MIXTURE AUTOREGRESSIVE MODELS WITH APPLICATIONS TO HETEROSKEDASTIC TIME SERIES

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2021

Davide Ravagli

School of Natural Sciences
Department of Mathematics

Contents

List of Tables

List of Figures

Abstract

This thesis presents advances in theory and applications of mixture autoregressive (MAR) models in both Bayesian and frequentist frameworks.

We improve the Bayesian analysis of mixture autoregressive models in the case of Gaussian components, by use of a sampling algorithm that allows to sample from the complete space of the posterior distribution of the parameters. In addition, we introduce a relabelling algorithm to deal with label switching, and propose density forecasts based on simulated Bayesian samples.

We generalise the methodology to MAR models with Student-t mixture components, which includes Gaussian MAR as a limit case. We tackle the challenge of treating the number of degrees of freedom of the Student-t distribution as parameters whose posterior distribution has to be estimated.

We propose using mixture vector autoregressive (MVAR) models for optimisation of portfolios of assets. The properties of MVAR models, combined with modern portfolio theory, allow in fact to analytically derive predictive distributions for portfolio returns at any time horizon. We also compare forecasting performance of MVAR models with other commonly used models in this context.

We introduce an uncorrelated version of MAR models. By applying a set of linear constraints on the autoregressive parameters, the resulting model represents a direct alternative to GARCH models, as they both assume an uncorrelated but dependent structure for the data. We also propose an application of the uncorrelated MAR to residuals of an econometric model.

All the data analysis is implemented in R, the majority of which is available in the package **mixAR**.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property

and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on presentation of Theses

Acknowledgements

I would like to thank my supervisor Dr. Georgi Boshnakov for his guidance throughout the development of this Ph.D. project. His encouragments, availability and willingness to share his knowledge has helped and shaped my improvements and results in this process.

Thanks go to the University of Manchester's Department of Mathematics for providing financial support during my studies, as well as all fellows Ph.D. students and staff in the department, who welcomed from the start, and have been a priceless source of support throughout.

Finally, a special mention goes to my family, who stand by my side unconditionally. They supported and encouraged me during hard times, and I thank them for giving me the strength to be where I am today.

Last, but not least, thanks go to Chris, Emma, Gustavo and Rhys, for being the most outstanding company and housemates, people who I could always rely on through these years.

Chapter 1

Introduction

Statistician strive to keep models as simple as possible. We do so by attempting to model our data under simple assumptions, such as thinking that all observations in a dataset come from a single, homogeneous population. Undoubtedly, simple models have some advantages, in that they have few parameters that can be estimated accurately and directly, with estimators that are simple to derive in closed form and which properties are well understood. In addition, the interpretation of the fitted model is straightforward in such cases. However, the assumption of data being generated from a single, simple distribution is often utopian in the real world. Data is instead asymmetric, multimodal and variable, as if it was generated from several heterogeneuous subgroups which exist, but cannot be observed.

The choice of a mixture of distributions comes naturally when the data shows signs of heterogeneity, multimodality or skewness. Finite mixture models (McLachlan and Peel, 2000) provide a theoretical base to approach a wide range of application in statistics, when data present this characteristic. Mixture autoregressive (MAR) models, first introduced by Wong and Li (2000), belong to this class. The flexibility of a mixture of distributions makes mixture autoregressive models attractive and suitable for nonlinear, nonstationary time series. Furthermore, thanks to conditional distributions that

depend on the recent past of the process, mixture autoregressive models can capture heteroskedasticity, multimodality and skewness in the data. Furthermore, they can inherently account for uncorrelatedness in the series of interest. This makes them particularly interesting for modelling financial time series, which is also the main focus of this thesis. However, their application is not limited to the financial field, as they have been used, for instance, in medicine or with environamental time series data.

Mixture autoregressive models may be seen as a particular case of regime switching models (or Markov switching models, Goldfeld and Quandt, 1973; Hamilton, 1989), a class of probabilistic models widely used in financial time series. In brief, regime switching models assume the underpinning presence of two or more processes, or regimes, that govern the data generating process, which alternate over time with probabilities that evolve according to a first-order discrete Markov chain. Thanks to this assumption, they can identify abrupt changes that may occur in the mean, variance, or other features of a time series of interest. In general, regime switching models are aimed to identify unobserved heterogeneity and overdispersion of a phenomenon over time, usually with a low number of change points. For example, they may be used to model periods of fast growth and low growth in the economy. In addition to this, mixture autoregressive models add the possibility of detecting outliers accurately by allowing for several change points in time. This makes MAR models attractive towards financial returns, in which sudden bursts may occur at any time point. However, a mixture of distributions can in practice approximate any distribution, a property that makes MAR models potentially useful to any time series for which the Normality assumption appears to be violated.

1.1 Main contributions

The main contributions of this thesis may find application in various fields within the framework of time series. Throughout this thesis, our main focus is in financial time series, which features very well serve the purpose of illustrating the features of mixture autoregressive models. However, they may be effectively generalised and applied to different types of time series data (as shown in the example of Canadian Lynx data in Chapter 3).

For the purpose of modelling financial returns, time series literature suggests that the assumption of a Normal distribution for the innovations is typically not appropriate. This type of data often presents heavy tails, meaning "extreme" observations are more likely to occur than those suggested by Normal distribution. Mixtures of distributions provide a flexible way to account for heavy-tailed or skewed data, making them a suitable modelling option in this scenario.

- The Bayesian analysis of mixture autoregressive models described in Chapter 3 is an improvement of the previous analyses by Sampietro (2006) and Hossain (2012), in that, unlike these and other existing literature, our method is able to cover the complete parameter space of the model. The issue of label switching, quite prominent in Hossain (2012), is also dealt with, using a relabelling algorithm a posteriori. The chapter displays a direct comparison between these methodologies, and explains in detail the shortcomings of existing methods in estimating parameter posterior distributions, which we successfully overcome. This analysis is also available as **arXiv** preprint (Ravagli and Boshnakov, 2020a).

  The main contribution of this chapter is to incorporate a check on whether or not a candidate set of parameters satisfies the stability region, without the need to truncate prior distributions, an operation that could result in a significant loss of

information. In addition, the use of a relabelling algorithm a posteriori solves the issues created by identifiability constraints, which sometimes affect convergence of the Markov Chain to its stationary distribution.

- Mixtures of Normal distributions can, in principle, approximate distributions with heavy tails. However, the number of components required to achieve that might be very large. Therefore, we present an extension and generalisation of the Bayesian methodology for MAR models (Chapter 4). This time, the assumption of Gaussian components is replaced by that of Student-t components (Wong et al., 2009). Although similar in concept, this extension requires introduction of an additional set of latent variables, and brings up more computational difficulties due to larger variability. On the other hand, Wong et al. (2009) argues that, because the tails of the Student-t distribution can be adjusted through the degrees of freedom, this mixture has a higher level of flexibility compared to that of the Gaussian model. The MAR model with Gaussian component is a particular case of the more general MAR with Student-t components, in which the degrees of freedom for each mixture component are sufficiently large that the distribution is approximately Gaussian.

  There are no cases in the literature of a Bayesian analysis of Student-t mixture autoregressive models, which makes the analysis presented in Chapter 4 a novel and original contribution.

- Another important contribution is in modelling the degrees of freedom of the Student-t distribution in the context of MAR models. The challenge here is in choosing suitable prior distributions for the degrees of freedom, as they are known to highly influence the posterior in this context (as seen in Geweke, 1994). We propose a different approach to that of Geweke (1994) for choosing prior distributions for the degrees of freedom, which allows to incorporate

prior information on the parameters more effectively and efficiently, and recur to a Metropolis-Hastings algorithm for simulation from the respective posterior distributions.

