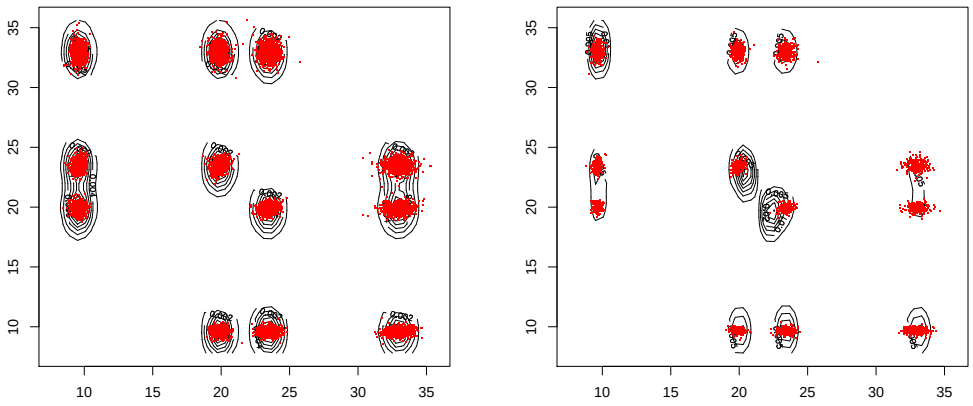

Figure 5.3: Left panel: sampled means for the model $c = 4$. Right panel: means after resampling.

Bibliography

Baldi, P., Russo, T., Parisi, A., Magnifico, G., Mariani, S. & Cataudella, S. (to be submitted). A new stochastic von bertalanffy model of fish growth, with application to population analysis. .

Bhattacharya, C. G. (1967). A simple method for resolution of a distribution into its gaussian components. *Biometrics* **23**, 115–135.

Cappé, O., Guillin, A., Marin, J. M. & Robert, C. P. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.* **13**, 907–929.

Cappé, O., Robert, C. P. & Rydén, T. (2003). Reversible jump, birth-and-death and more general continuous time markov chain monte carlo samplers. *Journal Of The Royal Statistical Society Series B* **65**, 679–700. Available at http://ideas.repec.org/a/bla/jorssb/v65y2003i3p679-700.html.

Celeux, G. & Diebolt, J. (1985). The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Comput. Statist. Quater.* **2**, 73–82.

Celeux, G., Hurn, M. & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957–970.

Celeux, G., Marin, J.-M. & Robert, C. P. (2006). Iterated importance sampling in missing data problems. *Comput. Statist. Data Anal.* **50**, 3386–3404.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 1313–1321.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via de EM algorithm. *The Journal of Royal Statistical Society* **39**, 1–37.

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

Douc, R., Guillin, A., Marin, J. M. & Robert, C. P. (2007). Convergence of adaptive mixtures of importance sampling schemes **35**. Comment: Published at http://dx.doi.org/10.1214/009053606000001154 in the Annals of Statistics (http://www.imstat.org/aos/) by the Institute of Mathematical Statistics (http://www.imstat.org).

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

Feller, W. (1943). On a general class of "contagious" distributions. *The Annals of Mathematical Statistics* **14**, 389–400.

Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* **96**, 194–209 (16).

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

Geweke, J. F. (1989). Bayesian inference of econometric models using monte carlo integration. *Econometrica* **57**, 1317–1339.

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

Guihenneuc-Jouyaux, C., Mengersen, K. & Robert, C. (1998). Mcmc convergence diagnostics: A "reviewww". Papers 9816, Institut National de la Statistique et des Etudes Economiques. Available at http://ideas.repec.org/p/fth/inseep/9816.html.

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

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27**, 887–906.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. IMS Monographs. Hayward, CA.

Lv, Q. & Pitchford, J. W. (2007). Stochastic von Bertalanffy models, with applications to fish recruitment. *J. Theoret. Biol.* **244**, 640–655.

Marin, J., Mengersen, K. & Robert, C. (2005). Bayesian modelling and inference on mixtures of distributions. In *Handbook of Statistics*, D. Dey & C. Rao, eds., vol. 25. Elsevier-Sciences.

McLachlan, G. & Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

Meng, X.-L. & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6**, 831–860.

Mengersen, K. & Robert, C. (1996). *Bayesian Statistics 5*, chap. Testing for mixtures: a Bayesian entropy approach. Oxford University Press,Oxford.

Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics* **10**, 35–57.

Pearson, K. (1893). Contributions to the mathematical theory of evolution. *Journal of the Royal Statistical Society* **56**, 675–679.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59**, 731–792.

Robert, C. P. & Casella, G. (2004). *Monte Carlo statistical methods*. Springer Texts in Statistics. New York: Springer-Verlag, 2nd ed.

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

ROEDER, K. & WASSERMAN, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**.

ROTHENBERG, T. J. (1971). Identification in parametric models. *Econometrica* **39**, 577–591.

RUBIN, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc.

RUBINSTEIN, R. Y. (1981). *Simulation and the Monte Carlo method*. New York: John Wiley & Sons Inc. Wiley Series in Probability and Mathematical Statistics.

TANNER, M. A. & WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. with discussion and with a reply by the authors. *Journal of the American Statistical Association* **82**, 528–550.

TEICHER, H. (1960). On the mixture of distributions. *The Annals of Mathematical Statistics* **31**, 55–73.

TEICHER, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics* **32**, 244–248.

TEICHER, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics* **34**, 1265–1269.

TITTERINGTON, D. M., SMITH, A. F. M. & MAKOV, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Chichester: John Wiley & Sons Ltd.

Venables, W. N. & Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus*. New York: Springer.

von Bertalanffy, L. (1938). A quantitative theory of organic growth. *Human Biol.* **10**, 181–213.

Yuan, C. & Druzdzel, M. (2005). How heavy should the tails be? In *Proceedings of the Eighteenth International FLAIRS Conference (FLAIRS-05)*, I. Russell & Z. Markov, eds. AAAI Press/The MIT Press, Menlo Park, CA.