Università della Svizzera Italiana
Facoltà di Scienze Economiche

Bayesian Analysis for Mixtures of Autoregressive Components with Application to Financial Market Volatility

A thesis submitted by
Stefano Sampietro
for the Degree of
Dottore in Scienze Economiche

February, 2004

Members of the jury:

Prof. F. Trojani, President of the jury
Prof. G. Arbia, Thesis director
Prof. G. Consonni, External member

Abstract

This thesis presents a Bayesian analysis of a non-linear time series model. In particular, we deal with a mixture of normal distributions whose means are linear functions of the past values of the observed variable. Since the component densities of the mixture can be viewed as the conditional distributions of different Gaussian autoregressive models, the model is referred as mixture of autoregressive components.

Bayesian perspective implies some advantages, especially in terms of model determination. First of all, it has been recognized that usual criterions like AIC and BIC are not satisfactory for the mixture of autoregressive components. In addition, these standard approaches do not take into account model uncertainty because they select a single model and then make inference based on this model. On the contrary, our Bayesian approach maintains consideration of several models, with the input of each into the analysis weighted by the model posterior probability.

Both parameter estimation and model selection do not lend themselves to analytic solutions and we use *Markov Chain Monte Carlo* (or *MCMC*) approximation methods, which have had a real explosion over the last years, especially in Bayesian statistics.

Our work takes into account the stationarity conditions of the autoregressive coefficients of the mixture components through a reparametrization in terms of partial autocorrelations.

Finally, this thesis addresses the important task of modelling and forecasting return volatility. Several stylized facts about volatility have been recognized and they are captured by the mixture of autoregressive components.

Acknowledgments

I would like to express my thanks to the people who contributed to this thesis: Prof. Giuseppe Arbia (G. d'Annunzio University, Pescara), Prof. Giovanni Barone-Adesi (University of Lugano), Prof. Guido Consonni (University of Pavia), Prof. Petros Dellaportas (Athens University of Economics and Business), Prof. Antonietta Mira (Insubria University, Varese), Prof. Fabio Trojani (University of Lugano).

Moreover I would like to thank postgraduates students at the University of Lugano and all the academic people I have held useful discussions with.

I also thank my family for creating an environment which allowed me to follow this path and Roberta for her essential support and encouragement.

Ringraziamenti

Desidero esprimere i miei ringraziamenti alle persone che hanno contribuito al presente lavoro: Prof. Giuseppe Arbia (Università G. d'Annunzio, Pescara), Prof. Giovanni Barone-Adesi (Università della Svizzera Italiana, Lugano), Prof. Guido Consonni (Università degli Studi di Pavia), Prof. Petros Dellaportas (University of Economics and Business, Atene), Prof. Antonietta Mira (Università deli Studi dell'Insubria, Varese), Prof. Fabio Trojani (Università della Svizzera Italiana, Lugano).

Voglio inoltre ringraziare gli studenti di dottorato dell'Università della Svizzera Italiana e tutti i ricercatori con cui ho intrattenuto utili discussioni.

Ringrazio anche la mia famiglia per aver creato un contesto che mi ha permesso di seguire questa strada e Roberta per il suo indispensabile supporto e incoraggiamento.

Models do not represent truth. Rather they are ways of viewing a system, its problems and their contexts.

(West and Harrison, 1989)

There are two things you are better off not watching in the making: sausages and econometric estimates.

(Leamer, 1983)

The people who don't know they are Bayesian are called non-Bayesian.

(Good, 1983)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The class of finite mixture models provides a mean for the formalization of heterogeneity, when the phenomena are too intricate to be described by simple probabilistic modelling through classical distributions. The assumption of conditional independent observable variables is usually made in this context, making a direct application to time series data inappropriate.

The mixture of autoregressive components we analyse relaxes this hypothesis. The resulting model is a non-linear time series tool (for instance, it takes into account changes in conditional distributions, which are not necessarily symmetric or unimodal) that is particularly suitable for financial data.

More precisely, we consider an application to financial market volatility. Several empirical facts about volatility have been recognized (persistence, clustering and threshold effects, non-symmetrical dependencies, etc.) and we believe that the mixture of autoregressive components captures these stylized features.

Parameter estimation and model selection are based on a Bayesian perspective. The advantages of this choice are particularly evident for the model selection problem. In general, Bayesian model selection does not ignore model uncertainty: while standard criterions, like AIC and BIC, select a single model and then make inference based on it only, a Bayesian approach maintains consideration of several models, with the input of each into the analysis weighted by their posterior probability. Influence of model uncertainty on financial models is an important factor and it has been recently investigated in financial literature.

In addition, a contingent reason to adopt the Bayesian approach is that it has been recognized that AIC and BIC criterions are not satisfactory in this mixture context.

Bayesian estimation and model selection for the mixtures of autoregressive components do not lead to analytical solution, thus we use *Markov Chain Monte Carlo* (or *MCMC*) approximation procedures. These methods have had a real explosion over the last years and they are often adopted in Bayesian statistics.

The thesis is organized as follows. In chapter 2 we shall present a brief introduction about Bayesian inference and a review of some univariate and multivariate distributions which will be used in the rest of the thesis.

Chapter 3 will start with some traditional methods based on stochastic simulation. After a presentation of some properties of a Markov chain, the class of MCMC methods will be illustrated. Eventually, the implementation of an algorithm and the use of its output to make statistical inference will be addressed.

Bayesian model selection will be treated in chapter 4. We shall present model selection techniques which are related to MCMC methods. As a matter of fact, some of them can be viewed as generalizations of such methods, while other techniques estimate model posterior probabilities by using a standard MCMC output.

In chapter 5 we shall illustrate some basics concepts about finite mixture models, with particular attention on the mixture of normal distributions. Bayesian estimation and model selection will be also treated.

The mixture of autoregressive components will be presented in chapter 6. We shall give the definition of the model and the prior structure. Also, we shall explain how to take into account the stationarity conditions on the autoregressive coefficients. Parameter estimation and model selection will be illustrated in details. Finally, the chapter will present the calculation of the predictive distributions.

The last chapter will show the application of the model to the return volatility. It will be a self-contained chapter, which will report definition and properties of the model and will summarize parameter estimation and model selection procedures.

Chapter 2

Bayesian inference

2.1 Introduction

Statistical theory is mainly devoted to derive an *inference* about the probability distribution underlying a random phenomenon from observations of this phenomenon.

Statistical inference can be viewed as a formalization step based on a *probabilistic modelling* with the purpose of interpreting the natural phenomena.

Different statistical approaches to inference are possible. For instance, we can distinguish between a *parametric* and a *nonparametric* approach. The former represents the distribution of an observed random variable y through a density or probability function $f(y|\theta)$, where only the parameter θ is unknown. On the contrary, the purpose of the second approach is to estimate the distribution under minimal assumptions, typically using functional estimation. In this work, we shall only consider parametric modelings.

The parameter θ can be considered as an index of a family of possible distributions for the observations. Thus θ is a quantity of interest whose estimation is necessary in order to obtain a description of the process.

The mathematical nature of θ leads to another distinction between competing statistical approaches. In the *classical* or *frequentist* statistical theory, the parameters are fixed (but unknown) quantities. Conversely, *Bayesian* approach consider θ as a random variable. Hence, θ is allowed to have a probability distribution called *prior distribution* or simply *prior.*

This feature implies an important consequence. It is likely that a researcher has some knowledge about the phenomenon under study, i. e. he can have some information *a priori* with respect to the experiment. It is scientifically recommended that this body of knowledge should be formally incorporated in the analysis. The Bayesian approach includes this kind of information through the prior distribution.

This chapter will present a brief overview of some concepts of Bayesian inference. For a more thorough discussion, many books are available: see for instance Bernardo and Smith (1994), Gelman et al. (1995) or Robert (2001).

In the sequel, the terminology will not distinguish between discrete and continuous quantities; thus, f or p can represent a density function or a probability function as well. The notation for the parameters will be the same for the univariate and multivariate case.

Whenever required, a distinction will be made in the text.

2.2 Bayes' theorem

The first two elements of a Bayesian statistical model are the observational (or sampling) distribution $f(y|\theta)$ and the prior distribution $p(\theta)$. If regarded as a function of θ, the observational distribution gives the likelihood function of θ, $L(\theta) = f(y|\theta)$. The parameters of the prior are called *hyperparameters* and they are initially assumed to be known.

Inference is based on the so called *posterior distribution*: it can be viewed as a compromise between likelihood and prior, i. e. between empirical and subjective information. The posterior distribution is formally obtained by means of Bayes' theorem:

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)d\theta} \tag{2.1}$$

Note that the posterior is a density for θ and then the denominator of (2.1) is simply a constant (*normalizing constant*). Bayes' theorem can then be written in a more compact form:

$$p(\theta|y) \propto L(\theta)p(\theta) \tag{2.2}$$

where $\propto$ means "proportional to".

It is worth anticipating that analytical calculation of the normalizing constant is not always possible. In the next chapter, we shall discuss this problem.

When θ is a multivariate parameter $\theta = (\theta_1, \ldots, \theta_k)'$, we can obtain marginal and conditional posterior distributions from the joint posterior density. The marginal posterior density of θ_j is:

$$p(\theta_j|y) = \int p(\theta_1, \ldots, \theta_k|y) d\theta_{-j}$$

where $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$.

Within the possible conditional distributions, an important role is played by the so called *full conditional*: roughly speaking, a full conditional is the distribution of a component conditional on the all remaining components. The full conditionals for the posterior distribution are:

$$p(\theta_j|\theta_{-j}, y) = \frac{p(\theta_1, \ldots, \theta_k|y)}{p(\theta_{-j}|y)} \propto p(\theta_1, \ldots, \theta_k|y) \tag{2.3}$$

for $j = 1, \ldots, k$. Of course, the above representation could be generalized for the case in which θ is partitioned into vector components.

Another important ingredient for Bayesian inference is the *predictive distribution*. Suppose y denotes an observed sample and let $\tilde{y}$ be an unknown observable variable. The distribution of $\tilde{y}$ conditional on y is called predictive distribution (or posterior predictive distribution) and it is equal to:

$$\begin{aligned} f(\tilde{y}|y) &= \int f(\tilde{y}, \theta|y) d\theta \\ &= \int f(\tilde{y}|\theta, y) p(\theta|y) d\theta \end{aligned}$$

The predictive distribution provides the expected distribution of $\tilde{y}$:

$$f(\tilde{y}|y) = E[f(\tilde{y}|\theta, y)]$$

where the expectation is taken with respect to the posterior.

2.3 Conjugate distributions

Some of the elements we introduced informally in the previous sections form a Bayesian statistical model $\mathcal{F}$:

$$\mathcal{F} = \{y \;\; ; \;\; L(\theta) \;\; ; \;\; p(\theta) \;\; ; \;\; \theta \in \Theta\}$$

In this work, a statistical model will also be denoted with the following notation:

$$y|\theta \sim f(y|\theta)$$
$$\theta \sim p(\theta)$$

where the symbol $\sim$ means "distributed as".

The case in which, given a statistical model, prior and posterior distributions belong to the same class of distributions shows a property called *conjugacy*. More precisely, a family of distribution P is conjugate to a statistical model $\mathcal{F}$ if for every prior $p \in P$ and for any observational distribution $f \in F$, the posterior belongs to P. Dealing with conjugate distributions assures analytic tractability: derivation of the posterior only requires a change in the hyperparameters with no additional calculation.

For example, consider the following *normal model* for a vector of observations $y = (y_1, \ldots, y_n)$:

$$y_i|\mu \overset{iid}{\sim} N(y|\mu, \sigma^2)$$
$$\mu \sim N(\mu|\lambda, \tau^2)$$

where $N(.)$ stands for the normal distribution. The hyperparameters σ^2, λ and τ^2 are assumed known. The posterior distribution for μ is shown to be a normal distribution:

$$p(\mu|y) = N\left(\mu\left|\frac{n\sigma^{-2}\bar{y} + \tau^{-2}\lambda}{n\sigma^{-2} + \tau^{-2}}, \frac{1}{n\sigma^{-2} + \tau^{-2}}\right.\right)$$

where $\bar{y}$ is the sample mean of y.

2.4 Hierarchical models

The prior distribution is the formal representation of the prior information. Unfortunately, prior information could not be rich enough to define a prior distribution exactly. In such situations, it is desirable to incorporate this uncertainty in the Bayesian model. To be precise, the model should be enriched with "additional"uncertainty because the concept of prior is itself a way to include uncertainty (about the parameters) in the analysis.

A *hierarchical* model decomposes the prior distribution into several conditional levels of distributions. In other words, the hyperparameters are no longer fixed and they become random variables. The simplest hierarchial model has only one additional level:

$$\mathcal{F} = \{y \;\; ; \;\; L(\theta) \;\; ; \;\; p(\theta|\theta_1) \;\; ; \;\; p_1(\theta_1) \;\; ; \;\; \theta \in \Theta, \theta_1 \in \Theta_1\}$$

The additional prior $p_1(\theta_1)$ is called *hyperprior*. The alternative notation shows the hierarchical structure of the model better:

$$\begin{aligned} y|\theta &\sim f(y|\theta) \\ \theta|\theta_1 &\sim p(\theta|\theta_1) \\ \theta_1 &\sim p(\theta_1) \end{aligned}$$

Through the specification of additional priors, a hierarchical analysis allows to reduce the arbitrariness of the hyperparameters choice.

The posterior distribution is obtained by successive application of Bayes' theorem:

$$p(\theta|y) = \int p(\theta|\theta_1, y)p(\theta_1|y)d\theta_1$$

where

$$\begin{aligned} p(\theta|\theta_1, y) &= \frac{f(y|\theta)p(\theta|\theta_1)}{f(y|\theta_1)} \\ f(y|\theta_1) &= \int f(y|\theta)p(\theta|\theta_1)d\theta \\ p(\theta_1|y) &= \frac{f(y|\theta_1)p_1(\theta_1)}{f(y)} \\ f(y) &= \int f(y|\theta_1)p_1(\theta_1)d\theta_1 \end{aligned}$$

As an example of hierarchical model, suppose any observation has a double index y_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. For instance, i could denote a group and j an individual. A *normal hierarchical model* can be specified as follows:

$$\begin{aligned} y_{ij}|\mu_i &\overset{ind}{\sim} N(y_{ij}|\mu_i, \sigma^2) \\ \mu_i|\lambda &\overset{iid}{\sim} N(\mu_i|\lambda, \tau^2) \\ \lambda &\sim N(\lambda|m, r^2) \end{aligned}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ and where hyperparameters σ^2, τ^2, m and r^2 are assumed known. The parameters μ_i can be viewed as group specific means, while λ is the overall mean.

An early illustration of the hierarchical Bayes analysis for the normal linear model is given in Lindley and Smith (1972).

More complex structures are possible. For instance, in *partition hierarchical models*, first level parameters are clustered into *partitions*: μ_i's are independent and identically distributed only if they belong to the same partition. See Malec and Sedransk (1992), Consonni and Veronese (1995) and Sampietro and Veronese (1998).

2.5 Some particular distributions

In this section, we shall present a review of some univariate and multivariate distributions which will be used in the next chapters.

2.5.1 The beta and the generalized beta distributions

A continuous random variable θ has a *beta* distribution with positive parameters α and β if its density function $\mathrm{Be}(\theta|\alpha, \beta)$ is

$$\mathrm{Be}(\theta|\alpha, \beta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

for $0 < \theta < 1$, where

$$c = \frac{1}{B(\alpha, \beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

where B stands for the *Beta function*:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

and where Γ is the *Gamma function*:

$$\Gamma(\alpha) = \int_0^\infty e^{-t}t^{\alpha-1}dt$$

It is possible to show that, if θ has a beta distribution, then:

$$E(\theta) = \frac{\alpha}{\alpha+\beta}$$

$$VAR(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

By considering the transformation $\lambda = a+(b-a)\theta$, where θ has a $\mathrm{Be}(\theta|\alpha,\beta)$ density, the beta distribution can be generalized to any finite interval (a,b) and the resulting density function $\mathrm{Be}_{(a,b)}(\lambda|\alpha,\beta)$ is:

$$\mathrm{Be}_{(a,b)}(\lambda|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\frac{(\theta-a)^{\alpha-1}(b-\theta)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}$$

for $a < \theta < b$. We shall refer to this distribution as *generalized beta*.

2.5.2 The gamma distribution

A continuous random variable θ has a *gamma* distribution with positive parameters α and β if its density function $\mathrm{Ga}(\theta|\alpha,\beta)$ is

$$\mathrm{Ga}(\theta|\alpha,\beta) = c\,\theta^{\alpha-1}e^{-\beta\theta}$$

for $\theta > 0$ and where c is a constant equal to:

$$c = \frac{\beta^\alpha}{\Gamma(\alpha)}$$

If θ has a gamma distribution, then:

$$E(\theta) = \frac{\alpha}{\beta}$$
$$VAR(\theta) = \frac{\alpha}{\beta^2}.$$

2.5.3 The inverted-gamma distribution

A continuous random variable θ has an *inverted-gamma* distribution with positive parameters α and β if its density function $\mathrm{Ig}(\theta|\alpha,\beta)$is

$$\mathrm{Ig}(\theta|\alpha,\beta) = c\,\theta^{-(\alpha+1)}e^{-\beta/\theta}$$

for $\theta > 0$ and where c is a constant equal to:

$$c = \frac{\beta^\alpha}{\Gamma(\alpha)}$$

If θ has an inverted-gamma distribution, then

$$E(\theta) = \frac{\beta}{\alpha-1}, \quad \alpha > 1,$$
$$VAR(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2.$$

The name of this distribution derives from the fact that if θ has a $\mathrm{Ga}(\theta|\alpha,\beta)$ density, then $\lambda = \theta^{-1}$ has an $\mathrm{Ig}(\lambda|\alpha,\beta)$ density.

2.5.4 The Dirichlet distribution

A continuous random vector $w = (w_1, \ldots, w_k)$ has a *Dirichlet* distribution of dimension k, with positive parameters $\delta_1, \ldots, \delta_k$, if its density function $\mathrm{Di}(w|\delta_1, \ldots, \delta_k)$ is

$$\mathrm{Di}(w|\delta_1, \ldots, \delta_k) = c\, w_1^{\delta_1 - 1} \cdots w_k^{\delta_k - 1}$$

for $0 < w_j < 1$, $j = 1, \ldots, k$, and $w_1 + \cdots + w_k = 1$. c is a constant equal to:

$$c = \frac{\prod_{j=1}^{k} \Gamma(\delta_j)}{\Gamma(\sum_{j=1}^{k} \delta_j)}$$

Let $\delta_0 = \sum_{j=1}^{k} \delta_j$. If w has a Dirichlet distribution, then

$$E(w_j) = \frac{\delta_j}{\delta_0},$$
$$VAR(w_j) = \frac{\delta_j(\delta_0 - \delta_j)}{\delta_0^2(\delta_0 + 1)}$$

2.5.5 The truncated normal distribution

In general, we can define a *truncated distribution* simply as the part of a distribution that is above or below some specified values. Suppose θ is a continuous random variable with probability density function $p(\theta)$; the density of θ truncated between two constants a and b is:

$$p(\theta|a < x < b) = \frac{p(\theta)}{\mathrm{Prob}(a < \theta < b)}$$

The density of the *truncated normal distribution* with mean μ and variance σ^2 will be denoted by $\mathrm{N}_{(a,b)}(\theta|\mu, \sigma^2)$ and it is:

$$\mathrm{N}_{(a,b)}(\theta|\mu, \sigma^2) = \frac{\mathrm{N}(\theta|\mu, \sigma^2)}{F_N(b|\mu, \sigma^2) - F_N(a|\mu, \sigma^2)}$$

where F_N is the normal cumulative distribution function.

Chapter 3

MCMC methods

3.1 Introduction

The class of approximate methods of inference consists in techniques useful when calculations cannot be performed analytically. It is possible to distinguish them in methods based on deterministic concepts and methods based on stochastic simulation. For instance, *Normal approximation*, *Laplace approximations* and *Gaussian quadrature* belong to the first category. For a review, see Evans and Swartz (1995).

This chapter is devoted to present the second class which, in turn, can be divided into traditional methods based on non-iterative simulation (section 3.2) and methods based on iterative simulation, essentially formed by Markov chain Monte Carlo or *MCMC* or again MC^2 algorithms (section 3.4).

Because of the subject of this thesis, the emphasis will be on Bayesian inference problems. Nevertheless, the methods we shall present are suitable for more general applications.

3.2 Traditional methods based on stochastic simulation

3.2.1 Monte Carlo calculations

In general, the idea of the methods based on stochastic simulation is to summarize information concerning a distribution using samples from the distribution itself.

Let X denote a random variable whose distribution π is the distribution of interest. Suppose we are able to generate N independent and identically distributed (i.i.d) random draws $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ from π. The Monte Carlo recipe is to estimate the expected value of X with respect to π simply with the empirical average:

$$\frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

More generally, let $E_\pi[g(X)]$ the expected value of a function g of X; its estimate is:

$$\bar{g} = \frac{1}{N}\sum_{i=1}^{N} g(x^{(i)}) \tag{3.1}$$

Note that $\bar{g}$ is an unbiased estimate of $E_\pi[g(X)]$ and has a sampling distribution that is approximately Gaussian.

3.2.2 Importance sampling

Suppose direct generation from π is not possible: in such a case, Monte Carlo technique seems to be useless. Luckily, *importance sampling* can often help us because it enables us to approximate $E_\pi[g(X)]$ if π is close to another distribution, say π^*, from which we have a random sample.

Assume that π and π^* are proportional to the functions h and h^* respectively:

$$\begin{aligned} \pi^*(x) &= c^*\, h^*(x) > 0, \qquad x \in S^* \\ \pi(x) &= c\, h(x) > 0, \qquad x \in S \end{aligned}$$

where only h and h^* are necessarily known and $S \subset S^*$.

Suppose we need $E_\pi[g(X)]$, for a given g, but we have a random sample $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ from π^*. First of all, note that:

$$\begin{aligned} E_{\pi^*}\left[\frac{g(X)h(X)}{h^*(X)}\right] &= \int_{S^*} \frac{g(x)h(x)}{h^*(x)}\; c^* h^*(x) dx \\ &= \frac{c^*}{c}\int_S g(x)\; c\; h(x) dx \qquad (3.2) \\ &= \frac{c^*}{c}\; E_\pi[g(X)] \end{aligned}$$

The constants c and c^* can be unknown but, as a special case of (3.2):

$$E_{\pi^*}\left[\frac{h(X)}{h(X)^*}\right] = \frac{c^*}{c}$$

Thus, the estimate for $E_\pi[g(X)]$ is

$$\frac{1}{N}\sum_{i=1}^{N}\frac{h(x^{(i)})g(x^{(i)})}{h^*(x^{(i)})}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{h(x^{(i)})}{h^*(x^{(i)})}\right]^{-1} \tag{3.3}$$

or

$$\sum_{i=1}^{N} w(x^{(i)})g(x^{(i)})$$

where

$$w(x^{(i)}) = \frac{h(x^{(i)})/h^*(x^{(i)})}{\sum_{i=1}^{N} h(x^{(i)})/h^*(x^{(i)})}$$

In practice, the efficiency of this methods depends on how π^* is close to π: in fact, the estimate is satisfactory if there are no large weights among the $w(x^{(i)})$'s.

As a final comment, note that we required distributions π and π^* known up to a constant. Keeping in mind Bayes' theorem (equation 2.1), it is immediate to apply the method in a Bayesian context (e.g. $h(\theta) = L(\theta)p(\theta)$), when we cannot calculate analytically the normalization constant.

3.3 Markov chains

As it will be clear later on, MCMC methods overcome the limit of Monte Carlo calculations (i.e. the need of an i.i.d sample from the target distribution π) without requiring another distribution close to π (as in importance sampling), but constructing a Markov chain with limit distribution π.

Thus, a brief presentation of Markov chains is now proposed. For a comprehensive treatment of this kind of stochastic processes, the reader is referred to the books by Cox and Miller (1965) or Ross (1996).

Markov dependence is a well known concept attributed to the Russian mathematician Andrei Andreivich Markov. For some set T, let $\{X^{(t)} : t \in T\}$ be a collection of random quantities defining a *stochastic process*. The set of the values that $X^{(t)}$ can assume is called the *state space*: it will be denoted by S and it will be initially assumed finite. T is the *index set*.

Roughly speaking, a process has the Markovian property if, given the present state, past and future states are independent. A stochastic process with a countable index set T is called a *discrete time* stochastic process. A *Markov chain* is a discrete time stochastic process which satisfies the Markovian property.

More formally, the process $\{X^{(t)} : t \in T\}$ is a Markov chain if T is countable and if:

$$\begin{aligned}
&\text{Prob}(X^{(t+1)} \in C | X^{(t)} = x, X^{(t-1)} \in C_{t-1}, \ldots, X^{(0)} \in C_0) \\
&= \text{Prob}(X^{(t+1)} \in C | X^{(t)} = x)
\end{aligned}$$

for all sets $C_0, \ldots, C_{t-1}, C \subset S$ and $x \in S$.

3.3.1 Transition probabilities

The probability that $X^{(t+1)} = z$ given $X^{(t)} = x$ is called *transition probability*. If it does not depend on t, it will be denoted by $p(x, z)$: in this case, the chain is said to be *homogeneous*.

Suppose to collect all the transition probabilities in a matrix P, with the (i, j)th element given by $p(x_i, x_j)$: P is called *transition probability matrix*. Clearly, P has non-negative elements and each row sums to one.

If the chain is homogeneous, it is possible to show that transition probabilities over m steps can be obtained by the matrix product of P m times. In other words, the resulting matrix P^m contains the probabilities of a chain moving from a state to another in exactly m steps.

3.3.2 Stationary distribution and ergodicity

Let $X^{(1)}, X^{(2)}, \ldots$ be a Markov chain with transition probability matrix P and state space S. In addition, let $p^{(0)}$ be the row vector representing the distribution of the initial state $X^{(0)}$. The marginal distribution of $X^{(t)}$ is given by:

$$p^{(t)} = p^{(0)} P^t, \qquad t = 0, 1, \ldots$$

If π is a probability vector satisfying the *general balance*:

$$\pi P = \pi \tag{3.4}$$

then π is a *stationary distribution* for P. Since P maintains π, if $p^{(0)} = \pi$, then $p^{(t)} = \pi$ for all $t = 1, 2, \ldots$.

Suppose $X^{(t)} = x$ for a given t. Avoiding formal details, we can say that another state z is *accessible* from x if, after m transitions, there is a positive probability that $X^{(t+1)} = z$. A Markov chain is called *irreducible* if all the states of the chain are accessible from each other. In irreducible chains there may still exist a periodic structure such that for each state $x \in S$, the set of possible return times to x when starting in x is a subset of the set $\{d, 2d, 3d, \ldots\}$, with $d \in \mathcal{N}$. The smallest number d with this property is the so-called *period* of the chain. An irreducible chain is called *aperiodic* if the period d equals 1.

An irreducible and aperiodic Markov chain is called *ergodic*. If the chain is ergodic, π is shown to be unique and, more important, $p^{(t)} \to \pi$ as $t \to \infty$, irrespective of $p^{(0)}$. Thus, if the chain is ergodic, π is also referred to as the *limit distribution*.

3.3.3 Detailed balance

We have already mentioned that an MCMC algorithm intends to produce a Markov chain with limit distribution π. To do that, the first step is to construct P's that satisfy general balance (3.4) with respect to π, i.e. we require that:

$$\sum_{x \in S} \pi(z) p(x, z) = \pi(z) \tag{3.5}$$

for all $z \in S$. Formula (3.5) involves a generally intractable summation over the state space S. Luckily, we can use a sufficient condition for general balance, namely *detailed balance*:

$$\pi(x) p(x, z) = \pi(z) p(z, x) \tag{3.6}$$

for all $x, z \in S$. Clearly, it is convenient to check the detailed balance rather than the general one.

Detailed balance is also known as *reversibility* condition. In fact, if a stationary Markov chain satisfies the (3.6), then it is "time reversible": for all states x and z, the rate at which the process goes from x to z is equal to the rate at which it goes from z and x.

3.4 MCMC methods

Over the last years, the considerable spreading of fast and powerful computers has entailed a real explosion of MCMC methods, especially in Bayesian statistics. Within a quite rich literature, we can suggest the books by Gilks et al. (1996), Gamerman (1997) and Robert and Casella (1999), or the papers by Casella and George (1992), Chib and Greenberg (1995) and Besag (2000).

In short, an MCMC algorithm constructs a Markov chain whose stationary distribution is our distribution of interest π. Once the ergodicity of the chain is proved, π can be considered as the limit distribution. Hence, the realized values of the chain are used to make inference about π. For example, the sequence of random variables corresponding to $\bar{g}$ (equation 3.1), still converges almost surely to $E_\pi[g(X)]$ as $m \to \infty$, by the so called ergodic theorem for Markov chains. The underlying theory is quite complicate and it does not concern this thesis: what is important is that we can use empirical averages to produce approximations to expectations under π for sufficiently large m.