- We introduce some useful novel application for MAR models (Chapter 5). First, we give an overview of mixture vector autoregressive models (MVAR), the multivariate version of MAR, and derive analyical expressions for multi-step predictive densities. Secondly, we propose an innovative application, which consists in combining MVAR models with modern portfolio theory (Markowitz, 1952) for portfolio optimisation. Not only the proposed methodology will derive predictive distributions on portfolio returns analytically, but also allows to estimate the risk associated with a portfolio of assets. The methodology finds its ground of comparison among multivariate GARCH models, as well as other conditional correlation models, well established techniques for estimation and prediction of multivariate financial time series data. This analysis is also available as arXiv preprint (Ravagli and Boshnakov, 2020b).

- Finally, we propose a constrained version of MAR models (Chapter 6). A set of linear constraints can be applied on the autoregressive parameters of a MAR model, which ensure uncorrelatedness while reducing the number of parameters to be estimated. Such constraints are particularly useful in cases where the data shows uncorrelatedness, and yet presents typical features of MAR processes. Since a constrained MAR process is still dependent, this provides an alternative to GARCH models. After introducing the linear constraints, we discuss how applying them reduces the standard error of the parameter estimates. This shows the advantage of fitting an uncorrelated MAR model when the data satsifies certain assumptions. Furthermore, we consider an application in econometrics, modelling residuals from a fitted model as a constrained MAR.

The contribution with uncorrelated MAR models is to provide a model for uncorrelated but dependent data, such as residuals of a time series model, and has potential to be used to test uncorrelatedness as an alternative to other tests.

- Most of the methodology presented throughout this work are implemented in the R package **mixAR** (Boshnakov and Ravagli, 2020), which has been developed alongside the progression of the research project.

1.2 Structure of the thesis

The thesis is structured as follows:

- Chapter 2 contains a review of MAR models and of the methodology used throughout the following chapters.
- Chapter 3 presents a fully Bayesian analysis of MAR models with Gaussian components.
- Chapter 4 presents a fully Bayesian analysis of MAR models with Student-t components.
- Chapter 5 introduces MVAR models and derivation of multi-step predictive densities for them. Afterwards, the derived formulas are combined with modern portfolio theory to analyse portfolios of assets.
- Chapter 6 presents a constrained version of the MAR model, discusses the advantages of using such constraints under suitable circumstances, and proposes an application to econometric data.

Chapter 2

Theoretical background

The methodology presented in this work finds its main application in the field of financial time series, although it is not limited to that. Financial time series have some peculiar properties which make them not suitable for being modelled with linear time series models: observations are generally uncorrelated or weakly correlated, while the squares of the observations tend to show significant autocorrelation; they often present conditional heteroskedasticity, meaning that, conditional on past information, the variance of the observations is not constant in time; outliers, or unlikely events, tend to occur more often than what is suggested by the Normal distribution, making heavy-tailed distributions more suitable than the Normal; data may sometimes be skewed or present multiple modes, as a sign, for instance, of a change of trend over time.

The two most popular classes of statistical models that account for these features of financial data are that of generalised autoregressive conditional heteroskedasticity (GARCH) models, and that of regime switching models. Mixture autoregressive models belong to the latter.

There has been growing interest in mixture models for financial data in the last few decades, since results such as that of Lanne and Saikkonen (2003) support the hypothesis that mixture models may in fact be better suited to fit this type of data than

other models commonly used.

2.1 Mixture autoregressive models

Mixture autoregressive models (Wong and Li, 2000) were introduced as a flexible way of modelling data that presents heteroskedasticity, asymmetry, multimodality. This makes them particularly suitable for modelling financial and econometric data.

A process $\{y_t\}$ is said to follow a Mixture autoregressive (MAR) model if its cumulative distribution function, conditional on past information, can be written as

$$F(y_t|\mathcal{F}_{t-1}) = \sum_{k=1}^{g} \pi_k F_k \left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k} \right), \tag{2.1}$$

where

- $\mathcal{F}_{t-1}$ is the sigma field generated by the process up to (and including) $t-1$. Informally, $\mathcal{F}_{t-1}$ denotes all the available information at time $t-1$, the most immediate past.
- g is the total number of autoregressive components, or regimes.
- $0 < \pi_k < 1$, $k = 1, \ldots, g$, are the mixing weights or proportions, specifying a discrete probability distribution. So, $\sum_{k=1}^{g} \pi_k = 1$ and $\pi_g = 1 - \sum_{k=1}^{g-1} \pi_k$. We denote the vector of mixing weights by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$. Each π_k is the unconditional probability of an observation to be generated by regime k at any given time t.
- F_k is the distribution function (CDF) of a standardised distribution with location parameter zero and scale parameter one. The corresponding density function is denoted by f_k.
- $\boldsymbol{\phi}_k = \left(\phi_{k1}, \ldots, \phi_{kp_k}\right)$ is the vector of autoregressive parameters for the k^{th} component, with ϕ_{k0} being the shift or intercept. Here, p_k is the autoregressive order

of component k, and we define $p = \max(p_k)$ to be the largest order among the components. A useful convention is to set $\phi_{kj} = 0$, for $p_k + 1 \leq j \leq p$. We may refer to a process with largest order p as a MAR process of order p.

- $\sigma_k > 0$ is the scale parameter for the k^{th} component. We denote by $\boldsymbol{\sigma} = (\sigma_1, \ldots \sigma_g)$ the vector of scale parameters. Furthermore, we define the precision, τ_k, of the k^{th} component by $\tau_k = 1/\sigma_k^2$.

- If the process starts at $t = 1$, then Equation (2.1) holds for $t > p$.

- The MAR model described in (2.1) is formally denoted as MAR$(g; p_1, \ldots, p_g)$, where g is the number of mixture components, and $p_1, \ldots, p_g$ are the autoregressive orders corresponding to the mixture components.

A nice feature of this model is that one-step predictive distributions are given directly by the specification of the model in (2.1). The h-step ahead predictive distribution at time t can be obtained by simulation (Wong and Li, 2000) or, in the case of Gaussian and α-stable components, analytically (Boshnakov, 2009). Furthermore, combining predictive distributions which depend on the recent history of the process, MAR models are very flexible in accommodating asymmetry, multimodality, heteroskedasticity and correlation in time series data.

2.1.1 Mixture autoregression as regime switching model

The idea behind regime switching models is that there may be more than one process, or regime, to govern the evolution of the data over time. It is not known in practice when, and how often regimes will switch, therefore a first order Markov chain is used to estimate the probability of any regime to occur at a given time point. Mathematically speaking, these probabilities are represented by a matrix, called *transition matrix*. Let a time series be described by a model with g distinct regimes. Assuming the most recent

observation was generated from regime i, $i = 1,\ldots,g$, then the present observation at time t is generated from regime j, $j = 1,\ldots,g$ as follows:

$$y_t = \begin{cases} a_1 + b_1 y_{t-1} + \varepsilon_{t1} & \text{with probability } \pi_{i1}, \\ a_2 + b_2 y_{t-1} + \varepsilon_{t2} & \text{with probability } \pi_{i2}, \\ \vdots & \\ a_j + b_j y_{t-1} + \varepsilon_{tj} & \text{with probability } \pi_{ij}, \\ \vdots & \\ a_{g-1} + b_{g-1} y_{t-1} + \varepsilon_{t(g-1)} & \text{with probability } \pi_{i,g-1}, \\ a_g + b_g y_{t-1} + \varepsilon_{tg} & \text{with probability } \pi_{ig}. \end{cases} \tag{2.2}$$

. The transition matrix for the occurrence of these regimes can be written as:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1(g-1)} & p_{1g} \\ p_{21} & p_{22} & \cdots & p_{2(g-1)} & p_{2g} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{(g-1)1} & p_{(g-1)2} & \cdots & p_{(g-1)(g-1)} & p_{(g-1)g} \\ p_{g1} & p_{g2} & \cdots & p_{g(g-1)} & p_{gg} \end{bmatrix} \tag{2.3}$$

where p_{ij}, $i,j = 1,\ldots,g$, denotes the probability that regime i will occur at time t, given that regime j has occured at time $t-1$. The above matrix is a representation of the probability of the next observation being generated from a specific regime, knowing the current state of the Markov chain. Note that (2.2) is only an example of what form regimes may assume, as in fact a variety of process may be considered.

A mixture autoregressive model may be seen as a particular case of regime switching model, in which regimes are distinct autoregressive processes, and they are allowed to alternate any number of times according to some constant probabilities. A MAR model is hence aimed at detection of outliers, as well as breaking points and changes

in the behavior of a time series.

To see the analogy with (2.2) and (2.3), we have that the regimes are:

$$y_t = \begin{cases} \phi_{10} + \sum_{i=1}^{p_1} \phi_{1i} y_{t-i} + \varepsilon_{t1} & \text{with probability } \pi_1, \\ \phi_{20} + \sum_{i=1}^{p_2} \phi_{2i} y_{t-i} + \varepsilon_{t2} & \text{with probability } \pi_2, \\ \quad \vdots & \\ \phi_{g-1,0} + \sum_{i=1}^{p_{g-1}} \phi_{g-1,i} y_{t-i} + \varepsilon_{t,g-1} & \text{with probability } \pi_{g-1}, \\ \phi_{g0} + \sum_{i=1}^{p_g} \phi_{gi} y_{t-i} + \varepsilon_{tg} & \text{with probability } \pi_g. \end{cases} \tag{2.4}$$

where $\varepsilon_{t1}, \ldots, \varepsilon_{tg}$ are independent white noise with equal means 0 and variances $\sigma_1^2, \ldots, \sigma_g^2$. In addition, the transition matrix can now be written as:

$$P = \begin{bmatrix} \pi_1 & \pi_2 & \ldots & \pi_{(g-1)} & \pi_g \\ \pi_1 & \pi_2 & \ldots & \pi_{(g-1)} & \pi_g \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \pi_1 & \pi_2 & \ldots & \pi_{(g-1)} & \pi_g \\ \pi_1 & \pi_2 & \ldots & \pi_{(g-1)} & \pi_g \end{bmatrix}. \tag{2.5}$$

meaning the unconditional probability of a given state to occur is constant over time, and does not depend on which of the regimes has previously occurred.

2.1.2 Stability of the MAR model

Stationarity conditions for MAR time series have some similarity to those for autoregressions, with some notable differences. Below we give the results we need, see Boshnakov (2011) and the references therein for further details.

Consider the companion matrices

$$A_k = \begin{bmatrix} \phi_{k1} & \phi_{k2} & \cdots & \phi_{k(p-1)} & \phi_{kp} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}, \quad k = 1, \dots, g.$$

We say that the MAR model is stable (Lütkepohl, 2007; Boshnakov, 2011) if and only if all eigenvalues of the matrix

$$A = \sum_{k=1}^{g} \pi_k A_k \otimes A_k$$

lie within the unit circle ($\otimes$ is the Kronecker product). What we mean by stable is that, when the condition is satisfied, $\{y_t\}$ is a bounded sequence,such that for some real number B, $P(t : |y_t| > B) = 0$. With this interpretation of stability as boundedness, the properties defined in Lütkepohl (2007) hold, and further assumptions can be made (NOTE: I thank the internal examiner for their constructive feedback and contribution on this particular insight). In fact, $\{y_t\}$ is in this case a well-defined stochastic process, in which distributions and joint distributions of the y_t's are uniquely determined by the distributions and joint distributions of the innovations, and such that $\mathrm{E}\, y_t \varepsilon_s = 0$ for $t < s$.

If a MAR model is stable, then it can be used as a model for stationary time series. The stability condition is sometimes called stationarity condition, as when this condition holds, the model is able to generate a stationary process with suitable initial conditions.

Notice that there may be instances in which the matrix A has eigenvalues outside the unit circle, and yet the model yields stationary solutions. We regard such models as

unstable. It is the case that some properties of the MAR model, such as the recursive equations for the autocorrelation function, do not hold if the model is unstable. For this reason, all of the analysis presented is based on the assumption that the underlying model is stable.

If $g = 1$, the MAR model reduces to an AR model and the above condition states that the model is stable if and only if $A_1 \otimes A_1$ is stable, which is equivalent to the same requirement for A_1. For $g > 1$, it is still true that if all matrices $A_1, \ldots, A_g$, $k = 1, \ldots, g$, are stable, then A is also stable. However, the inverse is no longer true, i.e. A may be stable even if one or more of the matrices A_k are not stable.

What the above means is that the parameters of some of the components of a MAR model may not correspond to causal AR models. It is convenient to refer to such components as "non-stationary".

Partial autocorrelations are often used as parameters of autoregressive models because they transform the stationarity region of the autoregressive parameters to a hypercube with sides $(-1, 1)$ (Barndorff-Nielsen and Schou, 1973; Sampietro, 2006). The above discussion shows that the partial autocorrelations corresponding to the components of a MAR model cannot be used as parameters if coverage of the entire stationary region of the MAR model is desired.

2.1.3 Likelihood function and the missing data formulation

It is straightforward to derive the conditional pdf of a MAR model from (2.1). This is:

$$f(y_t \mid \mathcal{F}_{t-1}) = \sum_{k=1}^{g} \frac{\pi_k}{\sigma_k} f_k \left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k} \right) \tag{2.6}$$

Given a time series $y_1, \ldots, y_n$, the likelihood function for the MAR model is the

product of the conditional densities

$$L(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\pi} | \boldsymbol{y}) = \prod_{t=p+1}^{n} \sum_{k=1}^{g} \frac{\pi_k}{\sigma_k} f_k \left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k} \right), \tag{2.7}$$

where $\boldsymbol{y}$ denotes the full data, i.e. $\boldsymbol{y} = (y_1, \ldots, y_n)$. Notice that the mixture components of a MAR model may not be interpreted, in general, as "real" underlying processes which, in turn, generate the observations, but rather as a way to bring more flexibility in fitting the data. This idea exploits the principle that a mixture of distribution can closley approximate any distribution (McLachlan and Peel, 2000).

Hence, the likelihood is the product of sums. It is often the case in practice that the likelihood function, as written in (2.7), is hardly tractable, so that it is not possible to derive estimators for the model parameters in closed form.

A common way to deal with this class of problems is to resort to the missing data formulation (Dempster et al., 1977). We assume that, at each time t, the corresponding observation y_t is a realisation of exactly one regime. Suppose it was known that y_t had been generated by regime k, then the pdf of y_t would be fully specified by $f_k(\cdot)$. This allows to rewrite the likelihood function as a product, which is simpler to deal with. This procedure is also referred to as *data augmentation*.

Formally, the idea of data augmentation is defined as follows. Let $\boldsymbol{Z}_t = (Z_{t1}, \ldots, Z_{tg})$ be a latent allocation random variable, where $\boldsymbol{z}_t$ is a g-dimensional vector with entry k equal to 1 if y_t was generated from the k^{th} component of the mixture, and 0 otherwise, and such that each $\boldsymbol{z}_t$ has exactly one entry equal to 1. Without further information on the process (i.e. unconditionally), we assume that the $\boldsymbol{z}_t$s are i.i.d. random variables from a discrete distribution with probabilities:

$$P(z_{tk} = 1 | g, \boldsymbol{\pi}) = \pi_k, \qquad k = 1, \ldots, g, \tag{2.8}$$

This setup, widely exploited in the literature (see, for instance Dempster et al., 1977; Diebolt and Robert, 1994) allows to rewrite the likelihood function in a much more tractable way as follows:

$$L(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\pi} | \boldsymbol{y}, \boldsymbol{z}) = \prod_{t=p+1}^{n} \prod_{k=1}^{g} \left(\frac{\pi_k}{\sigma_k} f_k \left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k} \right) \right)^{z_{tk}} \tag{2.9}$$

In practice, the $\boldsymbol{z}_t$s are not available, as we do not know which of the regimes has generated the data point. Both Bayesian and frequentist methods exist to effectively deal with this issue and estimate the latent variables. We introduce both in the following sections.

2.2 The Bayesian approach

Bayesian statistics aims to express the uncertainty about unknown quantities by use of probability distributions. The idea is that of treating parameters of interest as random variables rather than uknown fixed quantities to be estimated. A probability distribution, which represents a degree of belief in an event prior to observing the data, is attached to said random variables. Integrating this prior belief with the evidence arising from the data, one is able to build some posterior belief, which is later used to draw inference about events of interest.