3.4.1 Metropolis-Hastings algorithms

The name of this algorithm stems from the papers by Metropolis et al. (1953) and Hastings (1970). Originally, the first version of this algorithm was implemented to calculate properties of chemical substances.

Let Q be a transition probability matrix of a Markov chain with state space S such that:

$$q(x,z) > 0 \quad \Leftrightarrow \quad q(z,x) > 0$$

for all $x, z \in S$. Now define:

$$p(x,z) = q(x,z)\alpha(x,z), \qquad x \neq z \in S \tag{3.7}$$

where $\alpha(x,z) = 0$ if $q(x,z) = 0$ and otherwise

$$\alpha(x,z) = \min\left(1, \frac{\pi(z)q(z,x)}{\pi(x)q(x,z)}\right) \tag{3.8}$$

It is immediate to prove that (3.7) satisfies the detailed balance (3.6) for $x \neq z$. If $\pi(z)q(z,x) > \pi(x)q(x,z)$, $\alpha(x,z) = (\pi(z)q(z,x))/(\pi(x)q(x,z))$ while $\alpha(z,x) = 1$. Substituting in (3.7) and then in (3.6), the detailed balance is achieved. Otherwise, if $\pi(z)q(z,x) < \pi(x)q(x,z)$, then $\alpha(x,z) = 1$ and $\alpha(z,x) = (\pi(x)q(x,z))/(\pi(z)q(z,x))$ and the (3.6) is still verified.

It is important to note that π is a stationary distribution, despite the arbitrariness of q.

In practice, the algorithm works as follows. Suppose that the current value of the chain is $X^{(i)} = x$. The next value for the chain can be $X^{(i+1)} = X^{(i)}$ or $X^{(i+1)} = z$, where z is called *candidate* state and it is generated from $q(x,z)$. The probability that $X^{(i+1)} = z$, i. e. the candidate state is accepted, is $\alpha(x,z)$ defined in (3.8).

Up to now, we have dealt with finite state space Markov chain. Luckily, MCMC methods also work with continuous components. Although the theory must then be rewritten for general state space (e.g. Meyn and Tweedie, 1993), the modifications in practical terms are straightforward. Anyway, the terminology should reflect this change: Q's and P's become transition kernels rather than matrices, with elements that are densities rather than probabilities. For simplicity, the notation will be the same.

There are some particular case of Metropolis-Hastings algorithm. First of all, if we choose a proposal density q such that $q(x,z) = q(z,x)$ for all x and z, the acceptance probability (3.8) becomes:

$$\alpha(x,z) = \min\left(1, \frac{\pi(z)}{\pi(x)}\right)$$

This simpler version of the method is known as *Metropolis algorithm*, because it is the original algorithm in Metropolis et al. (1953).

Another particular case is the *independence Metropolis* algorithm, in which the proposal states are generated independently of the current ones. In other words, $q(x,z) = q(z)$ for all x and z, and

$$\alpha(x,z) = \min\left(1, \frac{\pi(z)q(x)}{\pi(x)q(z)}\right)$$

The success of Metropolis-Hastings algorithm in Bayesian statistic is due to the fact that, if the distribution of interest $\pi(\cdot)$ is a posterior distribution, in the ratio of (3.8) the normalizing constant is cancelled out and its calculation is then avoided. Keeping in mind equation (2.2), a meaningful representation of the ratio in the (3.8), which will be useful in the next chapters, is:

$$(\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio})$$

3.4.2 Componentwise algorithms

Consider now explicitly the multivariate case: $X^{(i)} = (X_1^{(i)}, \ldots, X_k^{(i)})'$, with $x = (x_1, \ldots, x_k)'$ and $z = (z_1, \ldots, z_k)'$ denoting possible values for $X^{(i)}$. In the form we presented Metropolis-Hastings algorithm, $X^{(i)}$ is updated in a single block i. e. at each iteration all the k components are changed simultaneously using the same q. In this section, we shall illustrate the possibility of assigning different transition mechanisms to each component $X_j^{(i)}$, for $j = 1, \ldots, k$.

Instead of a unique p (equation 3.7), consider k different transition probabilities p_j, $j = 1, \ldots, k$, each of them constructed as in simple Metropolis-Hastings and formed by q_j and α_j. The transition probability p_j only updates the jth component.

The acceptance probability α_j is shown to be:

$$\alpha_j(x_j, z_j) = \min\left(1, \frac{\pi(z_1, \ldots, z_k) q(z_j, x_j)}{\pi(x_1, \ldots, x_k) q(x_j, z_j)}\right) \tag{3.9}$$

The move determined by q_j only updates $X_j^{(i)}$, so the other components remain unchanged ($z_{-j} = x_{-j}$). Since $\pi(z_1, \ldots, z_k) = \pi(z_j | z_{-j}) \pi(z_{-j})$,

$$\frac{\pi(z_1, \ldots, z_k)}{\pi(x_1, \ldots, x_k)} = \frac{\pi(z_j | z_{-j})}{\pi(x_j | x_{-j})}$$

Hence, α_j can be simplified to:

$$\alpha_j(x_j, z_j) = \min\left(1, \frac{\pi(z_j | z_{-j}) q(z_j, x_j)}{\pi(x_j | x_{-j}) q(x_j, z_j)}\right) \tag{3.10}$$

Note that the (3.10) uses the full conditionals (section 2.2). If the distribution of interest is a posterior distribution, the normalizing constant of the (2.3) cancels out. Moreover, if the prior is designed with some independence structures, it is formed by products and the factors that do not involve $X_j^{(i)}$ annul each other.

In order to choose the component to update, different mechanisms have been suggested: for example, the choice can be made at random or in a fixed pre-specified order.

3.4.3 Gibbs sampler

The idea underlying the *Gibbs sampler* is quite old. The name of the algorithm comes from Geman and Geman (1984), where it was discussed in image analysis. Nevertheless, the Gibbs sampler can be viewed as a componentwise algorithm in which proposals are made from the full conditional themselves: $q_j(z_j, x_j) = \pi(x_j|x_{-j})$. Substituting in (3.10), the ratio becomes one and the candidate state is always accepted.

Operationally, the algorithm can be described in the following way:

1. Initialize the iteration counter $i = 1$ and set arbitrary initial values $X^{(1)} = (X_1^{(1)}, \ldots, X_k^{(1)})$.

2. Generate a new value $X^{(i+1)} = (X_1^{(i+1)}, \ldots, X_k^{(i+1)})$ from $X^{(i)}$ through successive generation of values

$$\begin{aligned}
X_1^{(i+1)} &\sim \pi_1(X_1^{(i+1)}|X_2^{(i)}, X_3^{(i)}, \ldots, X_k^{(i)}) \\
X_2^{(i+1)} &\sim \pi_2(X_2^{(i+1)}|X_1^{(i+1)}, X_3^{(i)}, \ldots, X_k^{(i)}) \\
&\cdots \\
X_k^{(i+1)} &\sim \pi_k(X_k^{(i+1)}|X_1^{(i+1)}, X_2^{(i+1)}, \ldots, X_{k-1}^{(i+1)})
\end{aligned}$$

3. Increase the counter $i = i + 1$ and return to step 2.

3.4.4 Implementation of MCMC algorithms

The choice of the algorithm is the first issue when dealing with MCMC methods. Unfortunately, there is not a general rule and it depends on the nature of the problem under study. For instance, when it is easy to write down the full conditionals, the Gibbs sampler is the most natural choice. Otherwise, the possibility to “tune”the proposals is often a desirable feature of the Metropolis-Hastings algorithm. Sometimes the best solution can be a *hybrid* algorithm,

i. e. a componentwise algorithm in which some components are updated with Gibbs type moves.

Once the algorithm is chosen, one of the most important problems is the convergence rate. Despite the theoretical results ensuring the convergence of an MCMC method, its practical implementation deals with topics like the sample size (how many iterations should the algorithm run?) and the *burn-in* (since the initial states of the chain are necessarily arbitrary, after how many iterations does the algorithm reach the convergence?). In literature, many *convergence diagnostics* have been proposed: a group of them are based on the study of the properties of the observed output from the chain. For instance, see Cowles and Carlin (1996) or Robert (1995).

3.4.5 Inference using MCMC output

After a simulation, whatever the MCMC algorithm chosen, a sample from the distribution of interest π is available. Suppose π is a posterior distribution for $\theta = (\theta_1, \dots, \theta_k)$ and suppose now to denote the sample by $\theta_1^{(i)}, \dots, \theta_k^{(i)}$, for $i = 1, \dots, N$. As we have already mentioned, we are able to estimate expected values of the form:

$$E_\pi[g(\theta)] = \int_\Theta g(\theta)\pi(\theta)d\theta \tag{3.11}$$

by the corresponding estimators based on the sample:

$$\frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)})$$

Some particular cases of the (3.11) allows us to obtain characteristics of π from the sample. First of all, if $g(\theta) = \theta$, then the (3.11) is the vector of the posterior means μ and it is estimated simply by:

$$\frac{1}{N}\sum_{i=1}^{N} \theta^{(i)}$$

If $g(\theta) = (\theta - \mu)(\theta - \mu)'$, then we have the posterior variance and covariance matrix. If $g(\theta) = I_C(\theta)$, where $I_C(\theta)$ stands for the indicator function (i.e. it is 1 if $\theta \in C$ and 0 otherwise), then we obtain the posterior probability of a set C. Furthermore, letting $g(\theta) = p(y|\theta)$, where y can denote a 'future' observation, we have the predictive distribution (section 2.2) for y.

Credibility intervals are similarly obtained by estimating the interval limits by the respective sample quantiles.

If we are interested in the marginal posterior density of a component θ_j, we can estimate it by (a smoothed version of) the histogram of sampled values of θ_j. Where additional information about π are available, better estimators can be obtained by using conditional distributions (see, Gelfand and Smith, 1990).

Chapter 4

MCMC and Bayesian model determination

4.1 Introduction

Up to now, we presented analysis and inference procedures that deal with evaluation of a given model. In this chapter, the problem of choosing between models is treated in a Bayesian perspective: the key idea is to index all the models under consideration and to view this index as another parameter.

More formally, consider a collection $\mathcal{M}$ of candidate models indexed by $m = 1, \ldots, M$. Let θ_m be the parameters related to the model m with $\theta_m \in \Theta_m$. The sampling distribution is now defined by $f(y|\theta_m, m)$: note that each model specifies this distribution apart from the unknown parameter vector θ_m.

Conditionally to a given model m, Bayes' theorem (equation 2.1) gives the posterior for θ_m:

$$p(\theta_m|m, y) = \frac{L(\theta_m, m)p(\theta_m|m)}{\int L(\theta_m, m)p(\theta_m|m)d\theta_m} \tag{4.1}$$

where $L(\theta_m, m)$ is the likelihood (i.e. the sampling distribution $f(y|\theta_m, m)$ regarded as a function of θ and m) and where $p(\theta_m|m)$ is the prior distribution conditional on model m.

To compare between models, it is natural to use the marginal posterior distribution of m which is derived by Bayes' theorem:

$$p(m|y) = \frac{f(y|m)p(m)}{\sum_{m=1}^{M} f(y|m)p(m)} \tag{4.2}$$

where $p(m)$ is a discrete prior for the models and where $f(y|m)$ is the *marginal likelihood*:

$$f(y|m) = \int L(\theta_m, m)p(\theta_m|m)d\theta_m \tag{4.3}$$

Unfortunately, the calculation of the normalizing constant of the (4.2) poses the usual problems in terms of analytical tractability, especially with a high number of possible models.

In the following, we present some methods based on simulations that are able to deal with model selection. We divide them into two main categories: *across-* and *within-* model simulation methods.

The across-model simulation approach is based on an MCMC simulation with states of the form (m, θ_m). The distribution of interest is the joint posterior of the parameters and the model index. The marginal posterior distribution of m is simply estimated by the proportions of m's, for $m = 1, \ldots, M$, in the sample obtained by the MCMC algorithm. By the conditional posterior distribution $p(\theta_m|m, y)$, we can make inference within each model. A sample of $p(\theta_m|m, y)$ is obtained considering only the sampled values for which the model is m. The acronym *MCMCMC* or MC^3 is often used to indicate this class of techniques and it stands for Markov chain Monte Carlo model composition. For a study about the connection between some of them, see Dellaportas et al. (2002).

In the within-model simulations, the aim of finding $p(m|y)$ for all m is reached by estimating all the marginal likelihoods $f(y|m)$. Once the $f(y|m)$ for all m are estimated, it is sufficient to normalize the products $f(y|m)p(m)$ to achieve the marginal posterior probabilities (4.2). As it will be clear later on, the idea of the within-model simulation methods is then to estimate the marginal likelihoods separately for each m, using samples for the within-model posteriors $p(\theta_m|y, m)$ from traditional MCMC algorithms.

In the analysis of the mixture of autoregressive models (chapter 6), model selection will be performed using a combination of an across- and a within-model simulation method.

It is important to note that Bayesian model determination entails some advantages with respect to other approaches. First of all, the most obvious one is the simplicity of the interpretation of the results: conclusions like "the (posterior) probabilities that m and m' are true are 0.87 and 0.13 respectively" are easy to interpret even with a limited statistical background.

In addition, Bayesian model determination acts as an automatic "Occam's razor", selecting a simpler model over a more complex model if both are compatible with the data.

Another fundamental feature is that one can account for model uncertainty. Standard approach selects a single model from a class of candidate models and then makes inference based on this model. This procedure ignores model uncertainty and it could provide small predictive precisions (see Draper, 1995, for a discussion). Conversely, Bayesian model determination can take into account model uncertainty because one can maintain consideration of several models, with the input of each into the analysis weighted by the model posterior probability.

4.2 Across-model simulation

The state space for an across-model simulation is $\Theta = \{\Theta_m \times \mathcal{M}\}$. It could seem a harmless generalization of the state space for the traditional MCMC algorithms of chapter 3: actually, this new space is non-standard because the dimension of parameters θ_m can depend on the model.

4.2.1 Independence sampler and pilot MCMC

The most natural approach to model determination using MCMC consists, anyway, in applying directly the Metropolis-Hastings algorithm (section 3.4.1) over the joint space of θ_m and m in order to simulate the posterior $p(m, \theta_m|y)$.

Suppose the current state of the chain is (θ_m, m). A proposal state $(\theta'_{m'}, m')$ is generated from the density $q((\theta_m, m), (\theta'_{m'}, m'))$ with respect to the natural measure on Θ and it is accepted with probability:

$$\begin{aligned}\alpha((\theta_m, m), (\theta'_{m'}, m')) &= \min\left(1, \frac{p(\theta'_{m'}, m'|y)q((\theta'_{m'}, m'), (\theta_m, m))}{p(\theta_m, m|y)q((\theta_m, m), (\theta'_{m'}, m'))}\right) \\ &= \min\left(1, \frac{L(\theta'_{m'}, m')p(\theta'_{m'}|m')p(m')q((\theta'_{m'}, m'), (\theta_m, m))}{L(\theta_m, m)p(\theta_m|m)p(m)q((\theta_m, m), (\theta'_{m'}, m'))}\right) \qquad (4.4)\end{aligned}$$

The first line of the (4.4) is a simple generalization of the (3.8); since the target distribution is a posterior, it is rewritten in the product of likelihood and prior, where the (joint) prior $p(\theta_m, m)$ is decomposed further in $p(\theta_m|m)p(m)$.

As in Gruet and Robert (1997), the proposal is usually constructed as a proposal for the model followed by a conditional proposal for the model parameters:

$$q((\theta_m, m), (\theta'_{m'}, m')) = q(m'|m, \theta_m)q(\theta'_{m'}|m', m, \theta_m)$$

Because of the general difficulty to find a proposal distribution of this kind, Tierney (1994) suggested a special case of this approach called *independence sampler*. The proposed values are independent with respect to the current ones and then:

$$q((\theta_m, m), (\theta'_{m'}, m')) = q(\theta'_{m'}, m')$$

The independence sampler is straightforward to implement but, apart from the rare case in which q is a reasonable approximation of the target distribution, it may be not efficient.

In order to choose q, *pilot MCMC* runs can be used. The idea is to simulate the conditional posteriors $p(\theta_m|m, y)$ for each model $m \in \mathcal{M}$ by a standard MCMC algorithm on the parameter space θ_m. These pilot runs construct approximations of $p(\theta_m|m, y)$ that are used in forming proposal $q(\theta'_{m'}|m')$. For instance, q can be a normal density with moments calculated on the MCMC output. This method should work well if the conditional posteriors are reasonable unimodal. Clearly, if the number of possible models is not small, this approach is not feasible.

4.2.2 Reversible jump algorithm

Green (1995) developed an MCMC strategy, called *reversible jump* algorithm, which allows the proposal to depend on the current values of the chain. This method creates a Markov chain which can "jump" between models with parameter spaces of different dimension.

Suppose to denote the dimension of a parameter vector θ_m with $d(\theta_m)$. In practice, the reversible jump works as follows. Let (θ_m, m) be the current value of the chain and let $g_{m,m'}$ be an invertible function:

- A proposal model m' is generated with probability $q(m'|m)$.
- Generate a random vector u of dimension $d(u)$ from a proposal density $q(u|\theta_m, m, m')$.
- Set $(\theta'_{m'}, u') = g_{m,m'}(\theta_m, u)$, with $d(u') = d(\theta_m) + d(u) - d(\theta'_{m'})$
- Accept the proposal values $(\theta'_{m'}, m')$ with probability:

  $$\alpha((\theta_m, m), (\theta'_{m'}, m')) = \min(1, R)$$

  with:

  $$\begin{aligned} R = & \frac{L(\theta'_{m'}, m')}{L(\theta_m, m)} \times \frac{p(\theta'_{m'}|m')p(m')}{p(\theta_m|m)p(m)} \times \frac{q(m|m')q(u'|\theta'_{m'}, m', m)}{q(m'|m)q(u|\theta_m, m, m')} \\ & \times \left| \frac{\partial g_{m,m'}(\theta_m, u)}{\partial(\theta_m, u)} \right| \end{aligned} \quad (4.5)$$

A useful representation of the ratio in the (4.5) is:

(likelihood ratio) × (prior ratio) × (proposal ratio) × (jacobian)

In some cases, the jacobian in the (4.5) is one. For instance, if all the parameters of the proposed model are generated directly from a proposal distribution, then $(\theta'_{m'}, u') = (\theta_m, u)$, with $d(\theta_m) = d(u')$ and $d(\theta'_{m'}) = d(u)$. Since the (4.5) becomes equivalent to the (4.4), the independence sampler can be viewed as a special case of reversible jump.

Note also that if $m = m'$, the move is a standard Metropolis-Hastings step.

Another simplified version of the algorithm occurs when $d(u') = 0$: in this case, $\theta'_{m'} = g_{m,m'}(\theta_m, u)$ and the "dimension matching" is achieved since $d(u) = d(\theta'_{m'}) - d(\theta_m)$. The ratio (4.5) of the acceptance probability becomes:

$$R = \frac{L(\theta'_{m'}, m')p(\theta'_{m'}|m')p(m')q(m|m')}{L(\theta_m, m)p(\theta_m|m)p(m)q(m'|m)q(u|\theta_m, m, m')} \left| \frac{\partial g_{m,m'}(\theta_m, u)}{\partial(\theta_m, u)} \right|$$

Furthermore, the proposal density q is often independent, i. e.:

$$q(u|\theta_m, m, m') = q(u)$$

The reversible jump algorithm constructs a Markov chain retaining detailed balance (equation 3.6) and then ensuring the correct limiting distribution. For the formal proof see Green(1995).

As a simple example, consider only two models m_1 and m_2 with parameters θ_1 and θ_2. Suppose that $d(\theta_2) > d(\theta_1)$ and $d(u) = d(\theta_2) - d(\theta_1)$. Suppose also that the random quantity u is generated independently from the state of the chain. The usual strategy to implement the reversible jump move is to design it in tandem, forming a reversible pair. Thus, the acceptance probability of the move from m_1 to m_2 is given by:

$$\min\left(1, \frac{L(\theta_2, m_2)p(\theta_2|m_2)p(m_2)q(m_1|m_2)}{L(\theta_1, m_1)p(\theta_1|m_1)p(m_1)q(m_2|m_1)q(u)}\left|\frac{\partial g(\theta_1, u)}{\partial(\theta_1, u)}\right|\right)$$

Likewise, the reverse move from m_2 to m_1 has an acceptance probability given by:

$$\min\left(1, \frac{L(\theta_1, m_1)p(\theta_1|m_1)p(m_1)q(m_2|m_1)q(u)}{L(\theta_2, m_2)p(\theta_2|m_2)p(m_2)q(m_1|m_2)}\left|\frac{\partial(\theta_1, u)}{\partial\theta_2}\right|\right)$$

The choice of the function g is crucial in terms of efficiency of the algorithm because of its role in forming the proposal values. g is usually chosen according to informal considerations suggesting a reasonable probability of acceptance (see for example the ” “moment matching” expedient of section 5.4.3).

For an illustration of some strategies for the construction of reversible jump proposals, see Brooks et al. (2003).

4.2.3 Carlin and Chib method

The following approach to model choice is based on the Gibbs sampler idea and it was introduced by Carlin and Chib (1995).

Let ψ be the collection of all the parameters for every model, $\psi = (\theta_1, \ldots, \theta_M)$. Note that the joint distribution of all random quantities is:

$$p(y, \psi, m) = f(y|\psi, m)p(\psi|m)p(m)$$

Now assume that:

$$p(\psi|m) = \prod_{i=1}^{M} p(\theta_i|m)$$

i. e. θ_m are conditionally independent given m.

The prior $p(\theta_i|m)$, for $i \neq m$, is called *pseudo prior* and it specifies the distribution of the parameters of model i given another model m.

As we mentioned in the introduction, $f(y|\psi, m) = f(y|\theta_m, m)$. Thus, the joint distribution becomes:

$$p(y, \psi, m) = f(y|\theta_m, m) \prod_{i=1}^{M} p(\theta_i|m)p(m)$$

and since $p(y, \psi, m) \propto p(\psi, m|y)$ we can propose the following strategy for a Gibbs sampler:

- Full conditional for θ:

$$p(\theta_m|\theta_{-m}, i, y) \propto \begin{cases} f(y|\theta_m, m)p(\theta_m|m), & i = m \\ p(\theta_m|i), & i \neq m \end{cases}$$

  for $m = 1, \ldots, M$.

- Full conditional for m:

$$p(m|\psi, y) \propto f(y, \psi, m) = c^{-1} f(y|\theta_m, m) \prod_{i=1}^{M} p(\theta_i|m)p(m)$$

  for $m = 1, \ldots, M$ and where:

$$c = \sum_{l=1}^{M} f(y|\theta_l, l) \prod_{i=1}^{M} p(\theta_i|l)p(l)$$

Clearly, it may not be possible to sample directly for some of the θ_m: in such situations, a Metropolis-Hastings move may be used.

This method is not free from difficulties. First of all, the choice of the pseudo priors affects the rate of convergence of the chain and it must be done with care. In addition, the algorithm shows sensitivity with respect to the model prior specifications and for certain priors the chain does not seem to move between models. Finally, the method is not applicable to the case of a countable number of models.

4.2.4 Other across-model approaches

In the literature, several across-model simulation methods were presented. Anyway, they are often closely related to reversible jump.

Grenander and Miller (1994) propose an algorithm with two kinds of move (between-model jumps and within-model diffusion) using a Langevian stochastic differential equation.

Another approach is based on *jump diffusions* (Phillips and Smith, 1996). The key idea is to take into account model uncertainty by introducing a joint prior probability distribution over both the set of possible models and the parameters of those models. The resulting posterior distribution is achieved by an iterative jump-diffusion sampling algorithm. A jump is a discrete transition between models of different dimensionality.

Geyer and Moller (1994) propose a Metropolis-Hastings sampler which constructs a continuous-time Markov chain in order to simulate certain spatial point processes (Ripley, 1977, originally investigated this idea using “birth and death” process).

The theory of point processes is also used by Stephens (2000): in particular, the idea is to view the parameters of the model as point process, in order to create a Markov birth-death process with an appropriate stationary distribution. For the connection between this method and the reversible jump see Cappé et al. (2001).

4.3 Within-model simulation

As we mentioned in the introduction of this chapter, within-model simulation techniques calculate the posterior distribution of the model $p(m|y)$ through the estimation of the marginal likelihood $f(y|m)$ for all m using a sample from the posterior $p(\theta_m|y, m)$. Then we suppose to have such a sample generated, for instance, by a traditional MCMC of chapter 3.

For notational convenience, the model index m will be suppressed in the rest of the chapter.

4.3.1 Importance sampling estimators of the marginal likelihood

From equation (4.3), the marginal likelihood can be viewed as the following expected value:

$$f(y) = \int L(\theta)p(\theta)d\theta = E_{prior}[L(\theta)] \tag{4.6}$$

A chance to estimate $f(y)$ consists in using the importance sampling method (section 3.2.2). Thus, if we have samples $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ from the auxiliary distribution $\pi^* = c^*\ h^*$, $E_{prior}[L(\theta)]$ is estimated by (see equation 3.3):

$$\frac{1}{N}\sum_{i=1}^{N}\frac{p(\theta^{(i)})L(\theta^{(i)})}{h^*(\theta^{(i)})}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{p(\theta^{(i)})}{h^*(\theta^{(i)})}\right]^{-1}$$

or in a more compact form:

$$\frac{\|pL/h^*\|_{\pi^*}}{\|p/h^*\|_{\pi^*}}$$

where $\|g\|_f = N^{-1}\sum_{i=1}^{N} g(\theta^{(i)})$, where g is a function and $\theta^{(i)}$ is a sample of size N from $f(\theta)$.

Different choices for h^* lead to different estimators for the marginal likelihood. The simplest one is the prior, that is $h^* = p$; the estimator is then:

$$\|L\|_{prior}$$

Note that it is a simple average of the likelihoods of a sample from the prior and, in other words, it is the Monte Carlo estimator of the (4.6). This estimator was investigated in particular cases by McCulloch and Rossi (1991). The problem with it is that most of $\theta_m^{(i)}$ have small likelihood values if the posterior is much more concentrated than the prior: as a consequence, the process is quite inefficient.

If we choose the posterior as the auxiliary distribution (that is, $h^* = Lp$), the estimator is:

$$\|1/L\|^{-1}_{\text{posterior}}$$

This harmonic mean of the likelihood values (Newton and Raftery, 1994) can be computationally unstable because of the occasional occurrence of $\theta_m^{(i)}$ with small likelihood.

Another possibility consists in using a mixture of the prior and posterior densities ($h^*(\theta) = wp(\theta_m) + (1-w)p(\theta_m|y,m)$), in order to implement a sort of compromise between the two previous estimators. Clearly, one must simulate from the prior as well as the posterior. Newton and Raftery (1994) show how to avoid that using an iterative scheme.

Gelfand and Dey (1994) mention a modification of the previous harmonic mean, i.e.:

$$\|f/Lp\|^{-1}_{\text{posterior}}$$

where f is a function of θ_m and it is any probability density. The efficiency of this estimator depends on how f is close to the posterior.

The *bridge sampling* technique, originally proposed by Meng and Wong (1996), is another specification of the importance sampling estimator of the marginal likelihood: if g is a positive function, their estimator is:

$$\frac{\|pLg\|_{prior}}{\|p\ g\|_{\text{posterior}}}$$

The optimal choice of g can be computed from an initial guess. The disadvantage is the need of simulating from the prior as well as the posterior.

4.3.2 Marginal likelihood from the Gibbs output

Chib (1995) proposed a method to estimate the marginal likelihood using the output generated by a Gibbs sampler (section 3.4.3).

The starting idea is that the marginal likelihood is the normalizing constant of the posterior density (see equations 4.3 and 4.1). Thus, we can write:

$$f(y) = \frac{L(\theta)p(\theta)}{p(\theta|y)} \tag{4.7}$$

The (4.7) is called *basic marginal likelihood identity.* Note that this identity is true for every θ: this means that we can estimate the marginal likelihood by finding an estimate of the posterior ordinate $p(\theta^*|y)$ in a single point θ^*. Thus, using the computationally convenient logarithm scale and indicating the estimate of $p(\theta^*|y)$ by $\bar{p}(\theta^*|y)$:

$$\ln\bar{f}(y) = l(\theta^*) + \ln p(\theta^*) - \ln\bar{p}(\theta^*|y) \tag{4.8}$$

where $\bar{f}(y)$ is the estimate of the marginal likelihood and $l(\theta^*)$ is the loglikelihood. For estimation efficiency, the point θ^* is taken to be a high-density point in the support of the posterior.