Originated in its raw form as early as 1763 in a paper by Thomas Bayes (hence *Bayesian*), Bayesian statistics has become more and more feasible to implement, and consequently more popular, with the introduction of increasingly fast computers and the creation of efficient sampling algorithms, in particular Markov Chain Monte Carlo (MCMC) methods.

2.2.1 Bayes' theorem

In his paper, Thomas Bayes derives what is nowadays known by everyone as Bayes' theorem. Bayes' theorem is a result in conditional probability. Let θ denote a parameter of interest, and y the observed data. The theorem states that:

$$p(\theta \mid y) = \frac{L(\theta \mid y)\, p(\theta)}{p(y)} \tag{2.10}$$

where $p(\theta)$ is the prior distribution (i.e. the prior belief) on the parameter of interest, $L(\theta \mid y)$ is the likelihood function of the data (i.e. the evidence arising from the data), $p(\theta \mid y)$ is the posterior distribution (i.e. the posterior belief) on the same parameter, and finally $p(y) \neq 0$ is the probability, or density, of observing the data (i.e. the probability of the evidence). If prior and posterior distribution belong to the same family, then we say that the prior distribution is a *conjugate prior* for the likelihood (Diaconis and Ylvisaker, 1979). Suppose for instance that the likelihood function is a product of Gaussian densities. By imposing a Gaussian prior distribution on the parameter of interest (e.g. the mean), we ensure that the posterior distribution will also be Gaussian. This implies that a Gaussian prior is a conjugate prior for a Gaussian likelihood.

In other words, the theorem proves that the posterior belief on a parameter is, up to proportionality, the product of the prior belief and the likelihood function of the observed data. It is common to rewrite the equality in (2.10) as a relationship of proportionality:

$$\begin{gathered} p(\theta \mid y) \propto L(\theta \mid y)\, p(\theta) \\ \text{or} \\ posterior \propto likelihood \times prior \end{gathered} \tag{2.11}$$

denoting proportionality between the posterior distribution and the product between

likelihood function and prior distribution. This is because many of the available sampling methods are able to approximate $p(\theta \mid y)$ without having to calculate $p(y)$ exactly, which may in most cases be cumbersome.

The theorem is also valid with a multidimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$. In this case, the full conditional posterior distribution of a generic element of the parameter vector, say θ_k, is not only conditional on the data y, but also on the remaining parameters, $\boldsymbol{\theta}_{-k}$. Hence, (2.11) becomes:

$$p(\theta_k \mid y, \boldsymbol{\theta}_{-k}) \propto L(\theta_k \mid y, \boldsymbol{\theta}_{-k})\, p(\theta_k) \tag{2.12}$$

The proportionality term in the denominator is now also dependent on $\boldsymbol{\theta}_{-k}$. However, this does not represent an issue, as it is not a function of θ_k, and therefore it is constant for all values of θ_k. This means that the relationship of proportionality still holds.

2.2.2 Markov chain Monte Carlo

It is often the case in practice that the forementioned posterior distributions are of a complex form, or are highly dimensional. Bayesian statistics requires evaluation of the expectation of these functions, exact derivation of which is rather infeasible. For this reason, Bayesian analyses have mostly been limited to conjugate cases in the past, where the posterior distribution could be easily derived. With the introduction of numerical methods for approximating functions and their expectations, and the increasing availability of poweful computers that could quickly implement them, Bayesian statistics has had a steep development in the last few decades, seeing its biggest revolution with Markov chain Monte Carlo (MCMC) methods, which we here discuss.

Markov chains Earlier in this chapter we introduced the notion of Markov chains. Markov chains are stochastic processes defined by *states*: a state denotes the occurrence of a certain event at a given time point. Each state is associated with a probability of occurrence. Markov chains are called such because they satisfy the so called *Markov property*: the probability of a certain event to occur next only depends on the latest event to have occured, the present. Formally, this means that the probability of a certain state to occur at the next time point only depends on the current state of the chain, and not at all on the path that led up to that. Thanks to this property, a Markov chain may be described by a transition matrix like that in (2.3).

Let $x_0, x_1, \ldots, x_n$ be the first n states of a Markov chain. For the next step x_{n+1}, the Markov property states that:

$$P(X_{n+1} = x_{n+1} \mid X_n = x_n, \ldots, X_1 = x_1, X_0 = x_0) = P(X_{n+1} = x_{n+1} \mid X_n = x_n) \quad (2.13)$$

A Markov Chain is said to be :

- *aperiodic* if there is no certainty that the same state will recur at regular intervals;
- *recurrent* if the probability of the chain returning to the same state in a finite number of steps is non-zero;
- *irreducible* if, at each step, all states have non-zero probability of occurrence.

A Markov chain that satisfies all three properties is called *ergodic*, and it is guaranteed to converge to an *equilibrium distribution* within a finite number of steps.

Monte Carlo method Monte Carlo is a method for numerical integration which uses random number generators. In principle, it can be used to approximate definite integrals by simulating random points at which the integrand is evaluated.

In Bayesian statistics, Monte Carlo is used to approximate posterior distributions. The idea is that any posterior distribution $p(\theta \mid y)$ may be approximated by simulating a sufficiently large sample of realisations from $p(\theta \mid y)$, directly or indirectly. Consequently, the summary statistics of the distribution are approximated by those of the obtained sample.

Markov chain Monte Carlo Combining the two concepts of Markov chains and Monte Carlo together, it is possible to approximate any posterior distribution of interest. This can be achieved by implementing a Markov chain with the desired distribution as its equilibrium, or stationary distribution. Once the stationary distribution has been achieved, a sufficiently large sample can be simulated to approximate such distribution and the statistics of interest.

Notice that it is not necessary to simulate directly from the target distribution, as this may be complicated in many occasions. In the next part, we review some of the simulation methods used througout this project.

2.2.3 Review of MCMC methods

Gibbs sampling Gibbs sampling (Geman and Geman, 1984) is a method for direct simulation from full conditional posterior distributions. Suppose we are interested in approximating the joint posterior distribution $p(\boldsymbol{\theta} \mid y)$, where $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_q\right)$. The conditional distribution of one variable, say θ_j, given all others is proportional to the joint posterior distribution:

$$p\left(\theta_j \mid y, \boldsymbol{\theta}_{-j}\right) \propto p\left(\theta_1, \ldots, \theta_q\right)$$

up to some normalisation constant. The idea of Gibbs sampling is that it is simpler to simulate from the full conditional posteriors than from the joint posterior distribution.

Given the above consideration, a Gibbs sampler can be implemented as follows:

1. Choose initial values for the parameters $\theta_1, \ldots, \theta_q$. Starting values may be chosen arbitrarily or, for instance, simulated from the prior distribution, since the choice may only affect how quickly the chain will reach the stationary distribution, but not convergence itself.

2. Given full conditional distributions of the parameters, simulate new values from their respective posterior distributions, conditional on the most up-to-date states of the remaining parameters. For instance, suppose the chain has completed m steps, $m \geq 0$, and the algorithm now moves to iteration $m+1$. We simulate $\theta_1, \ldots, \theta_q$ as:

$$
\begin{aligned}
&\text{draw } \theta_1^{m+1} \text{ from } p\left(\theta_1 \mid \theta_2^m, \theta_3^m, \ldots, \theta_{q-1}^m, \theta_q^m\right) \\
&\text{draw } \theta_2^{m+1} \text{ from } p\left(\theta_2 \mid \theta_1^{m+1}, \theta_3^m, \ldots, \theta_{q-1}^m, \theta_q^m\right) \\
&\text{draw } \theta_3^{m+1} \text{ from } p\left(\theta_3 \mid \theta_1^{m+1}, \theta_2^{m+1}, \ldots, \theta_{q-1}^m, \theta_q^m\right) \\
&\qquad\qquad \vdots \\
&\text{draw } \theta_{q-1}^{m+1} \text{ from } p\left(\theta_{q-1} \mid \theta_1^{m+1}, \theta_2^{m+1}, \theta_3^{m+1} \ldots, \theta_q^m\right) \\
&\text{draw } \theta_q^{m+1} \text{ from } p\left(\theta_q \mid \theta_1^{m+1}, \theta_2^{m+1}, \theta_3^{m+1} \ldots, \theta_{q-1}^{m+1}\right).
\end{aligned}
$$

3. Repeat step 2 until a large enough sample is obtained.

Notice that this procedure forms an ergodic Markov chain, and is hence guaranteed to reach the stationary distribution. By simply dropping some early draws (the so called *burn-in* period), the remaining sample will accurately approximate full conditional posterior distributions, so that summary statistics may be estimated using sample estimates.

Gibbs sampling is a particular case of Metropolis-Hastings algorithm, which we

discuss next.

Metropolis-Hastings algorithm Metropolis-Hastings algorithm is a method for obtaining a sample from a posterior distribution which is difficult to simulate from, very useful in particular with multi-dimensional distributions.

The method requires specification of a distribution $q(\cdot,\cdot)$, proportional to the target distribution $f(\cdot)$, which is easy to simulate from, and which values $q(\boldsymbol{\theta})$ can be calculated. $q(\cdot,\cdot)$ is called the *proposal distribution*, and it is used to generate candidate values for updating of the Markov chain. In general, $q(\cdot,\cdot)$ may be chosen to depend on the most recent draw. For instance, it could be a Normal distribution centered in the current state of the chain. However, this is not a necessary condition.

Once $q(\cdot,\cdot)$ is chosen, the algorithm is initiated by choosing an arbitrary starting point $\boldsymbol{\theta}^{(0)}$ for the model paramter. $\boldsymbol{\theta}^{(0)}$ can be univariate or multivariate. We then proceed as follows:

1. At a generic iteration $m+1$, $m \geq 0$, simulate a candidate value $\boldsymbol{\theta}^{(*)}$ from the proposal distribution.

2. Calculate the acceptance probability for $\boldsymbol{\theta}^{(*)}$, $\alpha\left(\boldsymbol{\theta}^{(m)},\boldsymbol{\theta}^{(*)}\right)$:

$$\alpha\left(\boldsymbol{\theta}^{(m)},\boldsymbol{\theta}^{(*)}\right) = \min\left\{1, \frac{f\left(\boldsymbol{\theta}^{(*)}\right) q\left(\boldsymbol{\theta}^{(m)},\boldsymbol{\theta}^{(*)}\right)}{f\left(\boldsymbol{\theta}^{(m)}\right) q\left(\boldsymbol{\theta}^{(*)},\boldsymbol{\theta}^{(m)}\right)}\right\} \tag{2.14}$$

   where $q(x,y)$ is the transition probability of moving to state y given that x is the current state.

3. The candidate value is retained with probability $\alpha\left(\boldsymbol{\theta}^{(m)},\boldsymbol{\theta}^{(*)}\right)$. The decision is made by simulating a value u such that $U \sim Un(0,1)$. If $u \leq \alpha\left(\boldsymbol{\theta}^{(m)},\boldsymbol{\theta}^{(*)}\right)$, $\boldsymbol{\theta}^{(*)}$ is retained, and set $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(*)}$; if $u > \alpha\left(\boldsymbol{\theta}^{(m)},\boldsymbol{\theta}^{(*)}\right)$, $\boldsymbol{\theta}^{(*)}$ is rejected, and set $\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)}$.

Random walk Metropolis Random walk Metropolis (RWM) is a particular case of Metropolis-Hastings algorithm in which the proposal distribution $g(\cdot)$ is symmetric. The symmetry in the proposal implies that the proposal ratio $\frac{q\left(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(*)}\right)}{q\left(\boldsymbol{\theta}^{(*)}, \boldsymbol{\theta}^{(m)}\right)}$ is always equal to 1. Therefore, the acceptance probability in (2.14) becomes:

$$\alpha\left(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(*)}\right) = \min\left\{1, \frac{f\left(\boldsymbol{\theta}^{(*)}\right)}{f\left(\boldsymbol{\theta}^{(m)}\right)}\right\} \tag{2.15}$$

For highest efficiency, it is common to set up RWM so that the proportion of accepted candidate values is between $20\% - 25\%$.

Independent sampler When there is not an obvious choice of what the proposal distribution should be, one choice is to generate candidate values from a fixed distribution, $g(\cdot)$, which does not depend on the current state of the chain. In this way, for a candidate value $\boldsymbol{\theta}^{(*)}$ and current state $\boldsymbol{\theta}^{(m)}$, the proposal distribution is $q\left(\boldsymbol{\theta}^{(*)}, \boldsymbol{\theta}^{(m)}\right) = g\left(\boldsymbol{\theta}^{(*)}\right)$. The acceptance probability for a candidate value then becomes:

$$\alpha\left(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(*)}\right) = \min\left\{1, \frac{f\left(\boldsymbol{\theta}^{(*)}\right) g\left(\boldsymbol{\theta}^{(m)}\right)}{f\left(\boldsymbol{\theta}^{(m)}\right) g\left(\boldsymbol{\theta}^{(*)}\right)}\right\} \tag{2.16}$$

In a Bayesian setting, the proposal distribution may be the prior distribution $p(\boldsymbol{\theta})$. With $f(\cdot)$ being the posterior distribution (up to proportionality), the acceptance probability for a candidate value is now:

$$\alpha\left(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(*)}\right) = \min\left\{1, \frac{L\left(\boldsymbol{\theta}^{(*)}\right) p\left(\boldsymbol{\theta}^{(*)}\right) p\left(\boldsymbol{\theta}^{(m)}\right)}{L\left(\boldsymbol{\theta}^{(m)}\right) p\left(\boldsymbol{\theta}^{(m)}\right) p\left(\boldsymbol{\theta}^{(*)}\right)}\right\} \tag{2.17}$$

which reduces to $\min\left\{1, \frac{L\left(\boldsymbol{\theta}^{(*)}\right)}{L\left(\boldsymbol{\theta}^{(m)}\right)}\right\}$ if the prior distribution is symmetric.

2.2.4 Bayesian model selection

When attempting to predict a certain event, it is important to measure the uncertainty associated with the prediction. For instance, there may be several models that describe the same event, and we may be interested in assessing which one provides the best fit given the evidence from the data, and how confident we are about that particular model being the best. In Bayesian statistics, this can be done by estimating posterior probabilities of candidate models.

Reversible jump MCMC Reversible jump MCMC (RJMCMC) is itself a type of Metropolis-Hastings algorithm. It allows simulation of posterior probabilities of models with different parameters (formally models with parameter spaces of varying dimensions). In a linear model, it may be used to assess which covriates are relevant towards predicting the response. Translated into time series, RJMCMC could be used, for example, to assess the ”best” order of an autoregressive model. This is also how we will use the methodology for MAR models.

Let $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_g\}$ be a set containing g distinct candidate models, each describing a particular event. Suppose that, at step m of the chain, $m \geq 0$, the current model is $\mathcal{M}^{(m)} = \mathcal{M}_i$, and that we propose a move to model $\mathcal{M}_j$. Because the two models have different parameter space, it is required to create a mapping between the two sets of paramters that allows the jump from one model to the other. This mapping is also called *dimension matching*, and is essentially a reparametrisation of model $\mathcal{M}_i$ into model $\mathcal{M}_j$. This is often done by generation of random variables. For an example relevant to our case, assume that the two models have parameter vectors respectively $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, where $\boldsymbol{\theta}_j$ has k addotional parameter. We may generate a realisation $\boldsymbol{u}$ from

an arbitrarily selected k-dimensional random variable U with density $g(u)$, and create the 1 to 1 mapping $\boldsymbol{\theta}_j \rightarrow (\boldsymbol{\theta}_i, \boldsymbol{u})$.