What it is necessary to do now is to produce the estimate $\bar{p}(\theta^*|y)$. After that, all (4.8) requires is the evaluation of the loglikelihood function and the prior. The estimate does not suffer from any instability problem. The estimation error is derived in Chib (1995).

Suppose θ is split into B blocks $\theta = (\theta_1, \theta_2, \ldots, \theta_B)$ and suppose we are able to write the full conditionals for each of them. We can also consider a latent variable, say z, which will be useful in the analysis of mixture models (chapters 5 and 6).

The complete set of full conditional is then:

$$\begin{aligned} &p(\theta_r|\theta_{-r}, z, y), \qquad r = 1, \ldots, B \\ &p(z|\theta, y) \end{aligned} \tag{4.9}$$

where $\theta_{-r} = (\theta_1, \ldots, \theta_{r-1}, \theta_{r+1}, \ldots, \theta_B)$.

Let $\eta_{r-1} = (\theta_1, \ldots, \theta_{r-1})$ and $\eta^{r+1} = (\theta_{r+1}, \ldots, \theta_B)$. The posterior density $p(\theta^*|y)$ can be decomposed as:

$$p(\theta^*|y) = \prod_{r=1}^{B} p(\theta_r^*|\eta_{r-1}^*, y) \tag{4.10}$$

Note that each term of the (4.10) is:

$$\begin{aligned} p(\theta_r^*|\eta_{r-1}^*, y) &= \int p(\theta_r^*, \eta^{r+1}, z|\eta_{r-1}^*, y)\ d\eta^{r+1}dz \\ &= \int p(\theta_r^*|\eta_{r-1}^*, \eta^{r+1}, z, y)p(\eta^{r+1}, z|\eta_{r-1}^*, y)\ d\eta^{r+1}dz \end{aligned}$$

and its Monte Carlo estimator is:

$$\bar{p}(\theta_r^*|\eta_{r-1}^*, y) = N^{-1} \sum_{i=1}^{N} p(\theta_r^*|\eta_{r-1}^*, \eta^{r+1,(i)}, z^{(i)}, y) \tag{4.11}$$

if $\{\eta^{r+1,(i)}, z^{(i)}\}$, for $i = 1, \ldots, N$, are samples from $p(\eta^{r+1}, z|\eta_{r-1}^*, y)$.

The method estimates each term of the (4.10) by successive "reduced" Gibbs samplers, i.e. with a decreasing number of full conditionals. Starting with $r = 1$, it can be represented by the following steps:

1. Sample $\{\eta^{r,(i)}, z^{(i)}\}$, for $i = 1, \ldots, N$, from the Gibbs sampler conditional to η_{r-1}^*, i.e. with distribution of interest $p(\eta^r, z|\eta_{r-1}^*, y)$.

2. Set:
$$\bar{p}(\theta_r^*|\eta_{r-1}^*, y) = N^{-1} \sum_{i=1}^{N} p(\theta_r^*|\eta_{r-1}^*, \eta^{r+1,(i)}, z^{(i)}, y)$$

3. $r = r + 1$ and go to step 1 until $r = B$.

In the end, the estimate of the loglikelihood is:

$$\ln\bar{f}(y) = l(\theta^*) + \ln p(\theta^*) - \sum_{r=1}^{B} \ln\bar{p}(\theta_r^*|\eta_{r-1}^*, y) \tag{4.12}$$

Since the samples $\{\eta^{r+1,(i)}, z^{(i)}\}$ are drawn from the Gibbs of step 1, they are marginally from $p(\eta^{r+1}, z|\eta_{r-1}^*, y)$ and the (4.11) is indeed the Monte Carlo estimator of $p(\theta_r^*|\eta_{r-1}^*, y)$.

4.3.3 Marginal likelihood from the Metropolis-Hastings output

To perform the previous method, it is necessary to have all the full conditionals in closed form. Chib and Jeliazkov (2001) provide a generalization which overcomes this problem and estimates the marginal likelihood using the output from a Metropolis-Hastings algorithm (section 3.4.1).

To illustrate and prove the method, we begin with the simple case in which the posterior density is sampled in one block by the Metropolis-Hastings algorithm. The produced sample is $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$, where $\theta^{(i)} = (\theta_1^{(i)}, \ldots, \theta_k^{(i)})$, for $i = 1, \ldots, N$. From equation (3.8), the acceptance probability of the algorithm is:

$$\alpha(\theta, \theta') = \min\left(1, \frac{L(\theta')p(\theta')}{L(\theta)p(\theta)} \frac{q(\theta', \theta)}{q(\theta, \theta')}\right)$$

where θ' is the candidate state and q is the usual proposal density. As we know, Metropolis-Hastings method satisfies the detailed balance (equation 3.6), thus we can write:

$$\alpha(\theta, \theta^*)q(\theta, \theta^*)p(\theta|y) = \alpha(\theta^*, \theta)q(\theta^*, \theta)p(\theta^*|y)$$

for any point θ^*. Integrating both sides of this expression with respect to θ, we obtain:

$$\begin{aligned} p(\theta^*|y) &= \frac{\int \alpha(\theta, \theta^*)q(\theta, \theta^*)p(\theta|y)d\theta}{\int \alpha(\theta^*, \theta)q(\theta^*, \theta)d\theta} \\ &= \frac{E_1[\alpha(\theta, \theta^*)q(\theta, \theta^*)]}{E_2[\alpha(\theta^*, \theta)]} \end{aligned}$$

where the expectation E_1 is with respect to the posterior $p(\theta|y)$ while the expectation E_2 is with respect to $q(\theta^*, \theta)$. The posterior ordinate is then estimated by the Monte Carlo estimator:

$$\bar{p}(\theta^*|y) = \frac{N^{-1}\sum_{i=1}^{N} \alpha(\theta^{(i)}, \theta^*)q(\theta^{(i)}, \theta^*)}{J^{-1}\sum_{j=1}^{R} \alpha(\theta^*, \theta^{(j)})}$$

where $\theta^{(i)}$, for $j = 1, \ldots, N$, are the samples from the posterior and $\theta^{(j)}$, for $j = 1, \ldots, J$, are draws from $q(\theta^*, \theta)$, given the fixed value θ^*.

The marginal likelihood is then estimated by

$$\text{ln}\bar{f}(y) = l(\theta^*) + \text{ln}p(\theta^*) - \text{ln}\bar{p}(\theta^*|y)$$

Consider now the general case in which θ is split into B blocks $\theta = (\theta_1, \theta_2, \ldots, \theta_B)$ as in the previous section. Consider also the latent variable z. Suppose a componentwise Metropolis-Hastings algorithm (section 3.4.2) is available: each block is updated with a probability of the form (equation 3.10):

$$\begin{aligned} \alpha(\theta_r, \theta') &= \min\left(1, \frac{p(\theta'|\theta_{-r}, z, y)q(\theta', \theta_r)}{p(\theta_r|\theta_{-r}, z, y)q(\theta_r, \theta')}\right) \\ &= \min\left(1, \frac{L(\theta', \theta_{-r}, z)p(\theta', \theta_{-r})q(\theta', \theta_r)}{L(\theta_r, \theta_{-r}, z)p(\theta_r, \theta_{-r})q(\theta_r, \theta')}\right) \end{aligned}$$

Let $\eta_{r-1} = (\theta_1, \dots, \theta_{r-1})$ and $\eta^{r+1} = (\theta_{r+1}, \dots, \theta_B)$. The posterior ordinate at a given point θ^* is decomposed as in (4.10) and, using an analogous argument to the previous single-block case, each term of it is equal to

$$p(\theta_r^*|\eta_{r-1}^*, y) = \frac{E_1[\alpha(\theta_r, \theta_r^*)q(\theta_r, \theta_r^*)]}{E_2[\alpha(\theta_r^*, \theta_r)]} \tag{4.13}$$

where the expectation E_1 is with respect to $p(\theta_r, \eta^{r+1}, z|\eta_{r-1}^*, y)$ and the expectation E_2 is with respect to the product $p(\eta^{r+1}, z|\eta_r^*, y)q(\theta_r^*, \theta_r)$.

Suppose now to have samples $\{\eta^{1,(i)}, z^{(i)}\}$, for $i = 1, \dots, N_1$, from the available componentwise Metropolis-Hastings algorithm. To estimate the two integrals in (4.13), the following steps must be performed (start with $r = 1$):

1. Set $\eta_r^* = (\eta_{r-1}^*, \theta_r^*)$ and sample $\{\tilde{\eta}^{r+1,(i)}, \tilde{z}^{(i)}\}$, for $i = 1, \dots, N_{r+1}$, from the reduced Metropolis-Hastings algorithm with distribution of interest $p(\eta^{r+1}, z|\eta_r^*, y)$. At each step of the sampling also draw $\tilde{\theta}_r^{(i)}$ from $q(\theta_r^*, \theta_r)$.

2. Set:
$$\bar{p}(\theta_r^*|\eta_{r-1}^*, y) = \frac{N_r^{-1}\sum_{i=1}^{N_r} \alpha(\theta_r^{(i)}, \theta_r^*)q(\theta_r^{(i)}, \theta_r^*)}{N_{r+1}^{-1}\sum_{i=1}^{N_{r+1}} \alpha(\theta_r^*, \tilde{\theta}_r^{(i)})} \tag{4.14}$$

3. Set $\eta^{r+1,(i)} = \tilde{\eta}^{r+1,(i)}$ and $z^{(i)} = \tilde{z}^{(i)}$, for $i = 1, \dots, N_{r+1}$.

4. Set $r = r + 1$ and go to step 1 until $r = B$.

Note that in equation (4.14), samples $\theta^{r,(i)}$ are from $p(\theta_r, \eta^{r+1}|\eta_{r-1}^*, y)$ while $\tilde{\theta}^{r,(i)}$ are from $p(\eta^{r+1}, z|\eta_r^*, y)q(\theta_r^*, \theta_r)$, thus the (4.14) is the Monte Carlo estimator of the (4.13).

When all the terms in (4.10) are estimated by these reduced simulations, the marginal likelihood on the log scale is estimated as in (4.12).

In the analysis of the mixture of autoregressive models of chapter 6, the set of parameters will be split into several blocks: some of them will be updated by Metropolis-Hastings moves and others by Gibbs moves. As a consequence, a combination of the last two methods will be used, depending on the type of the move.

Chapter 5

Mixture models

5.1 Introduction

Historically, the concept of finite mixture distributions dates back to the 19th century. Since the two pioneer works of Newcomb (1886) and Pearson (1894), the interest in this framework has been lively and sustained. Mixture modelling knew a wide range of applications: medical diagnostics, geography, agriculture, astronomy, economics, etc. It provides a natural mean for the formalization of heterogeneity, when the observed phenomena are too intricate to be described by simple probabilistic modelling through classical distributions. The aspects of mixture models establishes links with cluster analysis, latent structures, detection of outliers, robustness analysis, density estimation in semiparametric approaches and so on.

Nevertheless, statistical analysis of mixtures is not straightforward. In general, there is no explicit formulae for estimators of the various parameters. In addition, the geometry of the parameter space often poses non-standard problems.

The literature is quite rich. Possible references are Everitt and Hand (1981), Titterington et al. (1985) and Lindsay (1995).

5.2 Basic concepts

The basic *mixture model* for independent observations y_i is defined as:

$$y_i|w,\theta \overset{iid}{\sim} \sum_{j=1}^{k} w_j f(y_i|\theta_j), \qquad i = 1,\dots,n \tag{5.1}$$

where $w = (w_1,\dots,w_k)$, with:

$$w_j > 0, \quad j = 1,\dots,k; \qquad w_1 + \cdots + w_k = 1$$

and $\theta = (\theta_1,\dots,\theta_k)$. The densities $f(y_i|\theta_j)$ are called *component densities* and the parameters $w_1,\dots,w_k$ are the *mixing weights*, or simply *weights*, of the mixture model. k is the number of components of the mixture and it is initially assumed fixed (for this reason we shall omit it from the conditioning set of variables).

The so-called *direct application* of finite mixture models refers to situations in which we believe in the existence of k underlying categories and each of the observed variables y_i belongs to only one of these categories. From this point of view, $f(y_i|\theta_j)$ represents the probability density of y_i given that the observation comes from category j; the probability that each observation comes from the j-th component is w_j.

Typically, we do not usually observe the component of y_i directly. Consider now a latent variable $z = (z_1,\dots,z_n)$ which is a "component label": specifically, for $i = 1,\dots,n$, $z_i = j$ if the ith observation comes from the jth component. Probabilistically, we assume that z_i are discrete random variables independently drawn from the discrete distribution:

$$p(z_i|w) = \sum_{j=1}^{k} w_j I_{(z_i=j)}, \qquad z_i = 1,2,\dots,k \tag{5.2}$$

where $I_{(A)}$ denotes the indicator function of the event A. Alternatively, $Prob(z_i = j|w) = w_j$. Note that, given z_i and θ, the observations are drawn from their respective individual subpopolations:

$$y_i|z_i,\theta \sim \sum_{j=1}^{k} f(y_i|\theta_j) I_{(z_i=j)} \tag{5.3}$$

for $i = 1,\dots,n$.

5.3 Bayesian estimation

Several statistical approaches were implemented in order to make inference for mixture models but unfortunately all of them have to deal with a number of potential problems.

For instance, the method of moments may not allow explicit or unique solutions. Furthermore, it is possible to show that moment estimators may not be asymptotically efficient.

The maximum likelihood estimation often leads to non analytic treatment, posing computational problems that are not always straightforward. Typically, the analysis is based on numerical procedures like the *EM algorithm*, the *Newton-Raphson algorithm* and the *method of scoring*.

For a review of these and other non Bayesian approaches, see Titterington et al. (1985).

In the following, we shall focus on a Bayesian analysis of mixtures. If the prior distributions are proper (a distribution is called *improper* if it does not integrate to unity), Bayes estimators are well-defined. Nevertheless, it is only after the advent of MCMC methods that the implementation of the Bayesian approach has became practically feasible.

5.3.1 Analysis for the exponential family

Assume that all the component densities of the mixture belong to the *exponential family*:

$$f(y_i|\theta_j) = a(y_i)\exp\{y_i\,\theta_j - b(\theta_j)\} \tag{5.4}$$

for $j = 1, \ldots, k$. This allows us to carry out a conjugate analysis (section 2.3). Specifically assume that the parameters θ_j are independent with prior distribution equal to:

$$p(\theta_j) \propto \exp\{r_j'\,\theta_j - \lambda_j\,b(\theta_j)\} \tag{5.5}$$

where, for $j = 1, \ldots, k$, r_j is a known vector of the same length as θ_j and λ_j is a known scalar hyperparameter.

For the mixing weights, assume a Dirichlet distribution (section 2.5.4):

$$w \sim \mathrm{Di}(w|\delta_1, \ldots, \delta_k) \tag{5.6}$$

with known hyperparameters δ_j, $j = 1, \ldots, k$.

Note that the parameter set of the model is (w, θ) and the likelihood is given by:

$$L(w,\theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f(y_i|\theta_j) \tag{5.7}$$

By considering the allocation variable z, posterior expectations of (w, θ) can be written in closed form (Diebolt and Robert, 1990), but it leads to intractable calculations, since the posterior distribution takes into account all the k^n partitions of the sample (for instance, with only 50 observations, a simple model with two component requires the calculation of about $11 \cdot 10^{14}$ terms).

Luckily, Bayesian analysis can be implemented straightforwardly using MCMC, in particular with Gibbs sampler.

5.3.2 Gibbs sampler implementation

The full conditionals required by the Gibbs sampler (section 3.4.3) are easily calculated if we consider the allocation variable z. First of all, the likelihood becomes:

$$\begin{aligned} L(\theta, z) &= \prod_{i=1}^{n} \sum_{j=1}^{k} f(y_i|\theta_j) I_{(z_i=j)} \\ &= \prod_{j=1}^{k} \prod_{i:z_i=j} f(y_i|\theta_j) \end{aligned} \tag{5.8}$$

i.e., conditional on the information about the source of y_i, only the corresponding component is relevant.

The parameter set is augmented (w, θ, z), but the implementation of the Gibbs sampler is straightforward. Letting $n_j = \sum_{j=1}^{k} I_{(z_i=j)}$ and $n_j\,\bar{y}_j = \sum_{i:z_i=j} y_i$, for $j = 1, \ldots, k$, the complete set of full conditionals is:

- Full conditional for w:

$$p(w|\theta, z, y) = p_w(w|\delta_1 + n_1, \ldots, \delta_k + n_k) = \mathrm{Di}(w|\delta_1 + n_1, \ldots, \delta_k + n_k)$$

- Full conditional for θ: independently for $j = 1, \ldots, k$,

$$p(\theta_j\,|w, z, y) \propto \exp\left\{(r_j + n_j\bar{y}_j)\theta_j - (\lambda_j + n_j)b(\theta_j)\right\},$$

- Full conditional for z: independently for $j = 1, \ldots, k$,

$$p(z_i|w, \theta, y) = c^{-1} \sum_{j=1}^{k} w_j f(y_i|\theta_j)\, I_{(z_i=j)}, \tag{5.9}$$

  where $c = \sum_{s=1}^{k} w_s f(y_i|\theta_s)$.

For derivation of these results, see Appendix 5.A.

5.4 Mixture of normal distributions

5.4.1 Known number of components

As a special case of the previous analysis, Bayesian estimation of the mixture of normal distributions with a fixed number of components is easily established. The following analysis largely follows Diebolt and Robert (1994). The number of components k is initially assumed fixed. The next sections will generalize the model to an unknown k.

First of all, the distribution of the observable variable is now:

$$y_i|w, \mu, \sigma^2 \overset{iid}{\sim} \sum_{j=1}^{k} w_j \, \mathrm{N}(y_i|\mu_j, \sigma_j^2), \qquad i = 1, \ldots, n. \tag{5.10}$$

with $\mu = (\mu_1, \ldots, \mu_k)$ and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2)$.

It is easy to derive the moments for a mixture of normal distributions:

$$E(y_i) = \sum_{j=1}^{k} w_j \, \mu_j$$

$$E(y_i^2) = \sum_{j=1}^{k} w_j (\mu_j^2 + \sigma_j^2)$$

The likelihood of this model is:

$$L(\mu, \sigma^2, z) = \prod_{j=1}^{k} \prod_{i:z_i=j} \mathrm{N}(y_i|\mu_j, \sigma_j^2) \tag{5.11}$$

While the prior distribution for the weights still remains (5.6), priors for μ and σ^2 are:

$$\mu_j \overset{iid}{\sim} \mathrm{N}(\mu_j|\mu_0, \tau^2) \tag{5.12}$$

$$\sigma_j^2 \overset{iid}{\sim} \mathrm{Ig}(\sigma_j^2|\alpha, \beta) \tag{5.13}$$

for $j = 1, \ldots, k$, with hyperparameters μ_0, τ^2, α and β assumed known. This particular prior specification can be viewed as a particular case of the conjugate analysis for the exponential family: see Bernardo and Smith (1994) for details.

The set of full conditionals (appendix 5.B) is:

- Full conditional for w:

$$p(w|\mu, \sigma, z, y) = \mathrm{Di}(w|\delta_1 + n_1, \ldots, \delta_k + n_k) \tag{5.14}$$

- Full conditional for μ: independently for $j = 1, \ldots, k$,

$$p(\mu_j|w, \sigma, z, y) = \mathrm{N}\left(\mu_j \,\middle|\, \frac{n_j \bar{y}_j \tau^2 + \sigma_j^2 \mu_0}{n_j \tau^2 + \sigma_j^2}, \frac{\sigma_j^2 \tau^2}{n_j \tau^2 + \sigma_j^2}\right) \tag{5.15}$$

- Full conditional for σ^2: independently for $j = 1, \ldots, k$,

$$p(\sigma_j^2|w, \mu, z, y) = \mathrm{Ig}\left(\sigma_j^2 \,\middle|\, \alpha + \frac{1}{2} n_j, \beta + \frac{1}{2} \sum_{i:z_i=j} (y_i - \mu_j)^2\right) \tag{5.16}$$

- Full conditional for z: independently for $i = 1, \ldots, n$,

$$p(z_i|w, \mu, \sigma, y) = c^{-1} \sum_{j=1}^{k} \frac{w_j}{\sigma_j} \exp\left\{\frac{(y_i - \mu_j)^2}{2\sigma_j^2}\right\} I_{(z_i=j)} \tag{5.17}$$

  where $c = \sum_{s=1}^{k} \frac{w_s}{\sigma_s} \exp\left\{\frac{(y_i - \mu_s)^2}{2\sigma_s^2}\right\}$.

5.4.2 Unknown number of components

Up to now, we dealt with a fixed number of mixture components k. Nevertheless, mixture models should be suitable for situations where individual components are meaningless. Clearly, this issue can be viewed as a model selection problem. In the next sections, this problem will be treated using the reversible jump algorithm (section 4.2.2) and the method based on the estimation of the marginal likelihood (section 4.3.2). Here, we mention some former techniques.

Mengersen and Robert (1995) proposed a test based on the *Kullback-Leibler divergence* (or *entropy distance*). Suppose there are two competing sampling distributions, let's say $g(y)$ and $h(y)$, for a given sample y; the Kullback-Leibler divergence is a metric distance defined by:

$$ED[g,h] = \int \ln[g(y)/h(y)]g(y)dy$$

Suppose now we are interested in testing the presence of mixture model: we can compare a two-component mixture model to a single-component model (i. e. a normal distribution $\mathrm{N}(\mu, \sigma^2)$), which is the closest in terms of Kullback-Leibler divergence. Hence, for given μ_1, μ_2, σ_1^2 and σ_2^2, we choose μ and σ that minimizes:

$$ED[w\mathrm{N}(\cdot|\mu_1, \sigma_1^2) + (1-w)\mathrm{N}(\cdot|\mu_2, \sigma_2^2)\,,\, \mathrm{N}(\cdot|\mu, \sigma^2)] \tag{5.18}$$

It is possible to show that equation (5.18) is minimized when:

$$\begin{aligned} \mu &= w\mu_1 + (1-w)\mu_2 \\ \sigma^2 &= w\sigma_1^2 + (1-w)\sigma_2^2 + w(1-w)(\mu_1 - \mu_2)^2 \end{aligned}$$

If the distance (5.18) is less than a given bound, the parsimony principle indicates to discard a mixture (of normal distributions) model.

The *stochastic search variable selection* is a procedure proposed by George and McCulloch (1996) and developed for regression models: essentially, it puts a probability distribution on the set of all possible models such that "promising"models are given highest probability. A sample from this distribution is then obtained by Gibbs sampler.

Other approaches related to model choice can be founded in Gilks et al. (1996).

5.4.3 Reversible jump for mixture models

Previously, the choice of the number of components k and the parameter estimation (with k fixed) were treated separately. The following Bayesian analysis of mixtures of normal distributions with an unknown number of components, based on a Reversible Jump MCMC algorithm (section 4.2.2), allows us to deal simultaneously with estimation and model selection. The reference paper is Richardson and Green (1997).

First of all, since k is now a random variable, we need to specify a prior distribution. A reasonable choice is the discrete uniform distribution between 1 and k_{max}:

$$k|k_{max} \sim \mathrm{Un}(k|1, k_{max})$$

with hyperparameter k_{max} fixed.

The MCMC algorithm of section 5.4.1 is now enriched with a reversible jump type move. In practice, the parameters w, μ, σ and the allocation variable z are updated by Gibbs moves in the exact way of section 5.4.1 but, in addition, a move that updates k is introduced. The complete list of moves is summarized as:

i. Updating w
ii. Updating μ
iii. Updating σ
iv. Updating z
v. Updating k (reversible jump move)

For the moves from (i) to (iv), see the full conditionals (5.14),(5.15),(5.16) and (5.17).

Move (v) consists in splitting one mixture component into two, or combining two into one. Hence, a random choice between "split" and "combine" is made, with probabilities s_k and c_k respectively. These probabilities depend on k and they are such that $c_k = 1 - s_k$, with $c_1 = 0$ and $s_{k_{max}} = 0$.

The combine proposal begins by choosing two components (j_1, j_2) at random which are adjacent in terms of the current values of their means (i. e., $\mu_{j1} < \mu_{j2}$, with no other μ_j in the interval $[\mu_{j1}, \mu_{j2}]$). The reason of this constraint will be explained in section (5.5). If these two components are merged, then k is reduced by 1, forming a new component, say j^*. In order to propose values for w_{j^*}, μ_{j^*} and σ_{j^*}, consider the following transformation:

$$\begin{aligned} w_{j^*} &= w_{j_1} + w_{j_2} \\ w_{j^*}\mu_{j^*} &= w_{j_1}\mu_{j_1} + w_{j_2}\mu_{j_2} \\ w_{j^*}(\mu_{j^*}^2 + \sigma_{j^*}^2) &= w_{j_1}(\mu_{j_1}^2 + \sigma_{j_1}^2) + w_{j_2}(\mu_{j_2}^2 + \sigma_{j_2}^2) \end{aligned} \tag{5.19}$$

This proposal mechanism is based on a *moment matching* expedient: the zeroth, first and second moments of the new component (left side of the three lines of 5.19) are equal to the corresponding moments of the combination of the two merging components (right side). Equation (5.19) gives in a deterministic way the proposed values for the parameters.

Finally, observations y_i with $z_i = j_1$ or $z_i = j_2$ have to be reallocated by setting $z_i = j^*$.

For the split proposal, a component j^* is chosen at random and split into j_1 and j_2. The values for w_{j_1}, μ_{j_1}, σ_{j_1}, w_{j_2}, μ_{j_2} and σ_{j_2} are given by:

$$\begin{aligned}
w_{j_1} &= w_{j^*}u_1 \\
w_{j_1} &= w_{j^*}(1-u_1) \\
\mu_{j1} &= \mu_{j^*} - u_2\sigma_{j^*}\sqrt{w_{j_2}/w_{j_1}} \\
\mu_{j2} &= \mu_{j^*} + u_2\sigma_{j^*}\sqrt{w_{j_1}/w_{j_2}} \\
\sigma^2_{j1} &= u_3(1-u_2^2)\sigma^2_{j^*}w_{j^*}/w_{j1} \\
\sigma^2_{j2} &= (1-u_3)(1-u_2^2)\sigma^2_{j^*}w_{j^*}/w_{j2}
\end{aligned} \tag{5.20}$$

where u_1, u_2 and u_3 are random quantities between 0 and 1 (we will use beta distributions). The set of equations (5.20) satisfies the set (5.19): in other words, (5.19) and (5.20) define a one-to-one transformation. Once these values are proposed, it is necessary to check if the adjacency condition in terms of means is satisfied: if not, the move is rejected. The observations y_i such that $z_i = j^*$ are reallocated between j_1 and j_2 using the Gibbs move for the allocation variable (5.17).

For the split move the acceptance probability is min(1,R), where:

$$\begin{aligned}
R = \exp\Bigg\{ & (n_{j_1} + n_{j_2})\log\sigma_{j^*} - n_{j_1}\log\sigma_{j1} - n_{j_2}\log\sigma_{j2} + \frac{1}{2\sigma_{j^*}^2} \sum_{i:z_i=j^*} (y_i - \mu_{j^*})^2 \\
& - \frac{1}{2\sigma_{j2}^2} \sum_{i:z_i=j2} (y_i - \mu_{j2})^2 - \frac{1}{2\sigma_{j1}^2} \sum_{i:z_i=j1} (y_i - \mu_{j1})^2 \Bigg\} \\
& \times (k+1) w_{j1}^{\delta-1+l_1} w_{j2}^{\delta-1+l_2} w_{j^*}^{1-\delta-l_1-l_2} B(\delta, k\delta)^{-1} \\
& \times \sqrt{\frac{k}{2\pi}} \exp\left\{ -\frac{1}{2} k[(\mu_{j1} - \mu_0)^2 + (\mu_{j2} - \mu_0)^2 - (\mu_{j^*} - \mu_0)^2] \right\} \\
& \times \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\sigma_{j1}^{-2} \sigma_{j2}^{-2} \sigma_{j^*}^{2} \right)^{(\alpha+1)} \exp\{-\beta(\sigma_{j1}^{-2} + \sigma_{j2}^{-2} - \sigma_{j^*}^{-2})\} \\
& \times \frac{d_{k+1}}{b_k P_{alloc}} [\mathrm{Be}(u_1|2,2)\mathrm{Be}(u_2|2,2)\mathrm{Be}(u_3|1,1)]^{-1} \\
& \times \frac{w_{j^*} |\mu_{j1} - \mu_{j2}| \sigma_{j1}^2 \sigma_{j2}^2}{u_2(1-u_2^2) u_3 (1-u_3) \sigma_{j^*}^2}
\end{aligned} \tag{5.21}$$

where n_{j_1} and n_{j_2} are the numbers of observations proposed to be assigned to j_1 and j_2, B and Γ are the Beta and the Gamma functions respectively (section 2.5.1) and P_{alloc} is the probability that this particular allocation is made.