The acceptance rate of the proposed move from $\mathcal{M}_i$ to $\mathcal{M}_j$ involves a ratio of the posterior distributions of the parameters given the data y, multiplied by the ratio of proposal distributions, i.e. a function $q(x,y)$ that determines the probability of a proposed move from x to y:

$$\alpha\left(\mathcal{M}_i, \mathcal{M}_j\right) = \min\left\{1, \frac{L\left(\boldsymbol{\theta}_j \mid y\right)}{L\left(\boldsymbol{\theta}_i \mid y\right)} \times \frac{p\left(\boldsymbol{\theta}_j\right)}{p\left(\boldsymbol{\theta}_i\right)} \times \frac{q\left(\boldsymbol{\theta}_i\right)}{q\left(\boldsymbol{\theta}_j\right) g(u)} \times |J|\right\} \tag{2.18}$$

The last term of the product, $|J|$, is the determinant of the Jacobian matrix, that is determined by the mapping between the two models.

Steps for updating the "current" model are the same as for all other examples of Metropolis-Hastings algorithm seen so far.

Marginal likelihood from Gibbs and Metropolis-Hastings output All methods described so far estimate posterior distributions up to proportion. In fact, for any of those methods it is not necessary to calculate the normalising constant (the denominator in (2.10)), which is often very challenging. This normalising constant is also the marginal likelihood of the model.

The marginal likelihood of a model is the likelihood function after some or all of the parameters have been marginalised. Formally, given data $\boldsymbol{y} = (y_1, \ldots, y_n)$ where $\boldsymbol{y} \sim p(\boldsymbol{y} \mid \boldsymbol{\theta})$, $\boldsymbol{\theta}$ is a parameter vector and itself a random variable such that $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid x)$, the marginal likelihood is:

$$p(y_1, \ldots, y_n \mid x) = \int_{\boldsymbol{\theta}} p(y_1, \ldots, y_n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid x)\, d\boldsymbol{\theta} \tag{2.19}$$

which is the likelihood of the model after marginalising with respect to $\boldsymbol{\theta}$. Here, x denotes additional information about the model other than $\boldsymbol{\theta}$.

Chib (1995) and Chib and Jeliazkov (2001) proposed methods to calculate the marginal likelihood from the output of, respectively, Gibbs sampling and Metropolis-Hastings output.

Since the marginal likelihood is the normalising constant in the posterior density, we define the marginal likelihood identity by rearranging (2.11):

$$p(\boldsymbol{y} \mid x) = \frac{L(\boldsymbol{\theta} \mid \boldsymbol{y}, x)\, p(\boldsymbol{\theta} \mid x)}{p(\boldsymbol{\theta} \mid \boldsymbol{y}, x)} \tag{2.20}$$

This equality allows us to estimate the marginal likelihood at a single point $\boldsymbol{\theta}^*$, as long as the posterior distribution is available (known or estimated).

First, we discuss estimation of the marginal likelihood from Gibbs sampling. The output of Gibbs sampling provides an estimate of the posterior distribution $p(\boldsymbol{\theta} \mid x)$. Full conditional posterior distributions of $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_q\right)$ are

$$p(\theta_i \mid \boldsymbol{\theta}_{-i}, \boldsymbol{y}, x), \quad i = 1, \ldots, q,$$

which estimates are available via Gibbs sampling.

Now, let $\boldsymbol{\theta}^*$ be a single ordinate of the parameter vector. For efficiency reasons, this is normally taken as a high density point in the posterior distribution. The joint conditional posterior denisity at $\boldsymbol{\theta}^*$ can be factorised into:

$$p(\boldsymbol{\theta}^* \mid \boldsymbol{y}, x) = \prod_{i=1}^{q} p\left(\theta_i^* \mid \boldsymbol{\theta}_{-i}^*, \boldsymbol{y}, x\right) \tag{2.21}$$

with

$$p\left(\theta_i^* \mid \boldsymbol{\theta}_{-i}^*, \boldsymbol{y}, x\right) = \int p\left(\theta_i^* \mid \theta_1^*, \ldots, \theta_{i-1}^*, \theta_{i+1}, \ldots, \theta_q, \boldsymbol{y}, x\right) \times p\left(\theta_{i+1}, \ldots, \theta_q \mid \theta_1^*, \ldots, \theta_{i-1}^*, \boldsymbol{y}, x\right) d\theta_{i+1} \ldots, d\theta_q \tag{2.22}$$

Given a sample of size N obtained via Gibbs sampling, the integrals in (2.22) can then be estimated via Monte Carlo, so that:

$$\hat{p}\left(\theta_i^* \mid \theta_1^*, \ldots, \theta_{i-1}^*, \boldsymbol{y}, x\right) = \sum_j^N p\left(\theta_i^* \mid \theta_1^*, \ldots, \theta_{i-1}^*, \theta_{i+1}^{(j)}, \ldots, \theta_q^{(j)}, \boldsymbol{y}, x\right) \tag{2.23}$$

The methodology can be generalised to all Metropolis-Hastings algorithm, as shown by Chib and Jeliazkov (2001).

Let $\Omega_{i-1} = (\theta_1, \ldots, \theta_{i-1})$ and $\Omega_{i+1} = \left(\theta_{i+1}, \ldots, \theta_q\right)$. This time, suppose that for each θ_i the proposal distribution for a move from θ_i to $\theta_i^{'}$ is $q\left(\theta_i, \theta_i^{'} \mid \Omega_{i-1}, \Omega_{i+1}, \boldsymbol{y}\right)$. The acceptance probability of the proposed move is:

$$\alpha\left(\theta_i, \theta_i^{'} \mid \Omega_{i-1}, \Omega_{i+1}, \boldsymbol{y}\right) = \min\left\{1, \frac{f\left(\boldsymbol{y} \mid \theta_i^{'}\right) p\left(\theta_i^{'}\right) q\left(\theta_i^{'}, \theta_i \mid \boldsymbol{\theta}_{-i}, y\right)}{f\left(\boldsymbol{y} \mid \theta_i\right) p\left(\theta_i\right) q\left(\theta_i, \theta_i^{'} \mid \boldsymbol{\theta}_{-i}, \boldsymbol{y}\right)}\right\} \tag{2.24}$$

Now, because Metropolis-Hastings satisfies the principle of detailed balance of Markov chains, for any single ordinate $\boldsymbol{\theta}^*$ it holds that:

$$\begin{aligned}
&\alpha\left(\theta_i, \theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) q\left(\theta_i, \theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) p\left(\theta_i \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) = \\
&\quad \alpha\left(\theta_i^*, \theta_i \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) q\left(\theta_i^*, \theta_i \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) p\left(\theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right).
\end{aligned} \tag{2.25}$$

Finally, integrating both sides with respect to θ_i and rearranging the terms, we obtain:

$$\begin{aligned}
p\left(\theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) &= \frac{\int \alpha\left(\theta_i, \theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) q\left(\theta_i, \theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) p\left(\theta_i \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) d\theta_i}{\int \alpha\left(\theta_i^*, \theta_i \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) q\left(\theta_i^*, \theta_i \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) d\theta_i} \\
&= \frac{\mathrm{E}_1\left[\alpha\left(\theta_i, \theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) q\left(\theta_i, \theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right)\right]}{\mathrm{E}_2\left[\alpha\left(\theta_i^*, \theta_i \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right)\right]},
\end{aligned} \tag{2.26}$$

where E_1 is with respect to $p\left(\theta_i \mid \boldsymbol{y}\right)$ and E_2 is with respect to $q\left(\theta_i^*, \theta_i\right)$. Repeating this for $i = 1, \ldots, q$ we have the factors needed for estimation of the marginal likelihood.