The first two lines of expression (5.21) form the likelihood ratio. The third, fourth and fifth line correspond to the prior ratio. The sixth line is the proposal ratio and the final line is the jacobian. The derivation of these results is given in Appendix 5.C.

For the combine move, the acceptance probability move is min(1,R^{-1}), using the same expression for R but with some differences in the substitutions.

Actually, Richardson and Green (1997) add another reversible jump move for k (birth and death move) to increase the efficiency of the algorithm.

5.4.4 Marginal likelihood for mixture models

Chib (1995) uses his method to estimate the marginal likelihood from the Gibbs output to mixture of normal distributions (5.10). Consider the Gibbs sampler described in section 5.4.1 and write the posterior ordinate as:

$$p(\mu^*, \sigma^{2*}, w^*|y) = p(\mu^*|y) \times p(\sigma^{2*}|\mu^*, y) \times p(w^*|\mu^*, \sigma^{2*}, y)$$

A straightforward application of the method (section 4.3.2) leads to the following steps:

1. Sample $\{\mu^{(i)}, \sigma^{2(i)}, w^{(i)}, z^{(i)}\}$ from the full Gibbs run and set

$$\bar{p}(\mu^*|y) = N^{-1} \sum_{i=1}^{N} \prod_{j=1}^{k} p(\mu_j^*|\sigma^{2(i)}, z^{(i)}, y)$$

   where $p(\mu_j^*|\sigma^{2(i)}, z^{(i)}, y)$ is given by equation(5.15).

2. Sample $\{\sigma^{2(i)}, w^{(i)}, z^{(i)}\}$ from the reduced Gibbs conditional to μ^* and set:

$$\bar{p}(\sigma^{2*}|\mu^*, y) = N^{-1} \sum_{i=1}^{N} \prod_{j=1}^{k} p(\sigma_j^{2*}|\mu^*, z^{(i)}, y)$$

   where $p(\sigma_j^{2*}|\mu^*, z^{(i)}, y)$ is given by equation (5.16).

3. Sample $\{w^{(i)}, z^{(i)}\}$ from the reduced Gibbs conditional to (μ^*, σ^{2*}) and set:

$$\bar{p}(w^*|\mu^*, \sigma^{2*}, y) = N^{-1} \sum_{i=1}^{N} p(w^*|z^{(i)}, y)$$

   where $p(w^*|z^{(i)}, y)$ is given by equation (5.14).

Finally, equation (4.12) gives the estimate of the marginal loglikelihood:

$$\begin{aligned}\ln\bar{f}(y) = l(w^*, \mu^*, \sigma^{2*}) + \ln p(\mu^*) + \ln p(\sigma^{2*}) + \ln p(w^*) - \\ -\ln\bar{p}(\mu^*|y) - \ln\bar{p}(\sigma^{2*}|\mu^*, y) - \ln\bar{p}(w^*|\mu^*, \sigma^{2*}, y)\end{aligned}$$

The three priors in the first line are given by equations (5.6), (5.12) and (5.13). Note that the loglikelihood $l(w, \mu^*, \sigma^{2*})$ does not depend on the allocation variable z and, from equation (5.7), it is equal to:

$$l(w, \mu, \sigma^2) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{k} w_j N(y_i|\mu_j, \sigma_j^2) \right]$$

5.5 Label switching problem

Now we shall illustrate an important issue related to a mixture model. Consider equation (5.1):

$$y_i|w, \theta \overset{iid}{\sim} \sum_{j=1}^{k} w_j f(y_i|\theta_j)$$

and suppose to indicate all the parameters by $\psi = (w, \theta)$. We know from expression (5.7) that the likelihood is:

$$\begin{aligned} L(\psi) &= \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f(y_i|\theta_j) \\ &= \prod_{i=1}^{n} [w_1 f(y_i|\theta_1) + \cdots + w_k f(y_i|\theta_k)] \end{aligned} \tag{5.22}$$

For any permutation ν of $1, \ldots, k$, define the corresponding permutation of the parameter vector ψ by:

$$\nu(\psi) = ((w_{\nu(1)}, \ldots, w_{\nu(k)}), (\theta_{\nu(1)}, \ldots, \theta_{\nu(k)})) \tag{5.23}$$

The so-called *label switching problem* derives from the fact that the likelihood (5.22) is the same for all permutations of ψ.

In a Bayesian analysis, if we have no prior information that distinguishes between the components of the mixture (that is, the joint prior distribution is the same for all permutations of ψ), then the posterior distribution will be similarly symmetric (see Fruhwirth-Schnatter, 2001, for a proof). As a result, posterior shows artificial multimodality, which poses obvious problems in terms of parameter estimations.

The usual solution of the label switching problem consists in imposing an *identifiability constraint* on the parameter space, such as $w_1 < w_2 < \cdots < w_k$ or $\theta_1 < \theta_2 < \ldots, < \theta_k$. Actually, in the analysis of the mixture model with an unknown number of components of section 5.4.3, Richardson and Green (1997) restricted the component means μ_j in their increasing numerical order. This kind of constraints can be satisfied by only one permutation of ψ and this breaks the symmetry of the prior. In practice, the full conditional of μ_j (expression 5.15) is used to generate a proposal which is accepted only if the ordering is satisfied.

Fruhwirth-Schnatter (2001) proposes an MCMC estimation of models affected by the label switching problem based on a *permutation sampler*. Each iteration of the MCMC algorithm is concluded by a permutation of the current labelling of the states: the permutation is selected in such a way that the identifiability constraint is fulfilled. As a consequence, the algorithm doesn't reject forthwith parameter values when they do not satisfy the constrain, but it jumps between the various labelling subspaces in a balanced fashion. The permutation sampler is more efficient with respect to the simple rejection and we adopted it in our analysis.

Appendix 5.A

Full conditional for w

Using equations (5.2) and (5.6):

$$\begin{aligned} p(w|\theta, z, y) &\propto p(w, \theta, z, y) \\ &\propto p(z|w)p(w) \\ &\propto \prod_{i=1}^{n} \sum_{j=1}^{k} w_j I_{(z_i=j)} \prod_{j=1}^{k} w_j^{\delta_j - 1} \\ &\propto \prod_{j=1}^{k} w_j^{n_j} \prod_{j=1}^{k} w_j^{\delta_j - 1} \\ &= \mathrm{Di}(w|\delta_1 + n_1, \ldots, \delta_k + n_k) \end{aligned}$$

where $n_j = \sum_{j=1}^{k} I_{(z_i=j)}$, for $j = 1, \ldots, k$.

Full conditional for θ

Using equation (5.8):

$$\begin{aligned} p(\theta|w, z, y) &\propto L(\theta, z)p(\theta) \\ &\propto \prod_{j=1}^{k} \prod_{i:z_i=j} f(y_i|\theta_j) \prod_{j=1}^{k} p(\theta_j) \end{aligned}$$

Hence, using (5.4) and (5.5):

$$\begin{aligned} p(\theta_j|w, z, y) &\propto \left[\prod_{i:z_i=j} a(y_i) \exp\{y_i \, \theta_j - b(\theta_j)\} \right] \exp\{r_j \, \theta_j - \lambda_j \, b(\theta_j)\} \\ &\propto \exp\{(r_j + n_j \bar{y}_j)\theta_j - (\lambda_j + n_j) b(\theta_j)\} \end{aligned}$$

where $\bar{y}_j = \frac{1}{n_j} \sum_{i:z_i=j} y_i$, for $j = 1, \ldots, k$.

Full conditional for z

Using equations (5.8) and (5.2):

$$\begin{aligned} p(z|w, \theta, y) &\propto L(\theta, z)p(z|w) \\ &\propto \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f(y_i|\theta_j) I_{(z_i=j)} \end{aligned}$$

Hence:

$$p(z_i|w,\theta,y) \propto \sum_{j=1}^{k} w_j f(y_i|\theta_j) I_{(z_i=j)}$$

Calculating the normalization constant we obtain the (5.9).

Appendix 5.B

Full conditional for w

The derivation of the full conditional for w is the same of the previous appendix.

Full conditional for μ

Using equations (5.11) and (5.12):

$$\begin{aligned} p(\mu|w,\sigma^2,z,y) &\propto L(\mu,\sigma^2,z)p(\mu|\mu_0,\tau) \\ &= \prod_{j=1}^{k} \prod_{t:z_t=j} \mathrm{N}(y_i|\mu_j,\sigma_j^2) \prod_{j=1}^{k} \mathrm{N}(\mu_j|\mu_0,\tau^2) \end{aligned}$$

Hence:

$$p(\mu_j|w,\sigma^2,z,y) \propto \left[\prod_{t:z_t=j} \mathrm{N}(y_t|\mu_j,\sigma_j^2)\right] \mathrm{N}(\mu_j|\mu_0,\tau^2)$$

for $j = 1,\ldots,k$. Substituting the normal density function:

$$\begin{aligned} p(\mu_j|w,\sigma^2,z,y) &\propto \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i:z_i=j}(y_i-\mu_j)^2 - \frac{1}{2\tau^2}(\mu_j-\mu_0)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i:z_i=j}(y_i-\bar{y}_j+\bar{y}_j-\mu_j)^2 - \frac{1}{2\tau^2}(\mu_j-\mu_0)^2\right\} \\ &= \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i:z_i=j}[(y_i-\bar{y}_j)^2+n_j(\bar{y}_j-\mu_j)^2] - \frac{1}{2\tau^2}(\mu_j-\mu_0)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma_j^2}n_j(\bar{y}_j-\mu_j)^2 - \frac{1}{2\tau^2}(\mu_j-\mu_0)^2\right\} \\ &\propto \exp\left\{-\left[\left(\frac{n_j}{2\sigma_j^2}+\frac{1}{2\tau^2}\right)\mu_j^2 - \left(\frac{n_j\bar{y}_j}{\sigma_j^2}+\frac{\mu_0}{\tau^2}\right)\mu_j\right]\right\}, \end{aligned}$$

where $\bar{y}_j = \frac{1}{n_j}\sum_{i:z_i=j} y_i$. Now, if we let:

$$A = \frac{n_j}{2\sigma_j^2} + \frac{1}{2\tau^2} = \frac{1}{2}\frac{n_j\tau^2 + \sigma_j^2}{\sigma_j^2\tau^2} \tag{5.24}$$

$$D = \frac{1}{2}\left(\frac{n_j\bar{x}_j}{\sigma_j^2} + \frac{\mu_0}{\tau^2}\right) = \frac{1}{2}\frac{n_j\bar{x}_j\tau^2 + \mu_0\sigma_j^2}{\sigma_j^2\tau^2} \tag{5.25}$$

then:

$$\begin{aligned} p(\mu_j|w,\sigma^2,z,y) &\propto \exp\{-\left[A\mu_j^2 - 2D\mu_j\right]\} \\ &\propto \exp\left\{-A\left[\mu_j^2 - 2\frac{D}{A}\mu_j\right]\right\} \\ &\propto \exp\left\{-A\left[\left(\mu_j - \frac{D}{A}\right)^2 - \frac{D^2}{A^2}\right]\right\} \\ &\propto \exp\left\{-A\left(\mu_j - \frac{D}{A}\right)^2\right\} \end{aligned}$$

Finally, substituting the (5.24) and (5.25), we obtain:

$$\begin{aligned} p(\mu_j|w,\sigma^2,z,y) &\propto \exp\left\{\frac{1}{2}\frac{n_j\tau^2 + \sigma_j^2}{\sigma_j^2\tau^2}\left(\mu_j - \frac{n_j\,\bar{x}_j\,\tau^2 + \sigma_j^2\mu_0}{n_j\,\tau^2 + \sigma_j^2}\right)^2\right\} \\ &= \mathrm{N}\left(\mu_j \,\middle|\, \frac{n_j\,\bar{y}_j\,B\,\tau^2 + \sigma_j^2\mu_0}{n_j\,B^2\tau^2 + \sigma_j^2}, \frac{\sigma_j^2\tau^2}{n_j\,B^2\tau^2 + \sigma_j^2}\right). \end{aligned}$$

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

Full conditional for σ^2

Using equations (5.11) and (5.13):

$$\begin{aligned} p(\sigma^2|w,\mu,z,y) &\propto L(\mu,\sigma^2,z)p(\sigma^2) \\ &= \prod_{j=1}^{k}\prod_{i:z_i=j}\mathrm{N}(y_i|\mu_j,\sigma_j^2)\prod_{j=1}^{k}\mathrm{Ig}(\sigma_j^2|\alpha,\beta) \end{aligned}$$

Hence, for $j = 1, \dots, k$, the full conditional for σ_j^2 is:

$$\begin{aligned}
p(\sigma_j^2|w,\mu,z,y) &\propto \left[\prod_{i:z_i=j} \mathrm{N}(y_i|\mu_j,\sigma_j^2)\right] \mathrm{Ig}(\sigma_j^2|\alpha,\beta) \\
&\propto \sigma_j^{-n_j}\exp\left\{-\frac{1}{2}\sum_{i:z_i=j}\frac{(y_i-\mu_j)^2}{\sigma_j^2}\right\}\sigma_j^{-2(\alpha+1)}\exp\{-\beta/\sigma_j^2\} \\
&= \sigma_j^{-2(n_j/2+\alpha+1)}\exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i:z_i=j}(y_i-\mu_j)^2-\frac{\beta}{\sigma_j^2}\right\} \\
&= \mathrm{Ig}\left(\sigma_j^2\middle|\alpha+\frac{1}{2}n_j,\beta+\frac{1}{2}\sum_{i:z_i=j}(y_i-\mu_j)^2\right)
\end{aligned}$$

Full conditional for z

Using equations (5.11) and (5.2):

$$\begin{aligned}
p(z|w,\mu,\sigma^2,y) &\propto L(\mu,\sigma^2,z)p(z|w) \\
&= \prod_{i=1}^{n}\sum_{j=1}^{k} w_j\mathrm{N}(y_i|\mu_j,\sigma_j^2)I_{(z_i=j)}
\end{aligned}$$

Hence:

$$\begin{aligned}
p(z_i|w,\mu,\sigma^2,y) &\propto \sum_{j=1}^{k} w_j\mathrm{N}(y_i|\mu_j,\sigma_j^2)I_{(z_i=j)} \\
&\propto \sum_{j=1}^{k}\left[\frac{w_j}{\sigma_j}\exp\left\{\frac{(y_i-\mu_j)^2}{2\sigma_j^2}\right\}I_{(z_i=j)}\right]
\end{aligned}$$

calculating the normalization constant we obtain the (5.17).

Appendix 5.C

Expression (5.21) can be represented as:

$$R = (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio}) \times (\text{jacobian})$$

The component j^* is split into j_1 and j_2. We shall denote the k-dimension vectors of the current values of the parameters by μ, σ^2 and w. The proposal values, constructed in section 5.4.3, are μ', $\sigma^{2'}$ and w': remember they are of dimension $k+1$ and they differ from the current values only through the elements of positions j_1 and j_2. z and z' indicate the current and the proposal allocation variables.

- The likelihood ratio is:

$$
\begin{aligned}
\frac{L(\mu', \sigma^{2'}, z', k+1)}{L(\mu, \sigma^2, z, k)} &= \frac{\prod_{j=1}^{k+1} \prod_{i:z'_i=j} \mathrm{N}(y_i|\mu'_j, \sigma_j^{2'})}{\prod_{j=1}^{k} \prod_{i:z_i=j} \mathrm{N}(y_i|\mu_j, \sigma_j^{2})} \\
&= \frac{\prod_{i:z'_i=j_1} \mathrm{N}(y_i|\mu_{j_1}, \sigma_{j_1}^2) \prod_{i:z'_i=j_2} \mathrm{N}(y_i|\mu_{j_2}, \sigma_{j_2}^2)}{\prod_{i:z_i=j^*} \mathrm{N}(y_i|\mu_{j^*}, \sigma_{j^*}^2)} \\
&= \exp\Big\{ (n_{j_1}+n_{j_2})\mathrm{ln}\sigma_{j^*} - n_{j_1}\mathrm{ln}\sigma_{j1} - n_{j_2}\mathrm{ln}\sigma_{j2} \\
&+ \frac{1}{2\sigma_{j^*}^2} \sum_{i:z_i=j^*} (y_i-\mu_{j^*})^2 - \frac{1}{2\sigma_{j2}^2} \sum_{i:z_i=j2} (y_i-\mu_{j2})^2 \\
&- \frac{1}{2\sigma_{j1}^2} \sum_{i:z_i=j1} (y_i-\mu_{j1})^2 \Big\}
\end{aligned}
$$

where n_{j_1} and n_{j_2} are the numbers of observations proposed to be assigned to j_1 and j_2.

- The prior ratio is:

$$
\begin{aligned}
\frac{p(\mu', \sigma^{2'}, w', z', k+1)}{p(\mu, \sigma^2, w, z, k)} &= \frac{p(\mu'|+1')}{p(\mu|k)} \times \frac{p(\sigma^{2'}|k+1)}{p(\sigma^2|k)} \\
\times \frac{p(w'|k+1)}{p(w|k)} &\times \frac{p(z'|w', k+1)}{p(z|w,k)} \times \frac{p(k+1)}{p(k)}
\end{aligned}
$$

where:

$$\frac{p(\mu'|k+1)}{p(\mu|k)} = \frac{\prod_{j=1}^{k+1} \mathrm{N}(\mu'_j|\mu_0,\tau^2)}{\prod_{j=1}^{k} \mathrm{N}(\mu_j|\mu_0,\tau^2)} = \frac{\mathrm{N}(\mu_{j_1}|\mu_0,\tau^2)\mathrm{N}(\mu_{j_2}|\mu_0,\tau^2)}{\mathrm{N}(\mu_{j^*}|\mu_0,\tau^2)}$$
$$= \sqrt{\frac{k}{2\pi}}\exp\left\{-\frac{1}{2}k[(\mu_{j1}-\mu_0)^2+(\mu_{j2}-\mu_0)^2-(\mu_{j^*}-\mu_0)^2]\right\}$$

$$\frac{p(\sigma^{2'}|k+1)}{p(\sigma^2|k)} = \frac{\prod_{j=1}^{k+1} \mathrm{Ig}(\sigma_j^{2'}|\alpha,\beta)}{\prod_{j=1}^{k} \mathrm{Ig}(\sigma_j^2|\alpha,\beta)} = \frac{\mathrm{Ig}(\sigma_{j1}^2|\alpha,\beta)\mathrm{Ig}(\sigma_{j2}^2|\alpha,\beta)}{\mathrm{Ig}(\sigma_{j^*}^2|\alpha,\beta)} =$$
$$= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\sigma_{j1}^{-2}\sigma_{j2}^{-2}\sigma_{j^*}^{2}\right)^{(\alpha+1)}\exp\{-\beta(\sigma_{j1}^{-2}+\sigma_{j2}^{-2}-\sigma_{j^*}^{-2})\}$$

$$\frac{p(w'|k+1)}{p(w|k)} = \frac{\mathrm{Di}(w'|\delta_1,\ldots,\delta_{k+1})}{\mathrm{Di}(w|\delta_1,\ldots,\delta_k)} = \frac{w_{j1}^{\delta-1}w_{j1}^{\delta-1}\frac{\Gamma(\delta)^{k+1}}{\Gamma(k\delta+\delta)}}{w_{j^*}^{\delta-1}\frac{\Gamma(\delta)^k}{\Gamma(k\delta}} = \frac{w_{j1}^{\delta-1}w_{j1}^{\delta-1}}{w_{j^*}^{\delta-1}B(\delta,k\delta)}$$

$$\frac{p(z'|w',k+1)}{p(z|w,k)} = \frac{\prod_{i=1}^{n}\sum_{j=1}^{k+1} w'_j I_{(z'_i=j)}}{\prod_{i=1}^{n}\sum_{j=1}^{k} w_j I_{(z_i=j)}} = \frac{w_{j_1}^{n_{j_1}}w_{j_2}^{n_{j_2}}}{w_{j^*}^{n_{j_1}+n_{j_2}}}$$

Since we chose a uniform prior for k, the last factor $p(k+1)/p(k)$ is 1. Moreover, the $(k+1)$-factor in the third line of expression (5.21) comes from the order statistics densities for the means.

- The proposal ratio is formed by the probabilities of birth and death of a component, the probability of the allocation P_{alloc} and the densities of the variables u_1,u_2 and u_3:

$$\frac{d_{k+1}}{b_k P_{alloc}}\ [\mathrm{Be}(u_1|2,2)\mathrm{Be}(u_2|2,2)\mathrm{Be}(u_3|1,1)]^{-1}$$

- The transformation defined in (5.20) from $(w_{j^*},\mu_{j^*},\sigma_{j^*}^2,u_1,u_2,u_3)$ to

$(w_{j_1}, \mu_{j_1}, \sigma^2_{j_1}, w_{j_2}, \mu_{j_2}, \sigma^2_{j_2})$ leads to the following (6×6) Jacobian matrix:

$$J = \begin{bmatrix} u_1 & 0 & 0 & w_{j^*} & 0 & 0 \\ 0 & 1 & \frac{1}{2}u_2\sqrt{\frac{w_{j_2}}{w_{j_1}}\sigma^{-1}_{j^*}} & 0 & -\sqrt{\frac{w_{j_2}}{w_{j_1}}\sigma^2_{j^*}} & 0 \\ \frac{u_3(1-u_2^2)\sigma^2_{j^*}}{w_{j_1}} & 0 & u_3(1-u_2^2)\frac{w_{j^*}}{w_{j_1}} & 0 & -2u_2u_3\sigma^2_{j^*}\frac{w_{j^*}}{w_{j_1}} & (1-u_2^2)\sigma^2_{j^*}\frac{w_{j^*}}{w_{j_1}} \\ (1-u_1) & 0 & 0 & -w_{j^*} & 0 & 0 \\ 0 & 1 & 1\frac{1}{2}u_2\sqrt{\frac{w_{j_1}}{w_{j_2}}\sigma^{-1}_{j^*}} & 0 & \sqrt{\frac{w_{j_2}}{w_{j_1}}\sigma^2_{j^*}} & 0 \\ \frac{(1-u_3)(1-u_2^2)\sigma^2_{j^*}}{w_{j_2}} & 0 & (1-u_3)(1-u_2^2)\frac{w_{j^*}}{w_{j_2}} & 0 & -2u_2(1-u_3)\sigma^2_{j^*}\frac{w_{j^*}}{w_{j_2}} & -(1-u_2^2)\sigma^2_{j^*}\frac{w_{j^*}}{w_{j_2}} \end{bmatrix}$$

The final line of expression (5.21) is the absolute value of the determinant of this matrix:

$$|\det(J)| = \frac{w_{j^*}|\mu_{j1} - \mu_{j2}|\sigma^2_{j1}\sigma^2_{j2}}{u_2(1-u_2^2)u_3(1-u_3)\sigma^2_{j^*}}$$

Chapter 6

Mixture of autoregressive components

6.1 Introduction

In the mixture of normal distributions of section 5.4, the observable variables are supposed to be conditionally independent. In this chapter, we shall present a generalization of that model, in which the variables depends on their past values through an autoregressive structure. More precisely, we shall deal with a mixture of autoregressive components.

In a *linear* time series context, marginal and conditional distributions are usually unimodal and symmetric because of the assumption of Gaussian innovation terms. Conversely, the mixture of autoregressive components represents a *non-linear* time series tool: for instance, it is possible to take into account changes in conditional distributions, which can be bimodal or multimodal.

Wong and Li (2000) propose an analysis of this class of models based on the maximization of the likelihood function. The model selection problem is solved by the *AIC* (Akaike, 1973) and *BIC* (Schwarz, 1978) criterions. Nevertheless they conclude that these methods are not satisfactory. Moreover, their classical approach does not take into account model uncertainty. Influence of model uncertainty on financial models has been investigated by some recent papers (Barberis, 2000, MacKinley and Pastor, 2000, Pastor, 2000, Pastor and Stambaugh, 2000, Cremers, 2002).

We propose a Bayesian analysis, which allows to deal with both parameter estimation and model selection.

Our analysis also takes into account the stationarity conditions for each autoregressive component through a suitable prior specification.

In the next chapter, we shall apply the model to real financial data.

6.2 Definition of mixture of autoregressive components

An autoregressive (AR) process $\{y_t\}$ of finite order ρ and mean μ is usually defined as:

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \cdots + \phi_\rho(y_{t-\rho} - \mu) + \epsilon_t \tag{6.1}$$

where ϵ_t are independent and identically normally distributed random variables with mean 0 and variance σ^2.

It is well-known that a necessary and sufficient condition for the stationarity of the solution of (6.1) is that $\phi = (\phi_1, \ldots, \phi_\rho)$ belongs to:

$$\Phi = \{\phi \in \mathbf{R}^\rho | \phi(v) \neq 0, v \in \mathbf{C}, |v| \leq 1\} \tag{6.2}$$

where $\phi(v)$ is the usual characteristic polynomial:

$$\phi(v) = 1 - \phi_1 v - \cdots - \phi_\rho v^\rho$$

(see e.g., Anderson, 1970).

If we denote:

$$\nu_t = \mu + \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \cdots + \phi_\rho(y_{t-\rho} - \mu)$$

the conditional distribution of y_t can be written as:

$$y_t|\mu, \phi, \sigma^2, \rho, x_{t-1} \sim N(y_t|\nu_t, \sigma^2)$$

where $x_{t-1} = (y_1, \ldots, y_{t-1})$.

Consider now k different AR processes: each of them is characterized by a specific order ρ_j, a mean μ_j, a set of autoregressive parameters $\phi_j = (\phi_{1,j}, \ldots, \phi_{\rho_j,j})$ and a variance of the error term σ_j^2, for $j = 1, \ldots, k$. Suppose to adopt the concise notation $\rho = (\rho_1, \ldots, \rho_k)$, $\mu = (\mu_1, \ldots, \mu_k)$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2)$, $\phi = (\phi_1, \ldots, \phi_k)$.

The *mixture of autoregressive components* can be defined by:

$$y_t|\psi, \rho, k, x_{t-1} \sim \sum_{j=1}^{k} w_j \, \mathrm{N}(y_t|\nu_{j,t}, \sigma_j^2) \tag{6.3}$$

where $\psi = (w, \mu, \sigma^2, \phi)$, $w = (w_1, \ldots, w_k)$ and:

$$\nu_{j,t} = \mu_j + \phi_{1,j}(y_{t-1} - \mu_j) + \phi_{2,j}(y_{t-2} - \mu_j) + \cdots + \phi_{\rho_j,j}(y_{t-\rho_j} - \mu_j) \tag{6.4}$$

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

The mixing weights w_j satisfy the usual constraints, i. e.

$$w_j > 0, \quad j = 1, \ldots, k; \qquad w_1 + \cdots + w_k = 1$$

The mixture of autoregressive components (6.3) can be viewed as a generalization of two different model. First of all, as we have already noticed, it generalizes the mixture of normal distributions (5.1) because the observable variable depends on the past through the quantities $\nu_{j,t}$. Second, if the number k of the mixture components is equal to 1, the model reduces to a simple autoregressive model with Gaussian error term.

It is easy to calculate the conditional moments; the conditional expectation of y_t is:

$$E(y_t|\psi, \rho, k, x_{t-1}) = \sum_{j=1}^{k} w_j \nu_{j,t} \tag{6.5}$$

while the conditional variance is given by:

$$VAR(y_t|\psi, \rho, k, x_{t-1}) = \sum_{j=1}^{k} w_j \sigma_j^2 + \sum_{j=1}^{k} w_j \nu_{j,t}^2 - \left(\sum_{j=1}^{k} w_j \nu_{j,t} \right)^2$$

6.3 Model structure and priors

We set the following prior distributions for the parameters w, μ and σ^2:

$$\begin{aligned} w|k &\sim \mathrm{Di}(w|\delta, \delta, \ldots, \delta) \\ \mu_j &\overset{iid}{\sim} \mathrm{N}(\mu_j|\mu_0, \tau^2), \qquad j = 1, \ldots, k \\ \sigma_j^2 &\overset{iid}{\sim} \mathrm{Ig}(\sigma^2|\alpha, \beta), \qquad j = 1, \ldots, k \end{aligned} \tag{6.6}$$

with δ, μ_0, τ^2, α and β assumed known.

δ is set equal to 1, as in Richardson and Green (1997).

It seems natural to take the prior for μ_j to be rather flat over an interval of variation of the observed data. Let $R = \max(x_T) - \min(x_T)$; following Richardson and Green (1997), we choose $\mu_0 = \min(x_T) + R/2$ and $\tau = cR$ (the assignment of c will be discussed later on).

The knowledge of the range of the data could be useful in setting the hyperparameters of σ^2. In particular, β will be a small multiple of $1/R^2$. It is also possible to consider an additional hierarchical level for the variances (for instance, β could have a gamma distribution, as suggested by Richardson and Green, 1997).

About the coefficients ϕ_{ρ_j}, we are able to consider a prior distribution whose domain is the stationarity region. For this point, we refer to the section 6.4.

Finally, k (the number of the AR components) and $\rho = (\rho_1, \ldots, \rho_k)$ (the orders of the AR components) will be considered as stochastic quantities with the following discrete uniform priors:

$$\rho_j | \rho_{max} \overset{iid}{\sim} \text{Un}(\rho_j | 0, \rho_{max}), \qquad j = 1, \ldots, k \tag{6.7}$$

$$k | k_{max} \sim \text{Un}(k | 1, k_{max}) \tag{6.8}$$

for fixed hyperparameters ρ_{max} and k_{max}.

With the exception of the prior on the autoregressive coefficients, which we choose to take into account the stationarity regions, the choice of the prior setting is motivated by two kind of considerations. First of all, they give advantages of conjugacy (specifically, in terms of construction of the full conditionals), even though it is not actually needed when using MCMC. In addition, we wanted to consider a set-up without strong prior information on the mixture parameters. Unfortunately, fully non-informative priors don't lead to proper posterior distributions in a mixture context (independent improper priors cannot be used; see for instance Diebolt and Robert, 1994). Of course, there are situations in which subjective priors are preferable, and our prior structure could be modified accordingly.

Suppose now the observable sample consists of T variables $x_T = (y_1, \ldots, y_T)$. We shall consider the allocation variable (section 5.2) $z = (z_1, \ldots, z_T)$, where

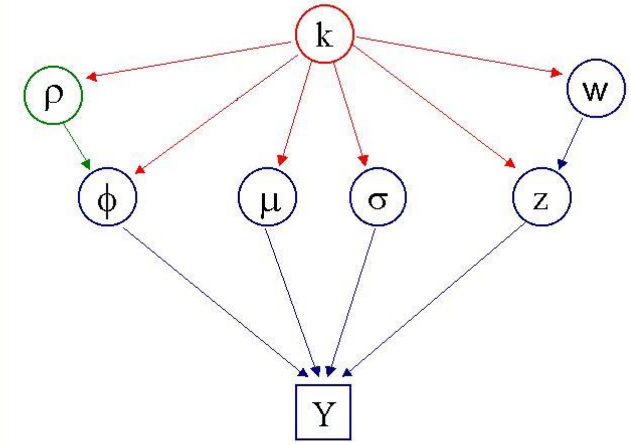

Figure 6.1: DAG of the mixture of autoregressive models.

$z_t = j$ if the tth observation comes from the jth component, for $t = 1, \ldots, T$ and $j = 1, \ldots, k$.

Assuming that there are ρ_{max} fixed observations before x_T, the (conditional) likelihood (Box et al., 1994) is given by:

$$L(\theta, z) = \prod_{j=1}^{k} \prod_{t:z_t=j} \mathrm{N}(y_t | \nu_{j,t}, \sigma_j^2) \tag{6.9}$$

for $t = 1, \ldots, T$, where $\theta = (\mu, \sigma^2, \phi)$.

The hierarchical structure of the model is displayed in figure 6.1, where we use a so called *directed acyclic graph* (DAG): observed quantities are in squares and circles include the unknowns; the directed arrows explain the dependence between variables.

6.4 Stationarity

The stationarity conditions on a model constrain the parameters to lie in regions. Apart from the cases of small orders, it is well-known that the form of these regions for an autoregressive model is very complex. Thus, a Bayesian approach is quite difficult because of the need of choosing a prior structure compatible with the stationarity regions.

Let Φ_j be the stationarity region for the jth autoregressive model. In general Φ_j is very complicated. Luckily, Barndorff-Nielsen and Schou (1973) proposed a reparametrization in terms of partial autocorrelations that simplifies the analysis.

First of all, we can define the *partial autocorrelation* between y_t and y_{t-h} as the correlation between y_{t-h} and y_t minus that part explained linearly by the intervening lags. A possible formalization of the partial autocorrelations is based on the coefficients in the linear projection of y_t on $(y_{t-1}, y_{t-2}, \dots, y_{t-h})$ (see e.g., Greene, 2000).

Let $\pi_{h,j}$ be the partial autocorrelation coefficient at lag h for the jth model. It is possible to show that $-1 < \pi_{h,j} < 1$, for $h = 1, \dots, \rho_j$, and $\pi_{h,j} = 0$, for $h > \rho_j$. In addition, $\pi_{1,j}$ equals the first autocorrelation coefficient.

For a given component j, suppose now to construct an instrumental variable φ in the following recursive way:

$$\begin{aligned}
\varphi_{1,1} &= \pi_{1,j} \\
\varphi_{h,m} &= \varphi_{h-1,m} - \varphi_{h,h}\varphi_{h-1,h-m}, && m = 1, \dots, h-1 \\
\varphi_{h,h} &= \pi_{h,j}, && h = 2, \dots, \rho_j
\end{aligned}$$

Finally, Barndorff-Nielsen and Schou (1973) proved that:

$$\phi_{hj} = \varphi_{\rho_j,h}, \quad h = 1, \dots, \rho_j$$

Hence, we established a very useful one-to-one transformation between $\phi_j = (\phi_{1,j}, \dots, \phi_{\rho_j,j})$ and $\pi_j = (\pi_{1,j}, \dots, \pi_{\rho_j,j})$.

For example, for $\rho_j = 1, 2, 3$ the mapping is given by:

- $\rho_j = 1:$

$$\phi_{1,j} = \pi_{1,j}$$

- $\rho_j = 2$:

$$\begin{aligned}
\phi_{1,j} &= \pi_{1,j}(1 - \pi_{2,j}) \\
\phi_{2,j} &= \pi_{2,j}
\end{aligned}$$

- $\rho_j = 3$:

$$\begin{aligned}\phi_{1,j} &= \pi_{1,j} - \pi_{1,j}\pi_{2,j} - \pi_{2,j}\pi_{3,j}\\ \phi_{2,j} &= \pi_{2,j} - \pi_{1,j}\pi_{3,j} - \pi_{1,j}\pi_{2,j}\pi_{3,j}\\ \phi_{3,j} &= \pi_{3,j}\end{aligned}$$

What is important to note here is that:

$$\phi_j \in \Phi_j \Longleftrightarrow |\pi_{h,j}| < 1, \quad h = 1, \ldots, \rho_j.$$

That is, the stationarity region for the jth component in terms of π_j is simply the hypercube $(-1, 1)^{\rho_j}$.

Furthermore, the prior specification is made in a straightforward way because of the following result (Jones, 1987):

$$\phi_j \sim \text{Uniform on } \Phi_j \Longleftrightarrow \pi_{i,j} \overset{ind}{\sim} \text{Be}_{(-1,+1)}\left(\pi_{i,j} \middle| \left[\frac{i+1}{2}\right], \left[\frac{i}{2}\right] + 1\right) \qquad (6.10)$$

for $i = 1, \ldots, \rho_j$ and $j = 1, \ldots, k$, where $\text{Be}_{(-1,+1)}(.)$ denotes a generalized beta distribution (section 2.5.1) defined on $(-1, +1)$ and where $[x]$ means "integer part of x". In other words, we can put an uniform prior distribution for the original parameters on the complicate stationarity region Φ_j, simply choosing a generalized beta prior for the $\pi_{i,j}$'s on $(-1, +1)$.

6.5 Parameter estimation

Suppose for the moment that k and ρ are known. We implemented an MCMC strategy based on a componentwise algorithm (section 3.4.2). The set of moves can be summarized by the following list:

i. Updating the weights w
ii. Updating the means μ
iii. Updating the partial autocorrelations π
iv. Updating the variances σ^2
v. Updating the allocation variable z

Each of these moves will be explained below; for details see Appendix 6.A.

- Move (i) is a Gibbs sampler move. The full conditional for w is the same as equation (5.14):

$$p(w|\mu, \sigma^2, \phi, z, x_T) = \mathrm{Di}(w|\delta_1 + n_1, \dots, \delta_k + n_k) \tag{6.11}$$

  where $n_j = \sum_{j=1}^{k} I_{(z_i=j)}$

- Move (ii) is also a Gibbs type move, with full conditional for μ_j given by:

$$p(\mu_j|w, \sigma^2, \phi, z, x_T) = \mathrm{N}\left(\mu_j \,\middle|\, \frac{n_j\, \bar{v}_j\, B\, \tau^2 + \sigma_j^2 \mu_0}{n_j\, B^2 \tau^2 + \sigma_j^2}, \frac{\sigma_j^2 \tau^2}{n_j\, B^2 \tau^2 + \sigma_j^2}\right) \tag{6.12}$$

  for $j = 1, \dots, k$, where $\bar{v}_j = \frac{1}{n_j} \sum_{t:z_t=j} v_{tj}$, with $v_{tj} = y_t - \phi_{j,1} y_{t-1} - \cdots - \phi_{j,\rho_j} y_{t-\rho_j}$, and where $B = 1 - \phi_{j,1} - \cdots - \phi_{j,\rho_j}$.

- Move (iii) updates π_j, for $j = 1, \dots, k$, by a Metropolis-Hastings mechanism. A candidate $\pi_{j,i}^*$ is generated by a normal density truncated in $(-1, +1)$ (section 2.5.5) and centered in the current state of the chain $\pi_{j,i}$:

$$q(\pi_{j,i}, \pi_{j,i}^*) = \mathrm{N}_{(-1,+1)}(\pi_{i,j}^* | \pi_{j,i}, \sigma_q^2) \tag{6.13}$$

  for $i = 1, \dots, \rho_j$ and for $j = 1, \dots, k$. The variance σ_q^2 is chosen in order to obtain a satisfactory acceptance rate.

  Let π_j^* be the proposal vector for the partial autocorrelations: using π_j^*, the corresponding parameters $\phi_j^* = (\phi_{j,1}^*, \dots, \phi_{j,\rho_j}^*)$ are derived trough the transformation of section (6.4).

  The acceptance probability is $\min(1, R)$, where R is given by:

$$\begin{aligned} R = &\exp\left\{ -\frac{1}{2\sigma_j^2} \sum_{t:z_t=j} \left[(y_t - \nu_{j,t}^*)^2 - (y_t - \nu_{j,t})^2 \right] \right\} \\ &\times \prod_{i=2}^{\rho_j} \frac{(\pi_{j,i}^* + 1)^{[(i-1)/2]} (1 - \pi_{j,i}^*)^{[i/2]}}{(\pi_{j,i} + 1)^{[(i-1)/2]} (1 - \pi_{j,i})^{[i/2]}} \\ &\times \prod_{i=1}^{\rho_j} \frac{F_N(1|\pi_{j,i}, \sigma_q^2) - F_N(-1|\pi_{j,i}, \sigma_q^2)}{F_N(1|\pi_{j,i}^*, \sigma_q^2) - F_N(-1|\pi_{j,i}^*, \sigma_q^2)} \end{aligned} \tag{6.14}$$

  where $\nu_{j,t}^* = \mu_j + \phi_{1,j}^*(y_{t-1} - \mu_j) + \cdots + \phi_{\rho_j,j}^*(y_{t-\rho_j} - \mu_j)$ and where F_N is the cumulative distribution function of the normal distribution. Note that the first line of the (6.14) corresponds to the likelihood ratio, the second one to the prior ratio and the third one to the proposal ratio.

- The Gibbs move (iv) is similar to the (5.16). The full conditional is:

$$p(\sigma_j^2|\beta, w, \mu, \phi, z, x_T) = \text{Ig}\left(\sigma_j^2 \middle| \, \alpha + \frac{1}{2}n_j, \beta + \frac{1}{2}\sum_{t:z_t=j}(y_t - \nu_{j,t})^2\right) \tag{6.15}$$

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

- Move (v) also is a Gibbs move with full conditional:

$$p(z_t = j|w, \mu, \sigma^2, \phi, x_T) \propto \frac{w_j}{\sigma_j} \; \exp\left\{-\frac{1}{2}\frac{(y_t - \nu_{j,t})^2}{\sigma_j^2}\right\}$$

  for $t = 1, \ldots, T$.

6.6 Model determination

We suppose that a single possible model is jointly specified by the number of the AR components k and the orders of the AR components ρ. As a consequence, Bayesian model determination is based on the evaluation of the posterior distributions $p(\rho|k, x_T)$ and $p(k|x_T)$. The procedure we implemented consists of two parts:

i. $p(\rho|k, x_T)$ is obtained by adding a reversible jump type move to the previously illustrated MCMC strategy which updates the orders ρ (section 6.6.1).

ii. The output from the step i is used to derive the marginal likelihood $f(x_T|k)$ and consequently the marginal posterior $p(k|x_T)$ by the methods described in sections 4.3.3 and 4.3.2 (section 6.6.2).

Once estimates of $p(\rho|k, x_T)$ and $p(k|x_T)$ are available, predictions of the observable variables, obtained by computing the predictive densities (section 6.7.1), can take into account contributions of every possible model (*Bayesian model averaging*). Nevertheless, one could be interested in selecting a single model, say (ρ^*, k^*). To do this, a possible strategy consists in choosing k^* as the value of k with the highest $p(k|x_t)$, and then selecting ρ^* as the value of ρ with the highest $p(\rho|k^*, x_t)$.

An alternative criterion could be proposed; specifically we could choose the values (ρ^*, k^*) with the highest *joint* posterior probability:

$$p(\rho, k|x_T) = p(\rho|k, x_T)p(k|x_T)$$

The two strategies are different and in principle they could not lead to the same results. Nevertheless, it is important to note that the second criterion penalizes models with an high number of components: as a matter of fact, the joint prior distribution is (see expressions 6.7 and 6.8):

$$p(\rho, k) = p(\rho|k)p(k) \propto \left(\frac{1}{\rho_{max} + 1}\right)^k$$

6.6.1 Order of the autoregressive components

The set of moves described in section 6.5 is augmented by a sixth move:

vi. Updating ρ

This reversible jump move starts by selecting a component, say j^*, randomly chosen in $\{1, \ldots, k\}$. The order of this component ρ_{j^*} increases by one with probability $b(\rho_{j^*})$ and decreases by one with probability $d(\rho_{j^*})$, where $b(\rho_j) = 1 - d(\rho_j)$, for $j = 1, \ldots, k$, $d(1) = 0$ and $b(\rho_{max}) = 0$. Formally, the proposal order $\rho^*_{j^*}$ is constructed as follow:

$$\rho^*_{j^*} = \begin{cases} \rho_{j^*} - 1 \ , & \text{with prob. } d(\rho_j) \\ \rho_{j^*} + 1 \ , & \text{with prob. } b(\rho_j) \end{cases}$$

It is now necessary to change the partial autocorrelation coefficients. Following Barbieri and O'Hagan (1997), if the order is decreased, the last partial autocorrelation is simply discarded. Otherwise, we need a new parameter $\pi^*_{\rho^*_{j^*}, j^*}$, which is generated from the beta prior (6.10).

That is, letting $\pi^*_{j^*}$ be the proposal vector of the partial autocorrelations:

- If $\rho^*_{j^*} = \rho_{j^*} - 1, \quad \pi^*_{j^*} = (\pi_{1,j^*}, \ldots, \pi_{\rho^*_{j^*}, j^*})$
- If $\rho^*_{j^*} = \rho_{j^*} + 1, \quad \pi^*_{j^*} = (\pi_{1,j^*}, \ldots, \pi_{\rho_{j^*}, j^*}, \pi^*_{\rho^*_{j^*}, j^*})$

with $\pi^*_{\rho^*_{j^*},j^*} \sim \mathrm{Be}_{(-1,+1)}\left(\pi_{i,j}|\left[\frac{i+1}{2}\right],\left[\frac{i}{2}\right]+1\right)$.

Note that in both cases all the autoregressive parameters are updated because of the reparametrization of section 6.4.

If $\rho^*_{j^*} = \rho_{j^*} - 1$, the acceptance probability ratio is $min(1, R)$, where R is given by:

$$R = \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{t:z_t=j}\left[(y_t-\nu_{j,t})^2-(y_t-\nu^*_{j,t})^2\right]\right\}\frac{b(\rho^*_{j^*})}{d(\rho_{j^*})} \tag{6.16}$$

where $\nu^*_{j^*,t} = \mu_{j^*} + \phi^*_{1,j^*}(y_{t-1}-\mu_{j^*}) + \cdots + \phi^*_{\rho_{j^*},j^*}(y_{t-\rho_{j^*}}-\mu_{j^*})$.

On the other hand, if $\rho^*_{j^*} = \rho_{j^*} + 1$:

$$R = \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{t:z_t=j}\left[(y_t-\nu_{j,t})^2-(y_t-\nu^*_{j,t})^2\right]\right\}\frac{d(\rho^*_{j^*})}{b(\rho_{j^*})} \tag{6.17}$$

Equations (6.16) and (6.17) are quite simple because of some cancellations between prior and proposal ratios. Furthermore, the jacobian is one because the matrix of derivatives of the transformation g (section 4.2.2) is the identity matrix (see appendix 6.B for details).

$p(\rho|k, x_T)$ is simply estimated by the proportions of every possible value for ρ in the sample obtained by the previous complete MCMC algorithm. In the following, the set of these seven moves will be referred as "complete MCMC ".

6.6.2 Number of the autoregressive components

Through Bayes' theorem, the marginal posterior distribution of k is:

$$p(k|x_T) \propto p(k)f(x_T|k)$$

where $p(k)$ is the prior on k and $f(x_T|k)$ is the marginal likelihood:

$$f(x_T|k) = \sum_{\rho}\int L(\psi,\rho,k)p(\psi,\rho|k)\,d\psi \tag{6.18}$$

with $\psi = (w, \mu, \sigma^2, \phi)$. Suppressing for notational convenience the model index k, we write the marginal likelihood (6.18) as:

$$\begin{aligned} f(x_T) &= \frac{L(\psi^*,\rho^*)p(\psi^*,\rho^*)}{p(\psi^*,\rho^*|x_T)} \\ &= \frac{L(\psi^*,\rho^*)p(\psi^*|\rho^*)p(\rho^*)}{p(\psi^*|\rho^*,x_T)p(\rho^*|x_T)} \end{aligned} \tag{6.19}$$

for a fixed point (ψ^*, ρ^*). Note that what we only need of the (6.19) is $p(\psi^*|\rho^*, x_T)$: we shall calculate the corresponding estimate $\bar{p}(\psi^*|\rho^*, x_T)$ by the methods of sections 4.3.3 and 4.3.2.

First of all, $\bar{p}(\psi^*|\rho^*, x_T)$ is factorized as:

$$\begin{aligned}\bar{p}(\psi^*|\rho^*,x_T) = {} & \bar{p}(\pi^*|\rho^*, x_T) \times \bar{p}(\mu^*|\pi^*, \rho^*, x_T)\times \\ & \times \bar{p}(\sigma^{2*}|\mu^*, \pi^*, \rho^*, x_T) \times \bar{p}(w^*|\sigma^{2*}, \mu^*, \pi^*, \rho^*, x_T) \end{aligned} \tag{6.20}$$

Suppose to have a sample $\{\psi^{(i)}, z^{(i)}\}$, for $i = 1, \ldots, N_1$, from the MCMC of section 6.5 for a given ρ^* (i.e. a sample from $p(\psi|\rho^*)$). Let $\eta_{j-1} = (\rho, \pi_1, \ldots, \pi_{j-1})$ and $\eta^{j+1} = (\pi_{j+1}, \ldots, \pi_k, \mu, \sigma^2, w)$. The terms of the (6.20) are estimated by the following steps:

1. Sample $\{\tilde{\eta}^{j+1,(i)}, \tilde{z}^{(i)}\}$, for $i = 1, \ldots, N_{j+1}$, from a reduced MCMC algorithm with distribution of interest $p(\eta^{j+1}, z|\eta_j^*, x_T)$. Also draw $\tilde{\pi}_j^{(i)}$ from $q_p(\pi_j^*, \pi_j) = \prod_{s=1}^{\rho_j} q(\pi_{s,j}^*, \pi_{s,j})$, where $q(.,.)$ is the proposal (6.13).

   Set:
   $$\bar{p}(\pi_j^*|\rho^*, \pi_1^*, \ldots, \pi_{j-1}^*) = \frac{N_j^{-1} \sum_{i=1}^{N_j} \alpha(\pi_j^{(i)}, \pi_j^*) q_p(\pi_j^{(i)}, \pi_j^*)}{N_{j+1}^{-1} \sum_{i=1}^{N_{j+1}} \alpha(\pi_j^*, \tilde{\pi}_j^{(i)})}$$
   where $\alpha(.,.) = \min(1, R)$ with R defined in equation (6.14).

   Set $\eta^{j+1,(i)} = \tilde{\eta}^{j+1,(i)}$ and $z^{(i)} = \tilde{z}^{(i)}$, for $i = 1, \ldots, N_{j+1}$.

   Repeat this step for $j = 1, \ldots, k$ and finally set:
   $$\bar{p}(\pi^*|\rho^*, x_T) = \prod_{j=1}^{k} \bar{p}(\pi_j^*|\rho^*, \pi_1^*, \ldots, \pi_{j-1}^*)$$

2. The second term is:
   $$\bar{p}(\mu^*|\pi^*, \rho^*, x_T) = N_{k+1}^{-1} \sum_{i=1}^{N_{k+1}} \prod_{j=1}^{k} p(\mu_j^*|\pi^*, \sigma^{2(i)}, z^{(i)}, \rho^*, x_T)$$
   where $(\sigma^{2(i)}, z^{(i)})$ are draws from the last iteration of the previous step (thus they are marginally from $p(\sigma^2, w, z|\pi^*, \rho^*, x_T)$) and $p(\mu_j^*|\pi^*, \sigma^{2(i)}, z^{(i)}, \rho^*, x_T)$ is given by equation(6.12).

3. Sample $\{\sigma^{2(s)}, w^{(s)}, z^{(s)}\}$, for $s = 1, \dots, S$, from a reduced MCMC algorithm with distribution of interest $p(\sigma^2, w, z|\pi^*, \mu^*, \rho^*, x_T)$ and set:

$$\bar{p}(\sigma^{2*}|\pi^*, \mu^*, \rho^*, x_T) = S^{-1} \sum_{s=1}^{S} \prod_{j=1}^{k} p(\sigma_j^{2*}|\pi^*, \mu^*, z^{(s)}, \rho^*, x_T)$$

   where $p(\sigma_j^{2*}|\pi^*, \mu^*, z^{(i)}, \rho^*, x_T)$ is given by equation (6.15).

4. Sample $\{w^{(v)}, z^{(v)}\}$, for $v = 1, \dots, V$, from a reduced MCMC algorithm with distribution of interest $p(w, z|\pi^*, \mu^*, \sigma^{2*}, \rho^*, x_T)$ and set:

$$\bar{p}(w^*|\pi^*, \mu^*, \sigma^{2*}, \rho^*, x_T) = V^{-1} \sum_{v=1}^{V} p(w^*|z^{(v)}, \rho^*, x_T)$$

   where $p(w^*|z^{(v)}, \rho^*, x_T)$ is given by equation (6.11)

6.7 Predictive distributions

6.7.1 One-step ahead predictions

The predictive distribution was defined in section 2.2. Consider an unknown observable future variable y_{T+1} and let (ρ^*, k^*) be a selected model. We can consider three different predictive distributions: conditional on (ρ^*, k^*), conditional on k^* and unconditional.

The predictive distribution conditional to (ρ^*, k^*) is:

$$\begin{aligned} f(y_{T+1}|\rho^*, k^*, x_T) &= \int f(y_{T+1}, \psi|\rho^*, k^*, x_T) d\psi \\ &= \int f(y_{T+1}|\psi, \rho^*, k^*, x_T) p(\psi|\rho^*, k^*, x_T) d\psi \qquad (6.21) \end{aligned}$$

where as usual $\psi = (w, \mu, \sigma^2, \phi)$ and where

$$f(y_{T+1}|\psi, \rho^*, k^*, x_T) = \sum_{j=1}^{k^*} w_j \, \mathrm{N}(y_{T+1}|\nu_{j,T+1}, \sigma_j^2) \qquad (6.22)$$

with $\nu_{j,T+1}$ defined in equation (6.4).

We estimate the (6.21) by the corresponding Monte Carlo estimator:

$$\bar{f}(y_{T+1}|\rho^*, k^*, x_T) = N^{-1} \sum_{i=1}^{N} f(y_{T+1}|\psi^{(i)}, \rho^*, k^*, x_T)$$

where $\psi^{(i)}$, for $i = 1, \ldots, N$, are samples from the posterior $p(\psi|\rho^*, k^*, x_T)$ and they are available from the MCMC output.

The predictive distribution unconditional to a model is more interesting because it takes into account the model uncertainty. As a matter of fact, different possible models are considered and they are weighted by their posterior probability.

The predictive distribution unconditional on the orders ρ is:

$$f(y_{T+1}|k^*, x_T) = \sum_{\rho} \int f(y_{T+1}|\psi, \rho, k, x_T) p(\psi, \rho|k, x_T) d\psi \tag{6.23}$$

which is estimated by:

$$\bar{f}(y_{T+1}|k^*, x_T) = N^{-1} \sum_{i=1}^{N} f(y_{T+1}|\psi^{(i)}, \rho^{(i)}, k, x_T) \tag{6.24}$$

Eventually, the predictive distribution unconditional on both k and ρ is:

$$f(y_{T+1}|x_T) = \sum_{k=1}^{k_{max}} f(y_{T+1}|k, x_T) p(k|x_T) \tag{6.25}$$

where $f(y_{T+1}|k, x_T)$ is given in (6.23). Monte Carlo estimator is simply:

$$\bar{f}(y_{T+1}|x_T) = \sum_{k=1}^{k_{max}} \bar{f}(y_{T+1}|k^*, x_T) p(k|x_T) \tag{6.26}$$

where $\bar{f}(y_{T+1}|k^*, x_T)$ is given in (6.24).

In order to achieve punctual predictions, we can calculate mean and variance of y_{T+1} with respect to the predictive densities. Conditionally on the model (ρ^*, k^*), the expected value is:

$$\begin{aligned} E(y_{T+1}|\rho^*, k^*, x_T) &= \int y_{T+1} f(y_{T+1}|\rho^*, k^*, x_T) dy_{T+1} = \\ &= \int y_{T+1} \left[\int f(y_{T+1}|\psi, \rho^*, k^*, x_T) p(\psi|\rho^*, k^*, x_T) d\psi \right] dy_{T+1} = \\ &= \int \left[\int y_{T+1} f(y_{T+1}|\psi, \rho^*, k^*, x_T) dy_{T+1} \right] p(\psi|\rho^*, k^*, x_T) d\psi = \\ &= \int U(\psi, \rho^*, k^*) p(\psi|\rho^*, k^*, x_T) d\psi \end{aligned}$$

where $U(\psi, \rho^*, k^*) = E(y_{T+1}|\psi, \rho^*, k^*, x_T)$ is defined in expression (6.5). The corresponding Monte Carlo estimator is then:

$$N^{-1} \sum_{i=1}^{N} U(\psi^{(i)}, \rho^*, k^*) \tag{6.27}$$

The variance is:

$$\begin{aligned} VAR(y_{T+1}|\rho^*, k^*, x_T) = E_\psi[VAR(y_{T+1}|\psi, \rho^*, k^*, x_T)]+ \\ + VAR_\psi[E(y_{T+1}|\psi, \rho^*, k^*, x_T)] \end{aligned}$$

where the first term is estimated by:

$$N^{-1} \sum_{i=1}^{N} VAR(y_{T+1}|\psi^{(i)}, \rho^*, k^*, x_T)$$

and the second one by:

$$(N-1)^{-1} \sum_{i=1}^{N} [U(\psi^{(i)}, \rho^*, k^*) - \bar{U}]^2$$

where $\bar{U} = N^{-1} \sum_{i=1}^{N} U(\psi^{(i)}, \rho^*, k^*)$.

The moments of y_{T+1} with respect to the unconditional predictive distributions are similarly estimated. Unconditional on ρ, the expected value $E(y_{T+1}|k^*, x_T)$ is estimated by:

$$N^{-1} \sum_{i=1}^{N} U(\psi^{(i)}, \rho^{(i)}, k^*)$$

where as usual $(\psi^{(i)}, \rho^{(i)})$ are samples from the posterior $p(\psi, \rho|k^*, x_T)$.

The variance is:

$$\begin{aligned} VAR(y_{T+1}|x_T) = E_{\psi,\rho}[VAR(y_{T+1}|\psi, \rho, k^*, x_T)]+ \\ + VAR_{\psi,\rho}[E(y_{T+1}|\psi, \rho, k^*, x_T)] \end{aligned}$$

where the first term is estimated by:

$$N^{-1} \sum_{i=1}^{N} VAR(y_{T+1}|\psi^{(i)}, \rho^{(i)}, k^*, x_T)$$

and the second one by:

$$(N-1)^{-1}\sum_{i=1}^{N}[U(\psi^{(i)},\rho^{(i)},k^{*})-\bar{U}]^{2}$$

where $\bar{U}=N^{-1}\sum_{i=1}^{N}U(\psi^{(i)},\rho^{(i)},k^{*})$.