When a sample of size N from the posterior distribution of $\theta_1, \ldots, \theta_q$ has been

obtained via Metropolis-Hastings, the factors in (2.26) are then estimated by Monte Carlo as:

$$\hat{p}\left(\theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) = \frac{\sum_{j=1}^{N} \alpha\left(\theta_i^{(j)}, \theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) q\left(\theta_i^{(j)}, \theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right)}{\sum_{j=1}^{N} \alpha\left(\theta_i^*, \theta_i^{(j)} \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right)} \tag{2.27}$$

and ultimately the marginal likelihood, evaluated at $\boldsymbol{\theta}^*$ is the product of these factors, i.e.:

$$\hat{p}\left(\boldsymbol{\theta}^* \mid \boldsymbol{y}\right) = \prod_{i=1}^{q} \hat{p}\left(\theta_i^* \mid \Omega_{i-1}^*, \Omega_{i+1}, \boldsymbol{y}\right) \tag{2.28}$$

2.2.5 The label switching problem

A common problem associated with Bayesian analysis of mixtures is that of label switching (see for instance Celeux, 2000), which derives from symmetry in the likelihood function. If no prior information is available to distinguish the mixture components, then the posterior distribution will also be symmetric. It is essential that label switching is detected and handled properly in order to obtain meaningful results. A common way to deal with this is to impose some sort of ordering of the mixture components through identifiability constraints, e.g. imposing in (2.1) that $\pi_1 > \pi_2 > \pi_g$. However, it is known that such constraints may lead to bias and other issues (think for instance of the case of two regimes with $\pi_1 = \pi = 2 = 0.5$). In the case of MAR models, Hossain (2012) showed that these constraints may affect convergence of the chain to the posterior distribution. More examples of this issue are given in the discussion to the paper by Richardson and Green (1997).

Throughout this work, label switching is dealt with using a *k*-means clustering algorithm proposed by Celeux (2000).

The algorithm works by first choosing the first m simulated values of the output

after convergence. The value m shall be chosen small enough for labels switch to not have occurred yet, and large enough to be able to calculate reliable initial values of cluster centres and their respective variances.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_g)$ be a subset of model parameters of size g, and N the size of the converged sample. For any centre coordinate θ_i, $i = 1, \ldots, g$, we calculate the mean and variance, based on the first m simulated values, respectively as:

$$\bar{\theta}_i = \frac{1}{m}\sum_{j=1}^{m}\theta_i^{(j)} \qquad \bar{s}_i^2 = \frac{1}{m}\sum_{j=1}^{m}\left(\theta_i^{(j)} - \bar{\theta}_i\right)^2$$

We set this to be the "true" permutation of the components, i.e. we now have an initial center $\bar{\boldsymbol{\theta}}^{(0)}$ with variances $\bar{s}_i^{(0)^2}$, $i = 1, \ldots, g$. The remaining $g! - 1$ permutations can be obtained by simply permuting these centres.

From these initial estimates, the r^{th} iteration ($r = 1, \ldots, N - m$) of the procedure consists of two steps:

- the parameter vector $\boldsymbol{\theta}^{(m+r)}$ is assigned to the cluster such that the normalised squared distance

$$\sum_{i=1}^{g}\frac{\left(\theta_i^{(m+r)} - \bar{\theta}_i^{(m+r-1)}\right)^2}{\left(s_i^{(m+r-1)}\right)^2} \tag{2.29}$$

  is minimised, where $\bar{\theta}_i^{(m+r-1)}$ is the i^{th} centre coordinate and $s_i^{(m+r-1)}$ its standard deviation, at the latest update $m + r - 1$.

- Centre coordinates and their variances are respectively updated as follows:

$$\bar{\theta}_i^{(m+r)} = \frac{m+r-1}{m+r}\bar{\theta}_i^{(m+r-1)} + \frac{1}{m+r}\theta_i^{(m+r)} \tag{2.30}$$

and

$$(s_i^{(m+r)})^2 = \frac{m+r-1}{m+r}(s_i^{(m+r-1)})^2 + \frac{m+r-1}{m+r}\left(\bar{\theta}_i^{(m+r-1)} - \bar{\theta}_i^{(m+r)}\right)^2 + \frac{1}{m+r}\left(\theta_i^{(m+r)} - \bar{\theta}_i^{(m+r)}\right)^2 \tag{2.31}$$

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

It is important to aknowledge that the choice of a subset is not always objective. In fact, clusters may appear in different subsets of the parameters for different datasets. As a result, certain choices of subsets may be ineffective for the purpose of relabelling. Translated into the the context of mixtures, clusters may sometimes be clear in the mixing weights $\pi_1, \ldots, \pi_g$ at times, or in the scale parameters $\sigma_1, \ldots, \sigma_g$ at others, and so on. Therefore this method requires graphical assistance, checking the raw output looking for the clearest group separation. For MAR models however, it is advisable not to use the autoregressive parameters, especially when the orders are different.

Once the selected subset has been relabelled, labels for the remaining parameters can be switched accordingly.

2.2.6 Label switching and marginal likelihood

We here discuss the possible effect of incorrect label switching on the methodology in Section 2.2.4 for calculation of the marginal likelihood of the data. Recall the formula:

$$p(\boldsymbol{y} \mid x) = \frac{L(\boldsymbol{\theta}^* \mid x)\, p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^* \mid \boldsymbol{y}, x)}$$

where $\boldsymbol{\theta}^*$ is a point of high density (ideally of highest density) according to its posterior distribution.

For mixture models, we have that the likelihood function $L(\boldsymbol{\theta}^* \mid x)$ is a product of sums. For simplicity, suppose the model is a mixture of two components, and $\boldsymbol{\theta} =$

$(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. It follows that the conditional likelihood is

$$L(\boldsymbol{\theta}^* \mid x) = \prod_{i=1}^{n} \pi_1 f(y_i \mid \theta_1) + \pi_2 f(y_i \mid \theta_2) = \prod_{i=1}^{n} \pi_2 f(y_i \mid \theta_2) + \pi_1 f(y_i \mid \theta_1)$$

which means that the likelihood is invariant with respect to the permutation of $\boldsymbol{\theta}$.

Under the same example, prior and posterior distributions for $\boldsymbol{\theta}$ may somewhat be affected by label switching. For prior distributions, this will happen when the practician sets up the experiment with informative priors, as this would bring the risk of evaluating a parameter under the wrong prior distribution. However, informative priors have the purpose of creating enough separation so that label switching does not in fact occur, as they incorporate prior belief on the distribution of the parameters (see Celeux, 2000). In the examples presented here, prior distributions are the same across all components for corresponing parameters (for instance, all precisions follow a priori the same Gamma distribution), and therefore label switching will not affect the result.

Posterior distributions are the most subject to the effect of label switching. However, we point few remarks in favor of the effectiveness of Chib (1995) and Chib and Jeliazkov (2001), even in the case of undetected label switching:

- The authors reassure that the methodology works effectively with a range of high density values under their respective posterior distributions. Returning to the two-component mixture example, suppose that there is undetected switching. The corresponding parameters in the two components, for example π_1 and pi_2, will show two modes. These modes will however correspond to the two highest density values, respectively, of π_1 and π_2. Therfore, it makes sense to believe that, ultimately, the choice of π_1^* and π_2^* will not change significantly, and high density values will be selected regardless.

- From (2.23) and (2.27), it is clear that undetected label switching could cause

issues in evaluation of the posterior density of $\boldsymbol{\theta}^*$. This brings forward two considerations: first of all, label switching may occur due to little separation between the groups, meaning the two posterior distributions shall not be too dissimilar and a wrong labelling of a few iterations may not affect significantly the evaluation. Secondly, even when incorrect labelling does have an effect, each iteration is dampened by a $1/N$ factor since we take an average over the entire sample.

- The algorithm in Section 2.2.4 sequentially fixes a set of parameters to their highest density values. This implies that, after very few parameters are fixed, label switching will definitely not occur for the remaining parameters. Going back to the two-component example, it is obvious that once we fix $\boldsymbol{\theta}_1^*$, there can no longer be label switching, since now we only draw a sample from $\boldsymbol{\theta}_2$.

- Finally, we must take into account that the contribution of the posterior distribution towards $p(\boldsymbol{y} \mid x)$ will in general be rather small compared to that of $L(\boldsymbol{\theta} \mid \boldsymbol{y}, x)$, which is "immune" to label switching.

All things considered, we therefore conclude that, while handled correctly throughout every example in this project, the effect of label switching could in general be neglectible in terms of correct model selection with marginal likelihood.