Unconditional on both ρ and k, the expected value $E(y_{T+1}|x_T)$ is estimated by:

$$\sum_{k=1}^{k_{max}}\left[N^{-1}\sum_{i=1}^{N}U(\psi^{(i)},\rho^{(i)},k)\right]p(k|x_T)$$

The variance is:

$$\begin{aligned}VAR(y_{T+1}|x_T)=E_{\psi,\rho,k}[VAR(y_{T+1}|\psi,\rho,k,x_T)]+\\+VAR_{\psi,\rho,k}[E(y_{T+1}|\psi,\rho,k,x_T)]\end{aligned}$$

where the first term is estimated by:

$$\sum_{k=1}^{k_{max}}\left[N^{-1}\sum_{i=1}^{N}VAR(y_{T+1}|\psi^{(i)},\rho^{(i)},k,x_T)\right]p(k|x_T)$$

and the second one by:

$$\sum_{k=1}^{k_{max}}\left[(N-1)^{-1}\sum_{i=1}^{N}[U(\psi^{(i)},\rho^{(i)},k)-\bar{U}]^{2}\right]p(k|x_T)$$

where $\bar{U}=N^{-1}\sum_{i=1}^{N}U(\psi^{(i)},\rho^{(i)},k)$.

6.7.2 Multiple-step ahead predictions

The above formulaes are devoted to estimate the predictive distributions for a single future observation. The computation of the m-step predictive distributions is less straightforward.

Suppose to consider the two-step predictive distributions. A first method consists in using a punctual one-step forecast, for instance the estimates of the expected values $E(y_{T+1}|\rho^{*},k^{*},x_T)$, $E(y_{T+1}|k^{*},x_T)$ or $E(y_{T+1}|x_T)$ as if it is the true value of y_{T+1}.

Thus, the two-step predictive distribution conditional to a model is:

$$f(y_{T+2}|\rho^{*},k^{*},x_T)=f(y_{T+2}|\rho^{*},k^{*},x_T,y_{T+1}=\bar{y}_{T+1})$$

where $\bar{y}_{T+1}$ is defined in (6.27). The unconditional predictive distribution is:

$$f(y_{T+2}|x_T) = f(y_{T+2}|x_T, y_{T+1} = \hat{y}_{T+1})$$

where $\hat{y}_{T+1}$ is previously defined.

This approach ignores any information from the shape of the one-step predictive distribution and it could be unsatisfactory, as pointed out by Wong and Li (2000). Alternatively, a better approach is to estimate the exact two-step predictive distribution trough Monte Carlo method, using samples from the one-step predictive distribution.

For the conditional case, we have:

$$\begin{aligned} f(y_{T+2}|\rho^*, k^*, x_T) &= \int f(y_{T+2}|\psi, \rho^*, k^*, x_T)p(\psi|\rho^*, k^*, x_T)d\psi = \\ &= \int \left[\int f(y_{T+2}|\psi, \rho^*, k^*, x_T, y_{T+1}) f(y_{T+1}|\psi, \rho^*, k^*, x_t) dy_{T+1} \right] \\ &\quad p(\psi|\rho^*, k^*, x_T)d\psi \end{aligned}$$

The Monte Carlo estimators is:

$$\bar{f}(y_{T+2}|\rho^*, k^*, x_T) = M^{-1} \sum_{j=1}^{M} \left[N^{-1} \sum_{i=1}^{N} f(y_{T+2}|\psi^{(i)}, \rho^*, k^*, x_T, y_{T+1}^{(j)}) \right]$$

where $\psi^{(i)}$, for $i = 1, \ldots, N$, are samples from the posterior $p(\psi|\rho^*, k^*, x_T)$ (available from the MCMC output) and $y_{T+1}^{(j)}$ are samples from the one-step predictive distribution (equation 6.22).

Similarly, the unconditional two-step predictive distribution is estimated by:

$$\bar{f}(y_{T+2}|\rho^*, k^*, x_T) = \sum_{k=1}^{k_{max}} \left\{ M^{-1} \sum_{j=1}^{M} \left[N^{-1} \sum_{i=1}^{N} f(y_{T+2}|\psi^{(i)}, \rho^{(i)}, k, x_T, y_{T+1}^{(j)}) \right] \right\} p(k|x_T)$$

Appendix 6.A

In the following, we shall use "$|\dots$" to denote conditioning on all other variables.

Move (i)

The full conditional for w is derived in a similar way as in Appendix 5.A:

$$\begin{aligned}
p(w|\dots) &\propto p(w,\mu,\phi,\sigma^2,\beta,z,\rho,x_T)\\
&\propto L(\mu,\phi,\sigma^2,\beta,z,\rho)p(z|w)p(w)p(\mu)p(\phi|\rho)p(\rho)p(\sigma^2|\beta)p(\beta)\\
&\propto p(z|w)p(w)\\
&\propto \prod_{j=1}^{k} w_j^{n_j}\prod_{j=1}^{k} w_j^{\delta_j-1}\\
&= \mathrm{Di}(w|\delta_1+n_1,\dots,\delta_k+n_k)
\end{aligned}$$

where $n_j = \sum_{j=1}^{k} I_{(z_t=j)}$, for $j=1,\dots,k$.

Move (ii)

Using equations (6.9) and (6.6):

$$\begin{aligned}
p(\mu|\dots) &\propto L(\mu,\phi,\sigma^2,\beta,z,\rho)p(\mu)\\
&= \prod_{j=1}^{k}\prod_{t:z_t=j} \mathrm{N}(y_t|\nu_{j,t},\sigma_j^2)\prod_{j=1}^{k}\mathrm{N}(\mu_j|\mu_0,\tau^2)
\end{aligned}$$

Hence:

$$p(\mu_j|\dots) \propto \left[\prod_{t:z_t=j}\mathrm{N}(y_t|\nu_{j,t},\sigma_j^2)\right]\mathrm{N}(\mu_j|\mu_0,\tau^2)$$

for $j=1,\dots,k$. Writing the normal density function and substituting the (6.4), we obtain:

$$\begin{aligned}
p(\mu_j|\dots) \propto \exp\{-\frac{1}{2\sigma_j^2}\sum_{t:z_t=j}[y_t-\mu_j-\phi_{1,j}(y_{t-1}-\mu_j)-\dots\\
\dots-\phi_{\rho_j,j}(y_{t-\rho_j}-\mu_j)]^2-\frac{1}{2\tau^2}(\mu_j-\mu_0)^2\}
\end{aligned}$$

Letting $v_{t,j} = y_t - \phi_{j,1}y_{t-1} - \cdots - \phi_{j,\rho_j}y_{t-\rho_j}$ and $B = 1 - \phi_{j,1} - \cdots - \phi_{j,\rho_j}$,

$$\begin{aligned}
p(\mu_j|\dots) &\propto \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{t:z_t=j}(v_{t,j} - \mu_j B)^2 - \frac{1}{2\tau^2}(\mu_j - \mu_0)^2\right\} \\
&= \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{t:z_t=j}(v_{t,j} - \bar{v}_j + \bar{v}_j - \mu_j B)^2 - \frac{1}{2\tau^2}(\mu_j - \mu_0)^2\right\} \\
&= \exp\left\{-\frac{1}{2\sigma_j^2}\sum_{t:z_t=j}[(v_{t,j} - \bar{v}_j)^2 + n_j(\bar{v}_j - \mu_j B)^2] - \frac{1}{2\tau^2}(\mu_j - \mu_0)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_j^2}n_j(\bar{v}_j - \mu_j B)^2 - \frac{1}{2\tau^2}(\mu_j - \mu_0)^2\right\} \\
&\propto \exp\left\{-\left[\left(\frac{n_j B^2}{2\sigma_j^2} + \frac{1}{2\tau^2}\right)\mu_j^2 - \left(\frac{n_j\bar{v}_j B}{\sigma_j^2} + \frac{\mu_0}{\tau^2}\right)\mu_j\right]\right\},
\end{aligned}$$

where $\bar{v}_j = \frac{1}{n_j}\sum_{t:z_t=j} v_{tj}$. Now, if we let:

$$A = \frac{n_j B^2}{2\sigma_j^2} + \frac{1}{2\tau^2} = \frac{1}{2}\frac{n_j B^2\tau^2 + \sigma_j^2}{\sigma_j^2\tau^2} \tag{6.28}$$

$$D = \frac{1}{2}\left(\frac{n_j\bar{v}_j B}{\sigma_j^2} + \frac{\mu_0}{\tau^2}\right) = \frac{1}{2}\frac{n_j\bar{v}_j B\tau^2 + \mu_0\sigma_j^2}{\sigma_j^2\tau^2} \tag{6.29}$$

then:

$$\begin{aligned}
p(\mu_j|\dots) &\propto \exp\{-\left[A\mu_j^2 - 2D\mu_j\right]\} \\
&\propto \exp\left\{-A\left[\mu_j^2 - 2\frac{D}{A}\mu_j\right]\right\} \\
&\propto \exp\left\{-A\left[\left(\mu_j - \frac{D}{A}\right)^2 - \frac{D^2}{A^2}\right]\right\} \\
&\propto \exp\left\{-A\left(\mu_j - \frac{D}{A}\right)^2\right\}
\end{aligned}$$

Finally, substituting the (6.28) and (6.29), we obtain:

$$\begin{aligned}
p(\mu_j|\dots) &\propto \exp\left\{\frac{1}{2}\frac{n_j B^2\tau^2 + \sigma_j^2}{\sigma_j^2\tau^2}\left(\mu_j - \frac{n_j\,\bar{v}_j\,B\,\tau^2 + \sigma_j^2\mu_0}{n_j\,B^2\tau^2 + \sigma_j^2}\right)^2\right\} \\
&= \mathrm{N}\left(\mu_j\middle|\,\frac{n_j\,\bar{v}_j\,B\,\tau^2 + \sigma_j^2\mu_0}{n_j\,B^2\tau^2 + \sigma_j^2}, \frac{\sigma_j^2\tau^2}{n_j\,B^2\tau^2 + \sigma_j^2}\right).
\end{aligned}$$

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

Move (iii)

From section (3.4.1), we know that the ratio in the acceptance probability for a Metropolis-Hastings algorithm can be represented as:

$$\text{(likelihood ratio)} \times \text{(prior ratio)} \times \text{(proposal ratio)}$$

In the case of move (iii), the likelihood ratio is simply given by (see equation 6.9):

$$\text{likelihood ratio } = \frac{\prod_{j=1}^{k} \prod_{t:z_t=j} \mathrm{N}(y_t|\nu_{j,t}^*, \sigma_j^2)}{\prod_{j=1}^{k} \prod_{t:z_t=j} \mathrm{N}(y_t|\nu_{j,t}, \sigma_j^2)} \tag{6.30}$$

where

$$\nu_{j,t}^* = \mu_j + \phi_{1,j}^*(y_{t-1} - \mu_j) + \cdots + \phi_{\rho_j,j}^*(y_{t-\rho_j} - \mu_j) \tag{6.31}$$

Substituting the normal density function and the equations (6.30) and (6.31), we obtain the first line of the (6.14).

All the parameters but the AR coefficients do not change and the prior ratio in terms of the partial autocorrelations π reduces to the ratio of two generalized beta (equation 6.10):

$$\text{prior ratio } = \frac{\prod_{i=1}^{\rho_j} \mathrm{Be}_{(-1,+1)}\left(\pi_{i,j}^*|\left[\frac{i+1}{2}\right], \left[\frac{i}{2}\right]+1\right)}{\prod_{i=1}^{\rho_j} \mathrm{Be}_{(-1,+1)}\left(\pi_{i,j}|\left[\frac{i+1}{2}\right], \left[\frac{i}{2}\right]+1\right)}$$

Substituting the density from section 2.5.1, the second line of the (6.14) is easily derived.

Eventually, from the equation (6.13) and from section 2.5.5, the proposal ratio is (third line of the (6.14)):

$$\begin{aligned}
\text{proposal ratio } &= \frac{\prod_{i=1}^{\rho_j} \mathrm{N}_{(-1,+1)}(\pi_{i,j}|\pi_{j,i}^*, \sigma_q^2)}{\prod_{i=1}^{\rho_j} \mathrm{N}_{(-1,+1)}(\pi_{i,j}^*|\pi_{j,i}, \sigma_q^2)} \\
&= \frac{\prod_{i=1}^{\rho_j} \frac{N(\pi_{j,i}|\pi_{j,i}^*,\sigma_q^2)}{F_N(1|\pi_{j,i}^*,\sigma_q^2)-F_N(-1|\pi_{j,i}^*,\sigma_q^2)}}{\prod_{i=1}^{\rho_j} \frac{N(\pi_{j,i}^*|\pi_{j,i},\sigma_q^2)}{F_N(1|\pi_{j,i},\sigma_q^2)-F_N(-1|\pi_{j,i},\sigma_q^2)}} \\
&= \prod_{i=1}^{\rho_j} \frac{F_N(1|\pi_{j,i}, \sigma_q^2) - F_N(-1|\pi_{j,i}, \sigma_q^2)}{F_N(1|\pi_{j,i}^*, \sigma_q^2) - F_N(-1|\pi_{j,i}^*, \sigma_q^2)}
\end{aligned}$$

where F_N is the normal cumulative distribution function.

Move (iv)

Using equations (6.9) and (6.6):

$$
\begin{aligned}
p(\sigma^2|\dots) &\propto L(\mu,\phi,\sigma^2,\beta,z,\rho)p(\sigma^2|\beta)\\
&= \prod_{j=1}^{k}\prod_{t:z_t=j}\mathrm{N}(y_t|\nu_{j,t},\sigma_j^2)\prod_{j=1}^{k}\mathrm{Ig}(\sigma_j^2|\alpha,\beta)
\end{aligned}
$$

Hence, for $j=1,\dots,k$, the full conditional for σ_j^2 is:

$$
\begin{aligned}
p(\sigma_j^2|\dots) &\propto \left[\prod_{t:z_t=j}\mathrm{N}(y_t|\nu_{j,t},\sigma_j^2)\right]\mathrm{Ig}(\sigma_j^2|\alpha,\beta)\\
&\propto \sigma_j^{-n_j}\exp\left\{-\frac{1}{2}\sum_{t:z_t=j}\frac{(y_t-\nu_{j,t})^2}{\sigma_j^2}\right\}\sigma_j^{-2(\alpha-1)}\exp\{-\beta/\sigma_j^2\}\\
&= \sigma_j^{-2(n_j/2+\alpha-1)}\exp\left\{\left[-\frac{1}{2}\sum_{t:z_t=j}(y_t-\nu_{j,t})^2\right]/\sigma_j^2\right\}\\
&= \mathrm{Ig}\left(\sigma_j^2\middle|\,\alpha+\frac{1}{2}n_j,\beta+\frac{1}{2}\sum_{t:z_t=j}(y_t-\nu_{j,t})^2\right)
\end{aligned}
$$

Move (v)

The full conditional for the allocation variable z is derived as in Appendix 4.B, with the only difference of $\nu_{j,t}$ instead of μ_j (equation 5.17).

Appendix 6.B

Move (vi)

For the reversible jump algorithm, we can represent the ratio in the acceptance probability as (section 4.2.2):

(likelihood ratio) × (prior ratio) × (proposal ratio) × (jacobian)

The likelihood ratio is the same of equation (6.30):

$$
\text{likelihood ratio } = \frac{\prod_{j=1}^{k}\prod_{t:z_t=j}\mathrm{N}(y_t|\nu_{j,t}^{*},\sigma_j^2)}{\prod_{j=1}^{k}\prod_{t:z_t=j}\mathrm{N}(y_t|\nu_{j,t},\sigma_j^2)}
$$

but here the conditional mean $\nu^*_{j,t}$ is:

$$\nu^*_{j,t} = \mu_j + \phi^*_{1,j}(y_{t-1} - \mu_j) + \cdots + \phi^*_{\rho^*_j,j}(y_{t-\rho^*_j} - \mu_j)$$

because the model order is updated by this move.

The parameters updated by this move are ρ_j and π_{ρ_j}; hence, the prior ratio is given by:

$$\text{prior ratio } = \frac{p(\rho^*_j)p(\pi^*_{\rho_j})}{p(\rho_j)p(\pi_{\rho_j})} = \frac{\prod_{i=1}^{\rho^*_j} p(\pi^*_{i,j})}{\prod_{i=1}^{\rho_j} p(\pi_{i,j})}$$

because of the uniform prior on the model orders. Since only the jth component is updated, if $\rho^*_j = \rho_j - 1$:

$$\text{prior ratio } = \frac{1}{p(\pi_{\rho_j,j})}$$

Otherwise, if $\rho^*_j = \rho_j + 1$:

$$\text{prior ratio } = p(\pi_{\rho^*_j,j})$$

To calculate the proposal ratio we have to take into account the probabilities $b(\cdot)$ and $d(\cdot)$ and the proposal density, which is chosen equal to the prior on π. If $\rho^*_j = \rho_j - 1$:

$$\text{proposal ratio } = \frac{b(\rho^*_j)p(\pi_{\rho_j,j})}{d(\rho_j)}$$

Otherwise, if $\rho^*_j = \rho_j + 1$:

$$\text{proposal ratio } = \frac{d(\rho^*_j)}{b(\rho_j)p(\pi_{\rho^*_j,j})}$$

If $\rho^*_j = \rho_j - 1$, then $pi^*_{i,j} = pi_{i,j}$, for $i = 1, \ldots, \rho^*_j$. On the other hand, if $\rho^*_j = \rho_j + 1$, then $\pi^*_{i,j} = pi_{i,j}$, for $i = 1, \ldots, \rho_j$, and $\pi^*_{\rho^*_j,j}$ is generated by the beta prior (6.10). As consequence, using the terminology of section 4.2.2, the invertible function $g(\cdot)$ is the identity function and the jacobian is equal to one.

Multiplying likelihood, prior and proposal ratios, equations (6.16) and (6.17) are easily derived.

Chapter 7

Return volatility

7.1 Introduction

This chapter shows the application of the mixture of autoregressive components to the return volatility, reporting definition and properties of the model and summarizing parameter estimation and model selection procedures in order to be a self-contained chapter.

Modelling and forecasting return volatility is one of the most important tasks in financial markets. Within a rich literature (for a recent review see Poon and Granger, 2003), several stylized facts have been recognized.

First of all, volatility is persistent (e.g. Poterba and Summers, 1986, Schwert, 1987, French *et al.*, 1987, and Hsieh, 1991) and it can have long memory properties (e.g. Ding *et al.*, 1993, and Bollerslev and Mikkelsen, 1996).

Second, observations of financial time series reveal volatility clustering. Autoregressive conditional heteroskedasticity (ARCH) models and stochastic volatility (SV) models (for a survey, see Bollerslev *et al.*, 1992, and Ghysels *at al.*, 1996, respectively) are well-known instruments proposed in literature and they are essentially built to mimic this volatility feature.

Moreover, volatility shows threshold effects, non-symmetrical dependencies and mean reversion. In particular, it has been argued that volatility adjustments follow a twin-speed process: low volatility state is more persistent with respect to high state (Longing, 1987, Jones *et al.*, 1998). In order to consider changing volatility persistence, models in a regime switching framework have been proposed (Hamilton, 1989, Cao and Tsay, 1992, Gray, 1996). Volatility asymmetry motivates several non-linear GARCH type models, like the exponential GARCH (Nelson, 1991), the quadratic GARCH (Engle, 1990) and the JGR-GARCH (Glosten *et al.*, 1993).

In this work we follow the approach which uses an observable proxy for the return volatility. This choice entails some empirical advantages, allowing to use methods directly based on observable variables.

We model the log volatility through the mixture of autoregressive components (chapter 6) which capture the previously mentioned stylized facts. The autoregressive nature on the mixture components explicitly formalizes the volatility intertemporal dependence. The clustering effect is also considered. In fact, the model assumes that, at each time t, the observable proxy for the volatility is drawn from one of a set of different autoregressive models (regimes) with probabilities equal to the mixture weight. In addition, a mixture model is a flexible technique to obtain departures from normality and the conditional distribution can be multimodal or non-symmetric.

As we shown in the previous chapter, Wong and Li (2000) introduced the mixture of autoregressive models, performing a numerical maximum likelihood estimation. A single model is jointly specified by the number of the autoregressive components of the mixture and by the orders of such autoregressive components. In order to solve the model selection problem, they consider the AIC and the BIC criterions and they conclude that these conventional approaches are not satisfactory in this context.

We propose a fully Bayesian approach, in which model determination is based on indexing all the models under consideration and considering this index as a variable. Since standard approaches select a single model and then make inference based on this model, model uncertainty is not taken into account. Conversely, even with a very high number of possible models, we do not ignore model uncertainty because we maintain consideration of several models, with the input of each into the analysis weighted by the model posterior probability. Influence of model uncertainty on financial models has been recognized as an important factor and it has been investigated by some recent papers (Barberis, 2000, MacKinley and Pastor, 2000, Pastor, 2000, Pastor and Stambaugh, 2000, Cremers, 2002).

A Bayesian approach in time series analysis is usually difficult because of the existence of stationary and invertible conditions on the model parameters. These constraints have been often ignored (see e.g. Zellner, 1971, Broemeling and Shaaraway, 1988). Instead, we consider a prior setting compatible with the stationarity conditions on the autoregressive parameters through a reparametrization in terms of partial autocorrelations (Barndorff-Nielsen and Schou, 1973, Barbieri and O'Hagan, 1997).

Chapter 6 treated the estimation procedure and the model selection which are based on *Markov Chain Monte Carlo* (or *MCMC*) methods. In particular, model selection is performed combining a *Reversible jump* sampler (Green, 1995) and a marginal likelihood estimation (Chib and Jeliazkov, 2001).

7.2 Mixture of autoregressive components

Let y_t be the observable variable, for $t = 1, \dots, T$. The mixture of autoregressive components can be defined by:

$$y_t | \dots \sim \sum_{j=1}^{k} w_j \; \mathrm{N}(y_t | \nu_{j,t}, \sigma_j^2) \tag{7.1}$$

where "$| \dots$" is used to denote conditioning on the past observations and on all other variables, $\mathrm{N}(\cdot|a, b)$ stands for the Normal distribution with mean a and variance b, and:

$$\nu_{j,t} = \mu_j + \phi_{1,j}(y_{t-1} - \mu_j) + \cdots + \phi_{\rho_j,j}(y_{t-\rho_j} - \mu_j) \tag{7.2}$$

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

Note that equation (7.2) is the conditional mean of a stationary autoregressive model, with order ρ_j, stationary mean μ_j and autoregressive coefficients $\phi_j = (\phi_{1,j}, \dots, \phi_{\rho_j,j})$.

The mixing weights w_j satisfy the constraints:

$$\begin{aligned} &w_j > 0, \quad j = 1, \dots, k \\ &w_1 + \cdots + w_k = 1 \end{aligned}$$

Clearly, the model captures the volatility persistence since the observable variable is formalized as a function of the past values. As we shown in chapter 6, we consider the orders of the autoregressive components ρ_j as random variables through a Bayesian perspective. As a consequence, inference is not based on a unique fixed level of persistence, but it considers different levels whose contributions are given by the model posterior probabilities.

In addition, the mixture of autoregressive components is a flexible nonlinear model which can capture other stylized facts of the volatility. In fact, the model assumes that, at each time t, the volatility is drawn from one of a set of different regimes with probabilities equal to the mixture weight. For instance, the twin-speed process (low volatility states are more persistent than high states, Longing, 1987, Jones *et al.*, 1998) is strongly recognized by our model (see section 7.7).

7.3 Bayesian analysis

Suppose to adopt the concise notation $\rho = (\rho_1, \ldots, \rho_k)$, $\mu = (\mu_1, \ldots, \mu_k)$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_k^2)$, $\phi = (\phi_1, \ldots, \phi_k)$. We set the following prior distributions for the parameters w, μ and σ^2:

$$w|k \sim \text{Di}(w|\delta, \delta, \ldots, \delta) \tag{7.3}$$

$$\mu_j \overset{iid}{\sim} \text{N}(\mu_j|\mu_0, \tau^2), \qquad j = 1, \ldots, k \tag{7.4}$$

$$\sigma_j^2 \overset{iid}{\sim} \text{Ig}(\sigma^2|\alpha, \beta), \qquad j = 1, \ldots, k. \tag{7.5}$$

where $\text{Di}(\cdot|\delta, \delta, \ldots, \delta)$ and $\text{Ig}(\cdot|\alpha, \beta)$ denote the Dirichlet and the Inverted-Gamma distributions respectively. The parameters δ, μ_0, τ^2, α and β are assumed to be fixed (see section 7.7).

The prior distributions (7.3), (7.4) and (7.5) are conventional choices for a mixture model (see e.g. Diebolt and Robert, 1994, Richardson and Green, 1997). Fully non-informative priors do not lead to proper posterior distributions in a mixture context and independent improper priors cannot be used (Diebolt and Robert, 1994).

Our analysis deals with two important issues. The first one is the so-called label switching problem, which derives from the symmetry in the likelihood of the parameters (chapters 5 and 6). In a Bayesian analysis, if we have no prior information that distinguishes between the components of the mixture (that is, the joint prior distribution is the same for all permutations of the parameters), then the posterior distribution will be similarly symmetric (see Fruhwirth-Schnatter, 2001, for a proof). As a result, the posterior distribution shows artificial multimodality, which poses obvious problems in terms of parameter estimations.

The usual solution of the label switching problem consists in imposing an *identifiability constraint* on the parameter space, such as $\mu_1 < \mu_2 < \ldots, < \mu_k$. This kind of constraints can be satisfied by only one permutation of ψ and this breaks the symmetry of the prior (e.g. Albert and Chib, 1993, McCulloch and Tsay, 1994, Engle and Kim, 1999).

The second issue we address is the existence of the stationarity regions for the autoregressive coefficients ϕ. Apart from the cases of small orders, it is well-known that the form of these regions for an autoregressive model is very complex. As a consequence, Bayesian analysis can be difficult in terms of prior specification and these constraints have been often ignored (see e.g. Zellner, 1971, Broemeling and Shaaraway, 1988).

Let Φ_j be the stationarity region for the *j*th autoregressive component and let $\pi_{h,j}$ be the partial autocorrelation coefficient at lag h for the *j*th model.

Through the reparametrization introduced by Barndorff-Nielsen and Schou (1973) (section 6.4), which establishes a one-to-one transformation between $\phi_j = (\phi_{1,j}, \ldots, \phi_{\rho_j,j})$ and $\pi_j = (\pi_{1,j}, \ldots, \pi_{\rho_j,j})$, and using the following result (Jones, 1987):

$$\phi_j \sim \text{Uniform on } \Phi_j \Longleftrightarrow \pi_{i,j} \overset{ind}{\sim} \text{Be}_{(-1,+1)}\left(\pi_{i,j} \middle| \left[\frac{i+1}{2}\right], \left[\frac{i}{2}\right]+1\right)$$

for $i = 1, \ldots, \rho_j$ and $j = 1, \ldots, k$, where $\text{Be}_{(-1,+1)}(.)$ denotes a generalized beta distribution defined on $(-1, +1)$ and where $[x]$ means "integer part of x", we can put a uniform prior distribution for the original parameters on the complicate stationarity region Φ_j, simply choosing a generalized beta prior for the $\pi_{i,j}$'s on $(-1, +1)$.

Finally, k and $\rho = (\rho_1, \ldots, \rho_k)$ will be considered as stochastic quantities with the following priors:

$$\rho_j|\rho_{max} \overset{iid}{\sim} \text{Un}(\rho_j|0, \rho_{max}), \qquad j = 1, \ldots, k \tag{7.6}$$

$$k|k_{max} \sim \text{Un}(k|1, k_{max}) \tag{7.7}$$

where $\text{Un}(x|a, b)$ denotes the discrete uniform distribution for $a \leq x \leq b$ and where ρ_{max} and k_{max} are fixed.

7.4 Parameter estimation

As we noticed in the introduction, k and ρ jointly specify a possible model. Let $\theta = (w, \mu, \sigma^2, \phi)$ be the complete parameters vector and let $m = (\rho, k)$ be the model index. The model determination problem will be discussed in the next section. For the moment, suppose to consider m as a fixed quantity.

The posterior distribution of θ conditionally on m is given by Bayes' theorem:

$$p(\theta|y, m) = \frac{f(y|\theta, m)p(\theta|m)}{\int f(y|\theta, m)p(\theta|m)d\theta} \tag{7.8}$$

where $f(y|\theta, m)$ is the likelihood function and $p(\theta|m)$ is the joint prior distribution for θ.

Because of the analytically intractable denominator of equation (7.8), we approximate the posterior distribution by a *Markov Chain Monte Carlo* (or *MCMC*) method. In particular, we implemented an MCMC strategy based on a componentwise algorithm (section 3.4.2). In short, suppose to split the parameter vector θ into L blocks $(\theta_1, \theta_2, \ldots, \theta_L)$. Each iteration of the algorithm is formed by L moves, each of them updates a single block θ_i. In practice, for $i = 1, \ldots, L$, a possible value for the chain, say θ'_i, is generated from a specific proposal density $q_i(\theta'_i|\theta_i)$. θ'_i is accepted with probability:

$$\alpha_i(\theta_i, \theta'_i) = \min\left(1, \frac{\pi(\theta')q_i(\theta_i|\theta'_i)}{\pi(\theta)q_i(\theta'_i|\theta_i)}\right) \tag{7.9}$$

where $\theta' = (\theta_1, \ldots, \theta'_i, \ldots, \theta_L)$. If θ'_i is rejected, the chain remains in θ_i.

The ith move related to equation (7.9) is referred as Metropolis-Hastings type move (Metropolis et al., 1953, Hastings, 1970). If the proposal density $q_i(\theta'_i|\theta_i)$ is equal to the full conditional $p(\theta'_i|\theta_{-i})$, where $\theta_{-i} = (\theta_1, \theta_2, \ldots, \theta_{i-1},$

$\theta_{i+1}, \ldots, \theta_L$), the probability (7.9) is shown to be 1 and the proposal value is always accepted. In this case, the move is said Gibbs type move (Geman and Geman, 1984).

In our analysis, θ is split into four blocks corresponding to the parameter (w, μ, σ^2, ϕ). In order to implement the MCMC strategy, it is necessary to consider an additional move that updates the allocation variable $z = (z_1, \ldots, z_T)$ (chapters 5 and 6). w, μ, σ^2 and z are updated by Gibbs type moves, while ϕ is updated by a Metropolis-Hastings move. For details see Appendix 7.A or section 6.5.

7.5 Model determination

As pointed out by Wong and Li (2000), traditional criterions like the AIC (Akaike, 1973) and the BIC (Schwarz, 1978) are not satisfactory for the mixture of autoregressive components, especially in selecting the number of components k. The difficulties arise from the non-standard application of these criterions, as in testing problems with nuisance parameters that are present only under the alternative hypothesis.

We propose a Bayesian model determination procedure. The key idea is to index all the models under consideration and consider this index as a further (stochastic) parameter. In general, this entails some advantages with respect to other approaches. First of all, results are simple to interpret: conclusions like "the (posterior) probabilities that m and m' are true are 0.87 and 0.13 respectively" are easy to understand even with a limited statistical background.

In addition, Bayesian model determination acts as an "Occam's razor", selecting a simpler model over a more complex one if both are compatible with the data.

Another fundamental feature is that model uncertainty is taken into account. Standard approach selects a single model from a class of candidate models and then makes inference based on this model. This procedure ignores model uncertainty and it could provide small predictive precisions (see Draper, 1995, for a discussion). Conversely, Bayesian model determination can account for model uncertainty because one can maintain consideration of several models, with the input of each into the analysis weighted by the model posterior probability.

To compare different models, it is natural to use the marginal posterior distribution of the model index which is derived by Bayes' theorem:

$$p(m|y) = \frac{f(y|m)p(m)}{\sum_m f(y|m)p(m)} \tag{7.10}$$

where $p(m)$ is the discrete prior for the models and where $f(y|m)$ is the marginal likelihood:

$$f(y|m) = \int f(y|\theta, m)p(\theta|m)d\theta \tag{7.11}$$

The calculation of the posterior (7.10) poses the usual problems in terms of analytical tractability and several methods that are able to deal with model selection were implemented in literature. It is possible to divide them into two main categories: *across*- and *within*- model simulation methods.

The across-model simulation approach is based on an MCMC simulation with states of the form (m, θ). The distribution of interest is the joint posterior of the parameters and the model index. The marginal posterior distribution of m is simply estimated by the proportions of m's in the sample obtained by the MCMC algorithm.

In the within-model simulations, the aim of finding $p(m|y)$ for all m is reached by estimating all the marginal likelihoods $f(y|m)$. Once $f(y|m)$ for all m is estimated, it is sufficient to normalize the products $f(y|m)p(m)$ to achieve the marginal posterior probabilities (see equation 7.10).

As we explained in chapter 6, our analysis is performed using a combination of these classes of methods. Specifically, we evaluate the two posterior distributions $p(\rho|k, y)$ and $p(k|y)$ by the following strategy:

i. $p(\rho|k, y)$ is obtained by adding a reversible jump type move (Green, 1995) to the previously illustrated MCMC strategy which updates the orders ρ (across-model simulation).

ii. The output from the step i is used to derive the marginal posterior $p(k|y)$ using the method by Chib and Jeliazkov (2001) (within-model simulation).

Reversible jump sampler (section 4.2.2) can be viewed as a Metropolis-Hastings method adapted for general state spaces. Suppose to denote the dimension of a parameter vector θ with $d(\theta)$. Let (θ, m) be the current value of the chain and let $g_{m,m'}$ be an invertible function. At each iteration, a candidate model m' is generated from $q(m'|m)$. Note that θ depends on m, so $d(\theta)$ can change. In order to have a dimension matching, a random vector u is generated from a proposal density $q(u|\theta, m, m')$. Set $(\theta', u') = g_{m,m'}(\theta, u)$, with $d(u') = d(\theta) + d(u) - d(\theta')$. The proposed value (θ', m') is accepted with probability:

$$\alpha[(\theta, m), (\theta', m')] = \min\left(1, \frac{h(\theta', m')}{h(\theta, m)} \times \frac{q(m|m')q(u'|\theta', m', m)}{q(m'|m)q(u|\theta, m, m')} \times \left|\frac{\partial g_{m,m'}(\theta, u)}{\partial(\theta, u)}\right|\right)$$

where h is the distribution of interest. Note the similarity with equation (7.9). The final term in brackets is a Jacobian arising from the change of variable from (θ, u) to (θ', u'). If $m = m'$, the move is a standard Metropolis-Hastings step. Appendix 7.B will summarize the implementation of this algorithm in our context.

The starting idea of the second step is that the marginal likelihood is the normalizing constant of the posterior density (equations 7.8 and 7.11). Thus, we can write:

$$f(y|m) = \frac{f(\theta|y, m)p(\theta|m)}{p(\theta|y, m)} \tag{7.12}$$

which is referred to as the *basic marginal likelihood identity*. Note that this identity is true for every θ: this means that we can estimate the marginal likelihood by finding an estimate of the posterior ordinate $p(\theta^*|y, m)$ in a single point θ^*. Anyway, for estimation efficiency, the point θ^* is taken to be a high-density point in the support of the posterior.

Chib and Jeliazkov (2001) propose an efficient method to produce the estimate $\bar{p}(\theta^*|y, m)$ using the output from a MCMC simulation with fixed m. After that, all (7.12) requires is the evaluation of likelihood and prior.

To illustrate the method, consider the simple case in which the posterior density is sampled in one block by the Metropolis-Hastings algorithm. The method is easily extended to our multiple parameter blocks case with a componentwise algorithm. Suppose the sample produced by the MCMC algorithm with fixed m is $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}\}$.

For notational convenience, the model index m will be suppressed in the rest of the section. Metropolis-Hastings method satisfies the detailed balance of a Markov chain (see e.g. Chib and Greenberg, 1995), thus we can write:

$$\alpha(\theta, \theta')q(\theta, \theta')p(\theta|y) = \alpha(\theta', \theta)q(\theta', \theta)p(\theta'|y)$$

for any point θ', where $\alpha()$ is the acceptance probability and $q()$ is the proposal density. Integrating both sides of this expression with respect to θ, we obtain:

$$\begin{aligned} p(\theta'|y) &= \frac{\int \alpha(\theta, \theta')q(\theta, \theta')p(\theta|y)d\theta}{\int \alpha(\theta', \theta)q(\theta', \theta)d\theta} \\ &= \frac{E_1[\alpha(\theta, \theta')q(\theta, \theta')]}{E_2[\alpha(\theta', \theta)]} \end{aligned}$$

where the expectation E_1 is with respect to the posterior $p(\theta|y)$ while the expectation E_2 is with respect to $q(\theta', \theta)$. The posterior ordinate is then estimated by the Monte Carlo estimator:

$$\bar{p}(\theta'|y) = \frac{N^{-1}\sum_{i=1}^{N} \alpha(\theta^{(i)}, \theta')q(\theta^{(i)}, \theta')}{J^{-1}\sum_{j=1}^{R} \alpha(\theta', \theta^{(j)})}$$

where $\{\theta^{(i)}\}$ are the samples from the posterior and $\{\theta^{(j)}\}$, for $j = 1, \dots, J$, are draws from $q(\theta', \theta)$, given the fixed value θ'.

In Appendix 7.C, this method is applied on the mixture of autoregressive components in order to calculate $p(k|y)$ (see also section 6.6.2 in the previous chapter).

7.6 Predictive distributions

Consider an unknown observable future variable y_{T+1}. A first kind of predictive distribution which can be computed is the one conditional on a given model m:

$$f(y_{T+1}|m, y) = \int f(y_{T+1}|\theta, m, y)p(\theta|m, y)d\theta \qquad (7.13)$$

where as usual $\theta = (w, \mu, \sigma^2, \phi)$, $f(y_{T+1}|\theta, m, y)$ is derived from (7.1) and $p(\theta|m, y)$ is the posterior distribution conditional on the model.

We estimate the (7.13) by the Monte Carlo estimator:

$$N^{-1}\sum_{i=1}^{N} f(y_{T+1}|\theta^{(i)}, m, y)$$

where $\theta^{(i)}$, for $i = 1, \dots, N$, are samples from the conditional posterior $p(\theta|m, y)$ and they are available from the MCMC output.

The unconditional predictive distribution takes into account the model uncertainty, since all the possible models are considered and they are weighted by their posterior probability (*Bayesian model averaging*). The unconditional predictive distribution is:

$$f(y_{T+1}|y) = \sum_{k=1}^{k_{max}} \left[\sum_{\rho} \int f(y_{T+1}|\theta, \rho, k, y) p(\theta, \rho|k, y) d\theta \right] p(k|y) \tag{7.14}$$

The corresponding Monte Carlo estimator of the (7.14) is:

$$\sum_{k=1}^{k_{max}} \left[N^{-1} \sum_{i=1}^{N} f(y_{T+1}|\theta^{(i)}, \rho^{(i)}, k, y) \right] p(k|y)$$

where $(\theta^{(i)}, \rho^{(i)})$, for $i = 1, \dots, N$, are samples from the conditional posterior $p(\theta, \rho^*|k^*, y)$ and they are available from the MCMC output.

In order to achieve punctual predictions, it is possible to calculate mean and variance of y_{T+1} with respect to the predictive densities by Monte Carlo estimators.

The computation of the m-step predictive distributions is less straightforward. Suppose to consider the two-step predictive distributions. A first method consists in using a punctual one-step forecast as if it is the true value of y_{T+1}. Alternatively, we can estimate the exact two-step predictive distribution trough Monte Carlo method, using samples from the one-step predictive distribution.

7.7 Return volatility

As we noticed in the introduction, our approach is based on considering an observable proxy variable for the return volatility. A standard choice is the daily squared returns which will be referred as volatility in the rest of the thesis.

We consider daily returns r_t on Nasdaq index from May 15, 2002 to May 15, 2003 (253 observations) and we take the logarithm of volatility: $y_t = ln(r_t^2)$. Time series plot (a) and histogram (b) of y_t are displayed in figure 7.1.

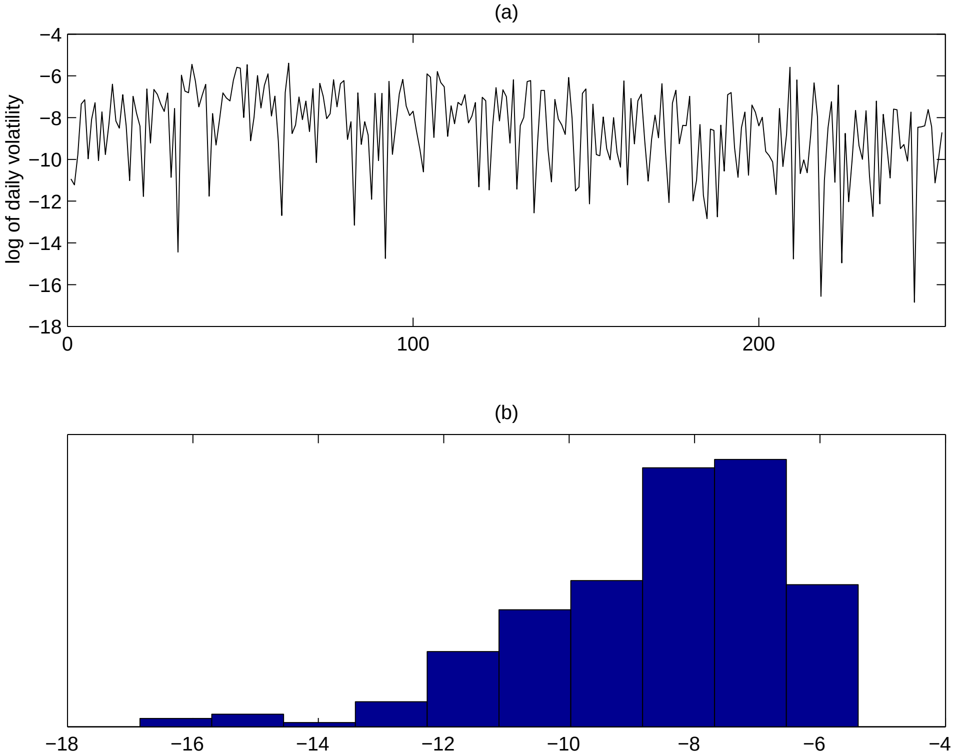

Figure 7.1: Plot (a) and histogram (b) of the logarithm of daily return volatility (Nasdaq index from May 15, 2002 to May 15, 2003)

Note that the empirical distribution shows a minor mode corresponding to low values of the volatility.

Some of the parameters involved are supposed to be fixed (section 7.3). We set $k_{max} = 5$ and $\rho_{max} = 9$. Higher values for these parameters will not change our results. Following Richardson and Green (1997), the prior for μ_j is chosen to be flat over an interval of variation of the observed data. Let $R = \max(y) - \min(y)$; we set $\mu_0 = \min(y) + R/2$ and $\tau = R$. We use the knowledge of the range of the data in setting the hyperparameters of σ^2. In particular, β will be a small multiple of R^{-2} ($\beta = 0.1 \cdot R^{-2}$) and $\alpha = 2$. Finally $\delta = 1$.

The number of iterations of the MCMC algorithm conditional on k was 50000, with a burn-in period of 10000. The successive reduced MCMC algorithms for the marginal likelihood were run with 15.000 iterations each one.

We believe that these numbers are sufficient to ensure the convergence of the algorithm.

Remember that prior distributions for the model indexes are uniform, as reported in equations (7.6) and (7.7). The resulting joint prior follows a parsimony principle, penalizing complex models with an high number of components:

$$p(\rho, k) = p(\rho|k)p(k) \propto \left(\frac{1}{\rho_{max}+1}\right)^k$$

where $\propto$ denotes "proportional to". Anyway it is easy to convert results to those corresponding to alternative priors by the identity:

$$\tilde{p}(\theta, m|y) \propto p(\theta, m|y)\frac{\tilde{p}(m)}{p(m)}$$

where $\tilde{p}(\cdot|y)$ is the posterior for a different prior $\tilde{p}$.

Table 7.1 shows the posterior probabilities of the number of mixture components k. The model with two components is largely preferred with a probability of 0.81. On the grounds of many different simulations, we noted a sensitivity of these results with respect to some hyperparameters. An obvious result is that smallest values of the prior variance τ increase the preference for the one-component model. Anyway, in all the simulations we performed the one-component model never had the highest posterior probability. $p(k|y)$ is also sensitive with respect to the prior on the variances σ^2. The posterior probabilities of the models with a number of components greater than 2 increases if the prior mean of σ^2 decreases. This is due to the fact that the smaller are the variance components, the higher is the number of components needed to fit the data.

k	$p(k\|y)$
1	0
2	0.81
3	0.18
4	0.01
5	0

Table 7.1: Posterior probabilities of the number of mixture components

Posterior probabilities of the orders of the autoregressive components are given in Table 7.2 for different values of k.

Information about the posterior distributions of some model parameters conditional on $k = 2$ are showed in table 7.3 (posterior means and variances) and in figure 7.2 (histograms of the posterior distributions).

$k = 1$		$k = 2$		$k = 3$	
(ρ_1)	$p(\rho_1 \vert k, y)$	(ρ_1, ρ_2)	$p(\rho_1, \rho_2 \vert k, y)$	(ρ_1, ρ_2, ρ_3)	$p(\rho_1, \rho_2, \rho_3 \vert k, y)$
(0)	0.85	(0,0)	0.46	(0,0,0)	0.09
(1)	0.09	(0,1)	0.03	(0,1,0)	0.04
(2)	0.03	(1,0)	0.24	(1,0,0)	0.08
(3)	0.02	(2,0)	0.06	(2,0,0)	0.05
		(3,0)	0.04	(3,0,0)	0.05
		(4,0)	0.03	(4,0,0)	0.04
		(5,0)	0.02	(5,0,0)	0.05
				(6,0,0)	0.04
				(7,0,0)	0.03
				(8,0,0)	0.03
				(9,0,0)	0.03

Table 7.2: Posterior probabilities (≥ 0.02) of autoregressive components' orders conditional on k.

	$\mathrm{E}(\cdot \vert k = 2, y)$	$\mathrm{Var}(\cdot \vert k = 2, y)$
w_1	0.31	0.01
w_2	0.69	0.01
μ_1	-11.07	4.97
μ_2	-7.69	0.04
σ_1^2	5.11	1.44
σ_2^2	1.65	0.13

Table 7.3: Posterior mean and variance of the model parameters conditional on $k = 2$

Tables 7.2 and 7.3 suggest several considerations. First of all, conditional on $k = 1$, order zero has a high posterior probability (0.85). In other words, the one-component model, that is a simple autoregressive model, does not seem able to capture volatility persistence. On the other hand, with $k \geq 2$,

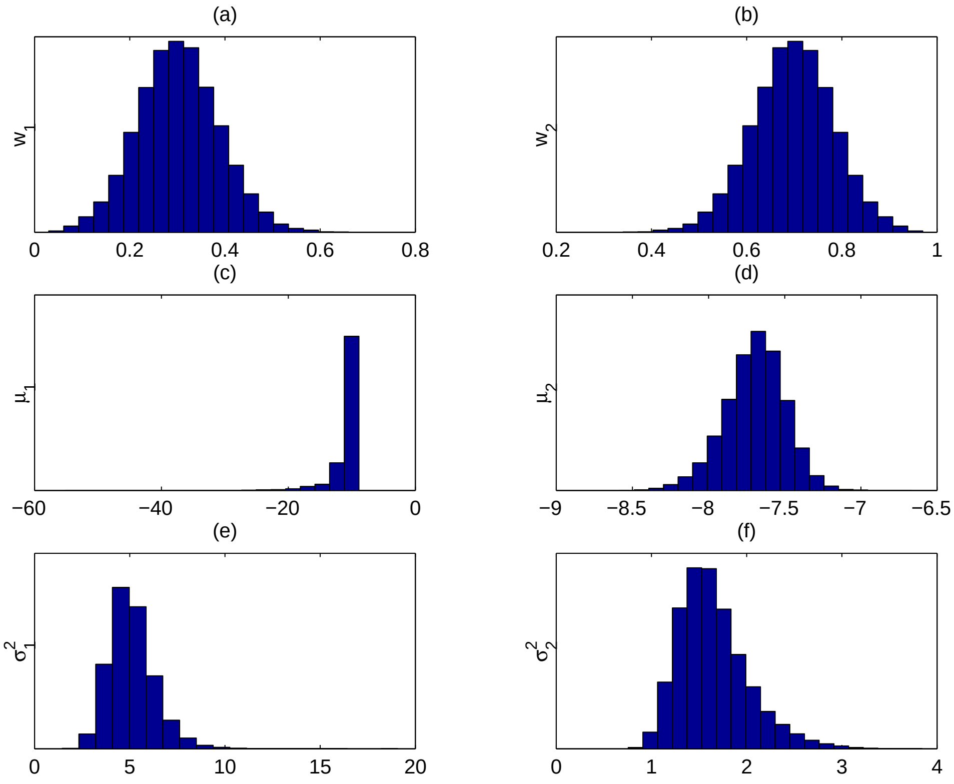

Figure 7.2: Posterior distributions of the weights [(a) and (b)], the means [(c) and (d)] and the variances [(e) and (f)] conditional on $k = 2$

the results show different levels of persistence according to a threshold effect. Consider the model formed by two components: the first one corresponds to low volatility values ($\mu_1 = -11.07$) and it shows higher persistence with respect to the second component. In fact, models of the form $(\rho_1, \rho_2) = (r, 0)$, with $r = 0, \ldots, 9$, have a total posterior probability of 0.883. Simulations with $k \geq 3$ lead to a similar structure, in which a first component formalizes low and persistent volatility values, while the remaining components are associated with zero autoregressive orders.

Table 7.4 shows posterior means and variances of the autoregressive coefficients ϕ conditional on $k = 2$ and on different values of ρ. Figure 7.3 shows the posterior distributions of ϕ conditional on $k = 2$ and on $\rho = (1, 0)$ and $\rho = (2, 0)$.

The one-step predictive distributions conditional on k are showed in figure

(ρ_1, ρ_2)	$\mathrm{E}(\phi_{1,\cdot}\|\rho, k, y)$	$\mathrm{E}(\phi_{2,\cdot}\|\rho, k, y)$	$\mathrm{Var}(\phi_{1,\cdot}\|\rho, k, y)$	$\mathrm{Var}(\phi_{2,\cdot}\|\rho, k, y)$
$(0,1)$	-	0.015	-	0.003
$(1,0)$	-0.25	-	0.042	-
$(2,0)$	-0.182	-	0.066	-
	0.111		0.045	
$(3,0)$	-0.161	-	0.068	-
	0.156		0.046	
	0.211		0.047	
$(4,0)$	-0.197	-	0.04	-
	0.126		0.034	
	0.208		0.039	
	0.278		0.044	
$(5,0)$	-0.205	-	0.042	-
	0.094		0.038	
	0.124		0.056	
	0.295		0.032	
	0.177		0.033	

Table 7.4: Posterior means and variances of the autoregressive coefficients ϕ conditional on $k = 2$ and on different values of ρ

7.4 . With a number of components greater than one, the predictive distribution is asymmetric. The unconditional one-step predictive distribution is given in figure 7.5.

An out of sample analysis is summarized in figure 7.6, where the one-step predictive distributions conditional on $k = 2$ for some values of t are displayed. Actual values of y_t are also shown at the times t-1, t and t+1.

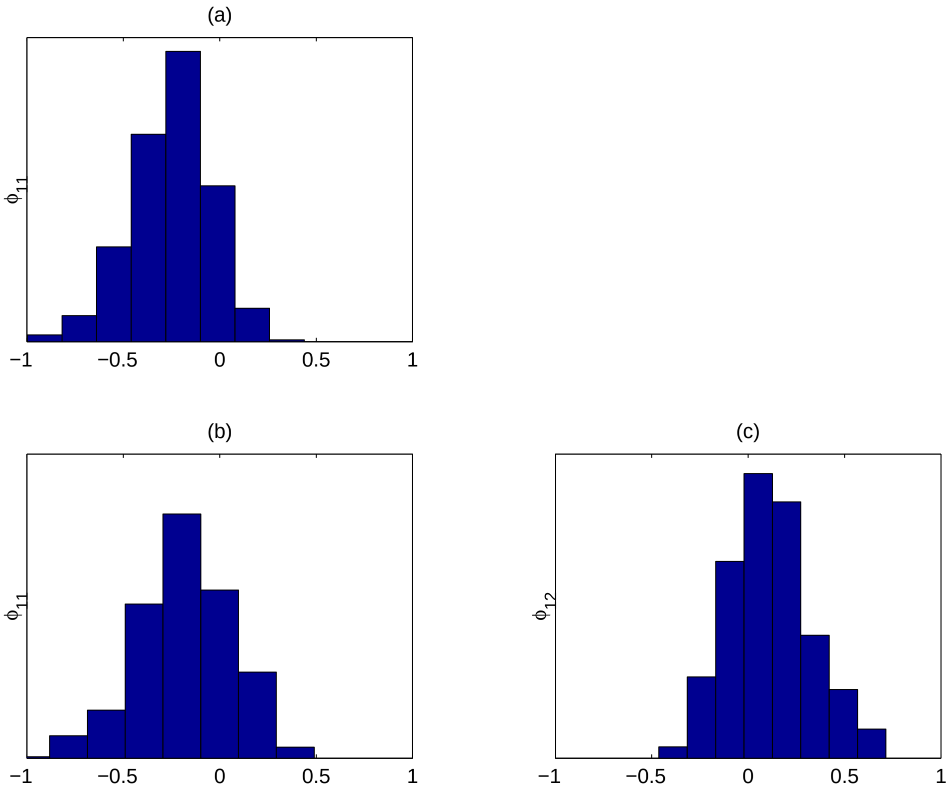

Figure 7.3: Posterior distributions of the autoregressive parameters conditional on $k = 2$ and $\rho = (1, 0)$ [(a)] and conditional on $\rho = (2, 0)$ [(b) and (c)]

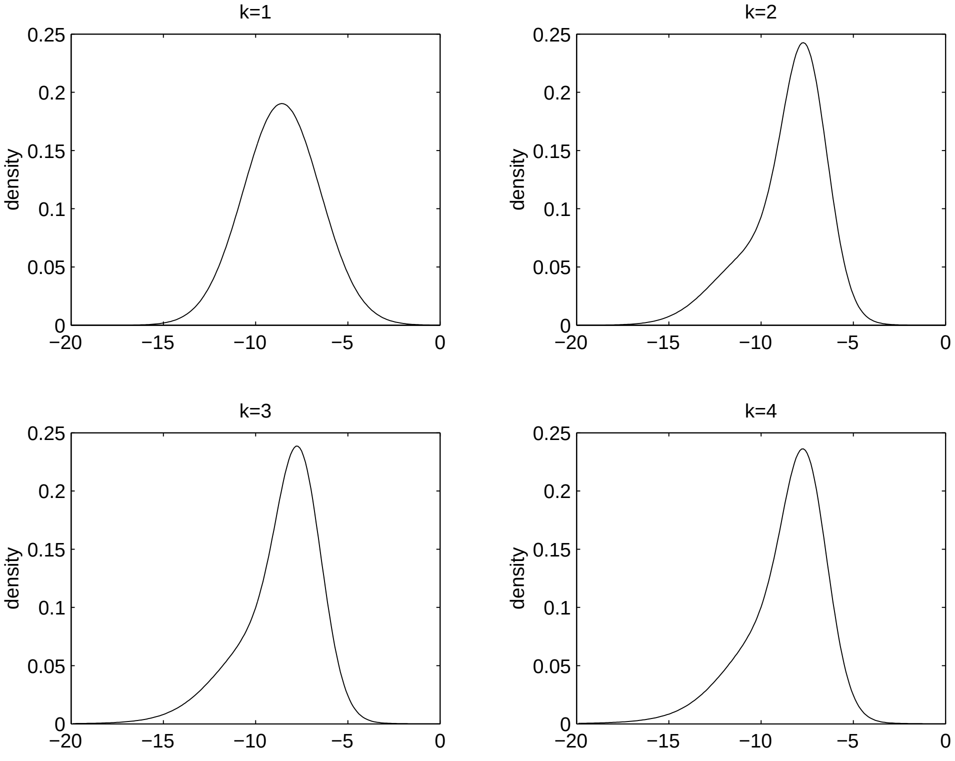

Figure 7.4: One-step predictive distributions conditional on the number of autoregressive components k

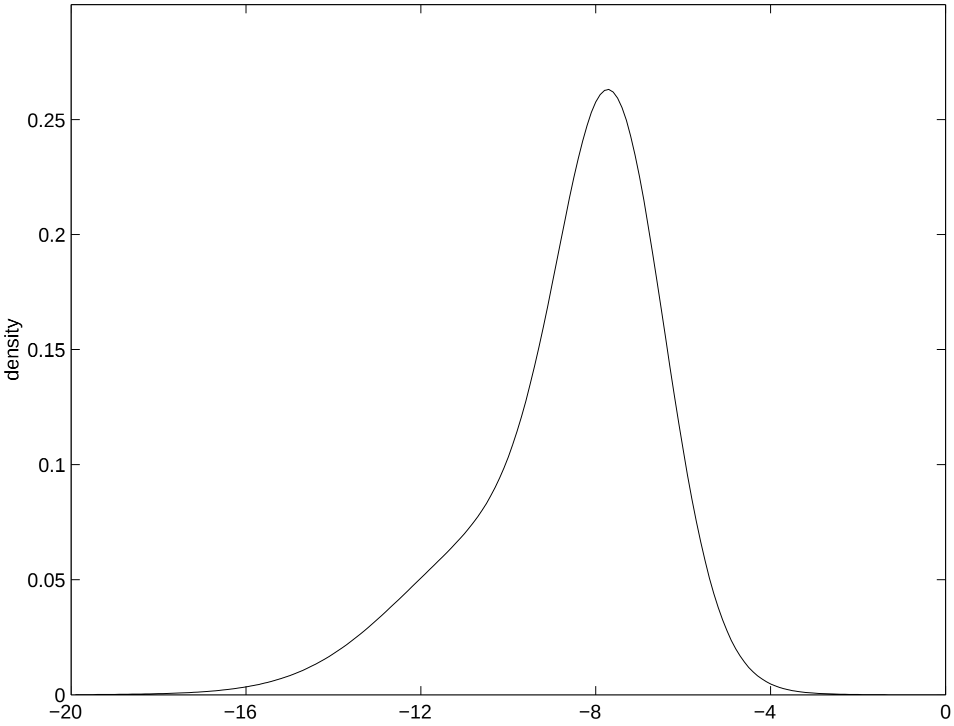

Figure 7.5: One-step unconditional predictive distribution

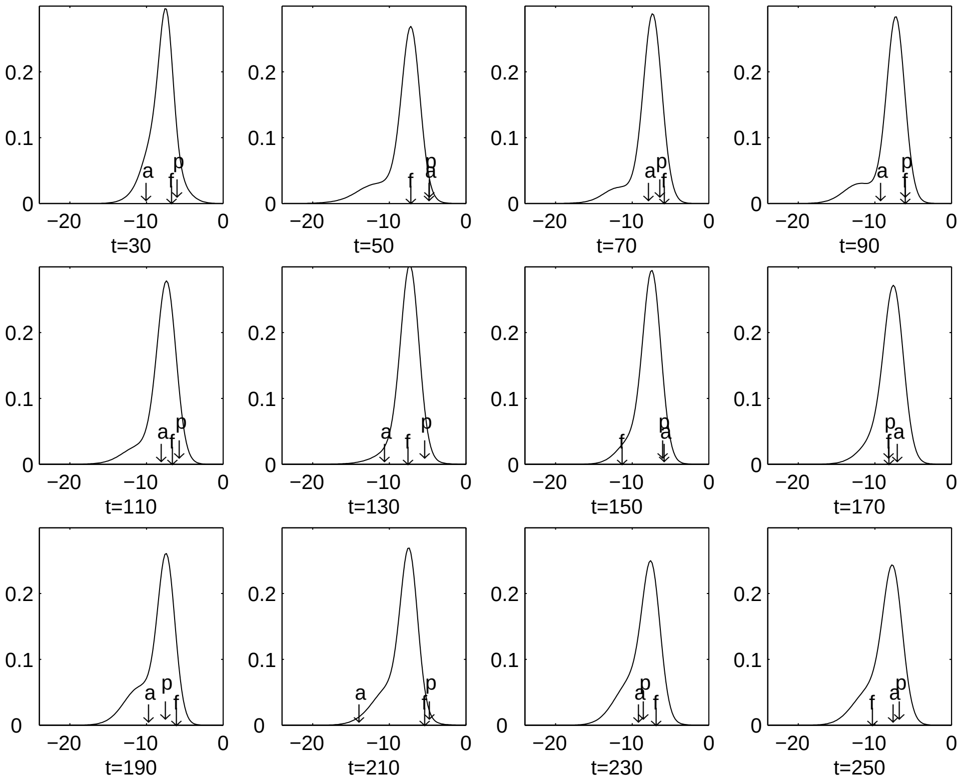

Figure 7.6: One-step predictive distributions conditional on $k = 2$ for different values of t. The actual values are shown (labels 'p', 'a' and 'f' stand for observations at times t-1, t and t+1 respectively)

7.8 Conclusions

In this chapter, we modelled an observable proxy of the return volatility of financial markets by the mixture of autoregressive components. Our results confirm some of the stylized facts about volatility which have been recognized in the literature, like the persistence and the non-symmetrical dependencies.

Moreover, our Bayesian perspective takes into account the important issue of model uncertainty, whose influence on financial models has been often ignored.

The mixture of autoregressive components can be extended in several ways. For instance, a mixture of autoregressive moving average (ARMA) components could be developed: the invertibility conditions of the moving average part of the model can be handled by a generalization of the reparametrization we used (Monaham, 1984).

The normality of the component densities can be also easily relaxed and different conditional distributions can be used. Note that considering other distributions implies opportune changes of the moves of the MCMC algorithm.

Another possible development is related to the stationarity conditions. In this work we consider local conditions, that is within each autoregressive components. This local stationarity seems to be a sufficient condition for the global stationarity but, as showed by Wong and Li (2000), a mixture of non-stationarity components can be stationary. A Bayesian analysis which takes into account global stationarity conditions could be done, nevertheless it does not appear straightforward.

Appendix 7.A

The MCMC algorithm for parameter estimation (section 7.4) consists in the following moves:

i. Updating the weights w
ii. Updating the means μ
iii. Updating the autoregressive coefficients ϕ
iv. Updating the variances σ^2
v. Updating the allocation variable z

Move i is a Gibbs sampler move. The proposed values for the weights w are drawn from the full conditional which is shown to be a Dirichlet density:

$$p(w|\mu, \sigma^2, \phi, z, y) = \mathrm{Di}(w|\delta_1 + n_1, \ldots, \delta_k + n_k) \tag{7.15}$$

where $n_j = \sum_{j=1}^{k} I_{(z_i=j)}$

Move ii is also a Gibbs type move, with full conditional for μ_j given by:

$$p(\mu_j|w, \sigma^2, \phi, z, y) = \mathrm{N}\left(\mu_j \left| \frac{n_j\, \bar{v}_j\, B\, \tau^2 + \sigma_j^2 \mu_0}{n_j\, B^2 \tau^2 + \sigma_j^2}, \frac{\sigma_j^2 \tau^2}{n_j\, B^2 \tau^2 + \sigma_j^2}\right.\right) \tag{7.16}$$

for $j = 1, \ldots, k$, where $\bar{v}_j = \frac{1}{n_j} \sum_{t:z_t=j} v_{tj}$, with $v_{tj} = y_t - \phi_{j,1} y_{t-1} - \cdots - \phi_{j,\rho_j} y_{t-\rho_j}$, and where $B = 1 - \phi_{j,1} - \cdots - \phi_{j,\rho_j}$.

Move iii updates ϕ through the partial autocorrelations π_j, for $j = 1, \ldots, k$ (section 7.3). The move is a Metropolis-Hastings type move. A candidate $\pi^*_{j,i}$ is generated by a normal density truncated in $(-1, +1)$ and centered in the current state of the chain $\pi_{j,i}$:

$$q(\pi_{j,i}, \pi^*_{j,i}) = \mathrm{N}_{(-1,+1)}(\pi^*_{i,j} | \pi_{j,i}, \sigma^2_q) \tag{7.17}$$

for $i = 1, \ldots, \rho_j$ and for $j = 1, \ldots, k$. The variance σ^2_q is chosen in order to obtain a satisfactory acceptance rate.

Let π^*_j be the proposal vector for the partial autocorrelations: using π^*_j, the corresponding parameters $\phi^*_j = (\phi^*_{j,1}, \ldots, \phi^*_{j,\rho_j})$ are derived through the transformation of section 7.3.

The acceptance probability is $\min(1, R)$ (equation 7.9), where R is given by:

$$\begin{aligned} R =& \exp\left\{ -\frac{1}{2\sigma_j^2} \sum_{t:z_t=j} \left[(y_t - \nu_{j,t}^*)^2 - (y_t - \nu_{j,t})^2 \right] \right\} \\ &\times \prod_{i=2}^{\rho_j} \frac{(\pi_{j,i}^* + 1)^{[(i-1)/2]}(1 - \pi_{j,i}^*)^{[i/2]}}{(\pi_{j,i} + 1)^{[(i-1)/2]}(1 - \pi_{j,i})^{[i/2]}} \\ &\times \prod_{i=1}^{\rho_j} \frac{F_N(1|\pi_{j,i}, \sigma_q^2) - F_N(-1|\pi_{j,i}, \sigma_q^2)}{F_N(1|\pi_{j,i}^*, \sigma_q^2) - F_N(-1|\pi_{j,i}^*, \sigma_q^2)} \end{aligned} \tag{7.18}$$

where $\nu_{j,t}^* = \mu_j + \phi_{1,j}^*(y_{t-1} - \mu_j) + \cdots + \phi_{\rho_j,j}^*(y_{t-\rho_j} - \mu_j)$ and where F_N is the cumulative distribution function of the normal distribution. The first two lines of the (7.18) correspond to the posterior ratio, expressed as likelihood ratio (first line) times prior ratio (second line). The third line is the proposal ratio.

The Gibbs move *iv* proposes a candidate value for σ^2 by the full conditional:

$$p(\sigma_j^2|\beta, w, \mu, \phi, z, y) = \text{Ig}\left(\sigma_j^2 \middle| \alpha + \frac{1}{2} n_j, \beta + \frac{1}{2} \sum_{t:z_t=j} (y_t - \nu_{j,t})^2 \right) \tag{7.19}$$

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

Finally, move v is a Gibbs move with full conditional:

$$p(z_t = j|w, \mu, \sigma^2, \phi, y) \propto \frac{w_j}{\sigma_j} \exp\left\{ -\frac{1}{2} \frac{(y_t - \nu_{j,t})^2}{\sigma_j^2} \right\}$$

for $t = 1, \ldots, T$.

Appendix 7.B

Order of the autoregressive components

In order to obtain the marginal posterior of the order of the autoregressive components $p(\rho|k, y)$, the set of moves described in appendix A is augmented by a sixth move based on a reversible jump mechanism (section 7.5).

The move starts by selecting a component, say j^*, randomly chosen in $\{1, \ldots, k\}$. The order of this component ρ_{j^*} increases by one with probability $b(\rho_{j^*})$ and decreases by one with probability $d(\rho_{j^*})$, where $b(\rho_j) = 1 - d(\rho_j)$, for $j = 1, \ldots, k$, $d(1) = 0$ and $b(\rho_{max}) = 0$. Formally, the proposal order $\rho^*_{j^*}$ is constructed as follow:

$$\rho^*_{j^*} = \begin{cases} \rho_{j^*} - 1 \ , & \text{with prob. } d(\rho_j) \\ \rho_{j^*} + 1 \ , & \text{with prob. } b(\rho_j) \end{cases}$$

It is now necessary to change the partial autocorrelation coefficients. Following Barbieri and O'Hagan (1997), if the order is decreased, the last partial autocorrelation is simply discarded. Otherwise, we need a new parameter $\pi^*_{\rho^*_{j^*},j^*}$, which is generated from the beta prior.

That is, letting $\pi^*_{j^*}$ be the proposal vector of the partial autocorrelations:

- If $\rho^*_{j^*} = \rho_{j^*} - 1, \quad \pi^*_{j^*} = (\pi_{1,j^*}, \ldots, \pi_{\rho^*_{j^*},j^*})$
- If $\rho^*_{j^*} = \rho_{j^*} + 1, \quad \pi^*_{j^*} = (\pi_{1,j^*}, \ldots, \pi_{\rho_{j^*},j^*}, \pi^*_{\rho^*_{j^*},j^*})$

with $\pi^*_{\rho^*_{j^*},j^*} \sim \text{Gb}\left(\pi_{i,j}| \left[\frac{i+1}{2}\right], \left[\frac{i}{2}\right] + 1\right)$.

Note that in both cases all the autoregressive parameters are updated because of the reparametrization of section (7.3).

If $\rho^*_{j^*} = \rho_{j^*} - 1$, the acceptance probability ratio is $\min(1, R)$, where R is given by:

$$R = \exp\left\{-\frac{1}{2\sigma^2_j} \sum_{t:z_t=j} \left[(y_t - \nu_{j,t})^2 - (y_t - \nu^*_{j,t})^2\right]\right\} \frac{b(\rho^*_{j^*})}{d(\rho_{j^*})} \tag{7.20}$$

where $\nu^*_{j^*,t} = \mu_{j^*} + \phi^*_{1,j^*}(y_{t-1} - \mu_{j^*}) + \cdots + \phi^*_{\rho_{j^*},j^*}(y_{t-\rho_{j^*}} - \mu_{j^*})$.

On the other hand, if $\rho^*_{j^*} = \rho_{j^*} + 1$:

$$R = \exp\left\{-\frac{1}{2\sigma^2_j} \sum_{t:z_t=j} \left[(y_t - \nu_{j,t})^2 - (y_t - \nu^*_{j,t})^2\right]\right\} \frac{d(\rho^*_{j^*})}{b(\rho_{j^*})} \tag{7.21}$$

The equations (7.20) and (7.21) are quite simple because of some cancellations between prior and proposal ratios. Furthermore, the jacobian is one because the matrix of derivatives of the transformation g (section 7.5) is the identity matrix.

$p(\rho|k, y)$ is simply estimated by the proportions of every possible value for ρ in the sample obtained by the previous complete MCMC algorithm.

Appendix 7.C

Number of the autoregressive components.

Through Bayes' theorem, the marginal posterior distribution of k is:

$$p(k|y) \propto p(k)f(y|k)$$

where $p(k)$ is the prior on k and $f(y|k)$ is the marginal likelihood:

$$f(y|k) = \sum_{\rho} \int L(\theta, \rho, k)p(\theta, \rho|k)\, d\theta \tag{7.22}$$

with $\theta = (w, \mu, \sigma^2, \phi)$. Suppressing for notational convenience the model index k, we write the marginal likelihood (7.22) as:

$$\begin{aligned} f(y) &= \frac{L(\theta^*, \rho^*)p(\theta^*, \rho^*)}{p(\theta^*, \rho^*|y)} \\ &= \frac{L(\theta^*, \rho^*)p(\theta^*|\rho^*)p(\rho^*)}{p(\theta^*|\rho^*, y)p(\rho^*|y)} \end{aligned} \tag{7.23}$$

for a fixed point (θ^*, ρ^*). Note that what we only need of the (7.23) is $p(\theta^*|\rho^*, y)$: we calculate the corresponding estimate $\bar{p}(\theta^*|\rho^*, y)$ by the method of section 7.5.

First of all, $\bar{p}(\theta^*|\rho^*, y)$ is factorized as:

$$\begin{aligned} \bar{p}(\theta^*|\rho^*{,}y) = {} & \bar{p}(\pi^*|\rho^*, y) \times \bar{p}(\mu^*|\pi^*, \rho^*, y) \times \\ & \times \bar{p}(\sigma^{2*}|\mu^*, \pi^*, \rho^*, y) \times \bar{p}(w^*|\sigma^{2*}, \mu^*, \pi^*, \rho^*, y) \end{aligned} \tag{7.24}$$

Suppose to have a sample $\{\theta^{(i)}, z^{(i)}\}$, for $i = 1, \dots, N_1$, from the MCMC algorithm for a given ρ^* (i.e. a sample from $p(\theta|\rho^*)$). Let $\eta_{j-1} = (\rho, \pi_1, \dots, \pi_{j-1})$ and $\eta^{j+1} = (\pi_{j+1}, \dots, \pi_k, \mu, \sigma^2, w)$. The terms of the (7.24) are estimated by the following steps:

1. Sample $\{\tilde{\eta}^{j+1,(i)}, \tilde{z}^{(i)}\}$, for $i = 1, \dots, N_{j+1}$, from a reduced MCMC algorithm with distribution of interest $p(\eta^{j+1}, z|\eta_j^*, y)$. Also draw $\tilde{\pi}_j^{(i)}$ from $q_p(\pi_j^*, \pi_j) = \prod_{s=1}^{\rho_j} q(\pi_{s,j}^*, \pi_{s,j})$, where $q(.,.)$ is the proposal (7.17).

   Set:

$$\bar{p}(\pi_j^*|\rho^*, \pi_1^*, \dots, \pi_{j-1}^*) = \frac{N_j^{-1} \sum_{i=1}^{N_j} \alpha(\pi_j^{(i)}, \pi_j^*) q_p(\pi_j^{(i)}, \pi_j^*)}{N_{j+1}^{-1} \sum_{i=1}^{N_{j+1}} \alpha(\pi_j^*, \tilde{\pi}_j^{(i)})}$$

where $\alpha(.,.) = \min(1, R)$ with R defined in equation (7.18).

Set $\eta^{j+1,(i)} = \tilde{\eta}^{j+1,(i)}$ and $z^{(i)} = \tilde{z}^{(i)}$, for $i = 1, \dots, N_{j+1}$.

Repeat this step for $j = 1, \dots, k$ and finally set:

$$\bar{p}(\pi^*|\rho^*, y) = \prod_{j=1}^{k} \bar{p}(\pi_j^*|\rho^*, \pi_1^*, \dots, \pi_{j-1}^*)$$

2. The second term is:

$$\bar{p}(\mu^*|\pi^*, \rho^*, y) = N_{k+1}^{-1} \sum_{i=1}^{N_{k+1}} \prod_{j=1}^{k} p(\mu_j^*|\pi^*, \sigma^{2(i)}, z^{(i)}, \rho^*, y)$$

   where $(\sigma^{2(i)}, z^{(i)})$ are draws from the last iteration of the previous step (thus they are marginally from $p(\sigma^2, w, z|\pi^*, \rho^*, y)$) and $p(\mu_j^*|\pi^*, \sigma^{2(i)}, z^{(i)}, \rho^*, y)$ is given by equation(7.16).

3. Sample $\{\sigma^{2(s)}, w^{(s)}, z^{(s)}\}$, for $s = 1, \dots, S$, from a reduced MCMC algorithm with distribution of interest $p(\sigma^2, w, z|\pi^*, \mu^*, \rho^*, y)$ and set:

$$\bar{p}(\sigma^{2*}|\pi^*, \mu^*, \rho^*, y) = S^{-1} \sum_{s=1}^{S} \prod_{j=1}^{k} p(\sigma_j^{2*}|\pi^*, \mu^*, z^{(s)}, \rho^*, y)$$

   where $p(\sigma_j^{2*}|\pi^*, \mu^*, z^{(i)}, \rho^*, y)$ is given by equation (7.19).

4. Sample $\{w^{(v)}, z^{(v)}\}$, for $v = 1, \dots, V$, from a reduced MCMC algorithm with distribution of interest $p(w, z|\pi^*, \mu^*, \sigma^{2*}, \rho^*, y)$ and set:

$$\bar{p}(w^*|\pi^*, \mu^*, \sigma^{2*}, \rho^*, y) = V^{-1} \sum_{v=1}^{V} p(w^*|z^{(v)}, \rho^*, y)$$

   where $p(w^*|z^{(v)}, \rho^*, y)$ is given by equation (7.15).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium in Information Theory* (eds. B. N. Petrov and F. Csaki), 267-281. Budapest: Akademiaia Kiado.

Albert, J. and Chib, S. (1993). Bayesian inference via Gibbs sampling of autoregressive time-series subject to Markov mean and variance shift . *Journal of Business and Economic Statistics*, **11**, 1-15.

Anderson, T.W. (1970). *The statistical analysis of time series.* New York: Wiley.

Barberis, N. (2000). Investing for the long run when returns are predictable. *Journal of Finance*, **55**, 225-264.

Barbieri, M.M. and O'Hagan, A. (1997). A reversible jump MCMC sampler for Bayesian analysis of ARMA time series. *Manuscript.*

Barndorff-Nielsen, O. and Schou, G. (1973). On the parametrization of autoregressive models by partial autocorrelation. *Journal of Multivariate Analysis*, **3**, 408-419.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian theory.* New York: Wiley.

Besag, J. (2000). *Markov chain Monte Carlo for statistical inference.* Working paper, Center for Statistics and the social sciences, University of Washington, USA.

Bollerslev, T., Chou, R.Y. and Kroner, K.P. (1992). ARCH modeling in finance: a review of the theory and empirical evidence. *Journal of Econometrics*, **52**, 5-59.

Bollerslev, T. and Mikkelsen, H.O.A. (1996). Modelling and pricing long memory in stock market volatility. *Journal of Econometric*, **73**, 151-184.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). *Time series analysis: forecasting and control*. New Jersey: Prentice-Hall.

Broemeling, L.D. and Shaaraway, S. (1988). Time series: a Bayesian analysis in the time domain. In *Bayesian analysis of time series and dynamic models*, 1-21, ed. J.C. Spall, New York: Marcel Dekker.

Brooks, S.P., Giudici, P. and Roberts, G.O. (2003). Efficient construction of reversible jump MCMC proposal distributions (with discussion). *Journal of the Royal Statistical Society Series B*, **65**, 3-55.

Cao, C.Q. and Tsay, R.S. (1992). Nonlinear time-series analysis of stock volatilities. *Journal of Applied Econometrics*, December, Supplement, **1**, 165-185.

Cappé, O., Robert, C.P. and Rydn, T. (2001). Discrete time and continuous time jump processes with application to hidden Markov models. Paris: Tech. report, CREST, INSEE.

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.

Casella G. and George, I.E. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167-174.

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

Chib S. and Greenberg E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**, 327-335.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **96**, 270-281.

Consonni, G. and Veronese, P. (1995). A bayesian method for combining results from several binomial experiments. *Journal of the American Statistical Association*, **90**, 935-944.

Cowles, M.K. e Carlin, B.P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883-904.

Cox, D.R. and Miller, H.D. (1965). *The theory of stochastic processes.* London: Chapman & Hall.

Cremers, K.J.M. (2002). Stock return predictability: a Bayesian model selection perspective. *Review of Financial studies*, **15**, 1223-1249.

Dellaportas, P., Forster, J.J. and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and computing*, **12**, 27-36.

Diebolt, J. and Robert, C.P. (1990). Estimation of finite mixture distributions through Bayesian sampling (Parts I and II). *Rapports Techniques*, 109, 100, LSTA, Université Paris 6.

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.

Ding, Z., Granger, J.W.J and Engle, R.E. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, **1**, 83-106.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B*, **57**, 45-98.

Engle, R.F. (1990). Discussion: Stock market volatility and the crash of 87. *The Review of Financial Studies*, **3**, 103-106.

Engle, C and Kim, C.J. (1999). The long run U.S./U.K. real exchange rate. *Journal of money, credit and banking*, **31**, 335-356.

Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, **10**, 254-72.

Everitt, B.S. and Hand, D.J. (1981). *Finite mixture distributions.* London: Chapman & Hall.

French, K.R., Schwert, G.W. and Stambaugh, R.F. (1987). Expected stock returns and volatility. *Journal of financial economics*, **19**, 3-29.

Fruhwirth-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.

Gamerman, D. (1997). *Markov Chain Monte Carlo.* London: Chapman & Hall.

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

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society Series B*, **56**, 501-514.

Gelman, A., Carlin J.B., Stern H.S. and Rubin D.B. (1995). *Bayesian data analysis.* New York: Chapman & Hall.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Institute of Electrical and Electronics Engineers, Transactions on pattern analysis and machine intelligence*, **6**, 721-741

George, E.I. and McCulloch, R.E. (1996). Stochastic search variable selection. In *Markov chain Monte Carlo in practice.* 203-214. London: Chapman & Hall.

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

Ghysels, E., Harvey, A. and Renault, E. (1996). Stochastic volatility. In *Handbook of statistics: statistical methods in finance*, **14**, 119-191, Amsterdam: Elsevier Science.

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

Glosten, L.R., Jagannathan, R. and Runke, D.E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, **48**, 1779-1801.

Good, I. J. (1983). *Good thinking: the foundations of probability and its applications.* Minneapolis: University of Minnesota Press.

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

Greene, W.H. (2000). *Econometric analysis.* New York: Prentice Hall.

Grenander, U. and Miller, M.I. (1994). Representation of knowledge in complex system. *Journal of the Royal Statistical Society Series B*, **56**, 549-603.

Gruet, M.A. and Robert, C.P. (1997). Comment on "On Bayesian analysis of mixtures with an Unknown number of components". *Journal of the Royal Statistical Society Series B*, **59**, 777.

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357-384.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chain and their applications. *Biometrika*, **57**, 97-109.

Hsieh, D.A. (1991). Chaos and nonlinear dynamics: application to financial markets. *Journal of Finance*, **46**, 1839-1877.

Jones, C., Lamont, O. and Lumsdaine, R. (1998). Macroeconomic news and bond market volatility. *Journal of Financial Economics*, **47**, 315-337.

Jones, M. C. (1987), Randomly choosing parameters from the stationarity and invertibility region of autoregressive-moving average models. *Applied Statistics*, **36**, 134-138.

Leamer, E. E. (1983), Let's take the con out of econometrics. *American Economic Review*, **73**, 31-43.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society Series B*, **34**, 1-41.

Lindsay, B.G. (1995). *Mixture models: theory, geometry and applications.* NSF-CBMS regional conference series in Probability and Statistics, **5**. Hayward: Institute of Mathematical Statistics.

Longin, F.M. (1997). The threshold effect in expected volatility: a model based on asymmetric information. *The Review of Financial Studies*, **10**, 837-869.

MacKinley, A.C. and Pastor, L. (2000). Asset pricing models: implications for expected returns and portfolio selection. *Review of Financial studies*, **13**, 883-916.

Malec, D. e Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, **79**, 593-601.

McCulloch, R.E. and Rossi, P.E. (1991). A Bayesian approach to testing the arbitrage pricing theory. *Journal of Econometrics*, **49**, 141-168.

McCulloch, R.E. and Tsay, R.S. (1994). Statistical analysis of economic time series via Markov switching models. *Journal of Time Series Analysis*, **15**, 523-539.

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

Mengersen, K.L. and Robert, C.P. (1995). *Testing for mixtures via entropy distance and Gibbs sampling.* In *Bayesian Statistics 5*, eds Berger, J.O., Bernardo, J.M., Dawid, A.P., Lindlye, D.V. and Smith, A.F.M. Oxford: Oxford University Press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, W. N. and Teller, A. H. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1091.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chain and stochastic stability.* London: Springer-Verlag.

Monahan, J.F. (1984). A note on enforcing stationarity in autoregressive-moving average models. *Biometrika*, **71**, 403-404.

Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, **59**, 347-370.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, **8**, 343-366.

Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B*, **56**, 3-48.

Pastor, L. (2000). Portfolio selection and asset pricing models. *Journal of Finance*, **55**, 179-224.

Pastor, L. and Stambaugh, R.F. (2000). Comparing asset pricing models: an investment perspective. *Journal of Financial Economics*, **56**, 353-381.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society Series A*, **185**, 71-110.

Phillips, D.B. and Smith, A.F.M. (1996). Bayesian model comparison via jump diffusion. In *Markov chain Monte Carlo in practice.* 203-214. London: Chapman & Hall.

Poon, S.H. and Granger, C.W.J. (2003). Forecasting volatility in financial markets: a review. *Journal of Economic Literature*, **41**, 478-539.

Poterba, J.M. and Summers, L.H. (1986), The persistence of volatility and stock markets fluctuations. *American Economic Review*, **76**, 1142-1151.

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**, 731-759.

Ripley, B. (1977). Modeling of spatial pattern. *Journal of the Royal Statistical Society Series B*, **39**, 172-192.

Robert, C.P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical science*, **10**, 231-253.

Robert, C.P. (2001). *The bayesian choice.* New York: Springer-Verlag.

Robert, C.P. and Casella, G. (1999). *Monte Carlo statistical methods.* New York: Springer-Verlag.

Ross, S.M. (1996). *Stochastic processes.* New York: John Wiley & Sons.

Sampietro S. and Veronese P. (1998). An MCMC algorithm for Bayesian analysis of hierarchical partition models. *Journal of the Italian statistical society*, **7**, 2, 209-220.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Schwert, G.W. (1987). Effects of model specification on tests for unit roots in macroeconomic data. *Journal of Monetary Economics*, **20**, 73-103.

Stephens, M. (2000). Bayesian methods for mixtures of normal distributions - An alternative to reversible jump methods. *Annals of Statistics*, **28**, 40-74.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701-1762.

Titterington D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical analysis of finite mixture distributions.* New York: Wiley.

West, M. and Harrison, J. (1989). *Bayesian forecasting and dynamic models.* New York: Springer-Verlag.

Wong, C. S. and Li, W. K. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society Series B*, **62**, 95-115.

Zellner, A. (1971). *An introduction to Bayesian inference and econometrics*, New York: Wiley.