2.3 The frequentist approach

Unlike Bayesian statistics, frequentist inference does not treat model parameters as random variables, but rather as unknown fixed quantities to be estimated. Furthermore, frequentist inference does not in any way incorporate subjective information (besides the choice of model) in the estimation process. Results are solely based on evidence

extracted from a given dataset, which is deemed to be representative of the population of interest.

While the origins of frequentist, or "classical", statistics date back to the 19^{th} century, it was formally introduced in the early 1900s, with the works of Fisher, Neyman and Pearson, who set the baseline theory for estimation and confidence intervals, as well as hypothesis and importance testing.

In the context of mixture models, frequentist parameter estimation is usually done by EM-algorithm, which we now introduce.

2.3.1 Expectation-Maximisation algorithm

The Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) is an iterative method to find local maximum likelihood estimates of the parameters of a statistical model in cases where the maximisation problem cannot be solved directly.

The idea of EM algorithm is to consider the likelihood function of the data as incomplete. To "complete" the likelihood, a set of latent, unobservable variables are introduced of the type defined in Section 2.1.3. The latent variables function as missing values from the original data. Since these variables are not observable, they need to be included in the estimation process.

Let $\boldsymbol{y}$ be a dataset, $\boldsymbol{\theta}$ the model parameter vector, and $\boldsymbol{z}$ the latent variables. In addition, denote $L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z})$, the likelihood function of the data. The EM algorithm aims at maximising the marginal likelihood

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \int L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z})\, d\boldsymbol{z} \tag{2.32}$$

Assume however that this marginal likelihood, as it is, is intractable. The EM algorithm is used to find a local MLE to $L(\boldsymbol{\theta};\boldsymbol{y})$, with a two-step iteration involving Z:

- **E step:** calculate the expected value of the loglikelihood $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right)$:

$$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = E_{\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(t)}}\left[L\left(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z}\right)\right] \tag{2.33}$$

  This step is used to predict the latent variables $\boldsymbol{z}$ through their expectations, conditional on the known parameter vector $\boldsymbol{\theta}$;

- **M step:** Maximise $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right)$ conditional on the current conditional distribution of the complete data $(\boldsymbol{y},\boldsymbol{z})$:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}}\{Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right)\} \tag{2.34}$$

The two steps are iterated until convergence.

An important property of the EM algorithm is that convergence to a local maximum is always guaranteed. However, in the case of multiple local maxima, convergence to the global maximum is not guaranteed. Therefore one must be alert of the likelihood function possibly having several local maxima (a common feature in mixture models), and carefully choose appropriate starting values for implementation of the algorithm.

2.4 Diagnostics for MAR models

For diagnostics of MAR models, we calculate two types of errors:

- The first set of errors, ε_t follows the "standard" definition, calculated as the difference between the observation and its conditional expectation or conditional predictor:

$$\varepsilon_t = y_t - \mathrm{E}\left[y_t \mid \mathcal{F}_{t-1}\right] = y_t - \hat{y}_t. \tag{2.35}$$

  While no distribution assumption can be made about ε_t, these errors should still

be uncorrelated (hence lack of serial correlation is a first sign of good model fit). Furthermore, it can be proved that ε_t form a *martingale difference sequence* with mean 0 and positive, finite variance with an upper bound at the unconditional variance of y_t. (see Akinyemi, 2013, for details).

We can perform some transformations on the y_t to assess goodness of fit of the model. Smith (1985) suggests that, under correct model specification, the following assumptions on transformations of y_t are correct:

$$\begin{aligned} U_t &= F\left(y_t \mid \mathcal{F}_{t-1}\right) \sim U\left(0,1\right) \\ V_t &= \boldsymbol{\Phi}^{-1}\left(U_t\right) \sim N\left(0,1\right) \end{aligned} \tag{2.36}$$

One can therefore calculate the series U_t and V_t from the fitted model, and test their respective distributional assumption.

- The second set of errors, denoted $\tilde{\varepsilon}_t$, arises from exploiting the assumption of a mixture model. The idea for estimation of such errors is that if the model is correctly specified, then every observation shall be assigned to its "true" mixture component it was generated from. In this sense, we make use of the conditional expectations of the latent allocation variables, z_t. Recall the formula:

$$\mathrm{E}\left[z_{tk} \mid \boldsymbol{y}, \boldsymbol{\mu}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\pi}}\right] = \hat{\tau}_{tk} = \frac{\hat{\pi}_k f_k\left(\frac{\hat{e}_{tk}}{\hat{\sigma}_k}\right)}{\sum_{l=1}^{g} \hat{\pi}_l f_l\left(\frac{\hat{e}_{tl}}{\hat{\sigma}_l}\right)}$$

where $\hat{e}_{tk} = y_t - \hat{\phi}_{k0} - \sum_{i=1}^{p} \hat{\phi}_{ki} y_{t-i}$ for $k = 1, \ldots, g$.

Naturally, the largest $\hat{\tau}_{tk}$ will occur for that mixture component k which generated y_t. Hence, we choose k, the "true" mixture component to have generated y_t, as the one for which $\hat{\tau}_{tk}$ is the largest. For that component k, we calculate the

standardised error as if the model was now an $AR(p_k)$:

$$\tilde{\varepsilon}_t = \frac{\hat{e}_{tk}}{\hat{\sigma}_k} \tag{2.37}$$

Under correct model specification, each residual $\hat{e}_{tk}$ will be standardised by its "true" standard deviation, up to variance of the estimator $\hat{\sigma}_k$. Therefore, the series of residuals $\tilde{\varepsilon}_t$ will appoximately follow the standardised distribution of interest. For instance, if $f_k \equiv \phi$, the standard Normal distribution for $k = 1, \ldots, g$, then the distribution of $\tilde{\varepsilon}_t$ would be standard Normal under correct model specification. This provides us with one further tool for assessing the goodness of fit of the model, as we can calculate $\tilde{\varepsilon}_t$, and test for uncorrelatedness and distribution of this set of residuals.

A limitation of this method is that it depends on correct classification of the observations, therefore it may not be accurate when mixture components are not well separated. However, if mixture components are so indistinguishable that correct classification of observations is not possible, one should question whether the right number of mixture components has been chosen, or even if the choice of a mixture model was correct in the first place.

2.5 Prediction with density forecasts

A density forecast of the realisation of a random variable at some future time horizon is an estimate of the probability distribution of the possible values that random variable may assume. It hence provides a measure of the uncertainty associated with a prediction, as opposed to point forecasts, which by themselves do not give any description of uncertainty.

In the context of mixture models, density forecasts are often more attractive than

point predictors and prediction intervals. This is because the qualitative features of a predictive distribution, such as multiple modes or skewness, are more intuitive and useful than just a forecast and its associated prediction interval, which are unable to catch such behavior. Think for example of the point prediction for a symmetric bi-modal density: a point prediction would fall exactly between the two modes, in a point of low density, and would therefore be misleading. In addition, when the predictive distribution is available, prediction intervals can easily be obtained by extracting the quantiles of interest from the distribution (Boshnakov, 2009; Lawless and Fredette, 2005).

2.5.1 Prediction with mixture autoregressive models

Boshnakov (2009) showed that the h-steps ahead predictive distribution of MAR models can be derived analytically. For a mixture of g components, the density forecast at horizon h is a mixture of g^h components, which essentially accounts for every possible permutation of components up until time $t+h$. In particular, this holds true for any α-stable distribution, including Gaussian. Details on derivation can be found in Boshnakov (2009). Derivation of the same properties in the case of multivariate Normal mixtures is shown in Chapter 5.

2.5.2 Scoring rules

A scoring rule is a function that assigns a numerical value to a pair of a forecast distribution F and an observation y. In general, it is convention for scoring rules that a lower value denotes a better forecast.

Let F be the true distribution of y, and G be any distribution. A scoring rule $\mathcal{S}$ is a *proper scoring rule* if:

$$\mathrm{E}\left[\mathcal{S}\left(F,y\right)\right] \leq \mathrm{E}\left[\mathcal{S}\left(G,y\right)\right] \tag{2.38}$$

This means that, on expectation, the scoring rule will be minimised when the true distribution of y is used. In addition, $\mathcal{S}$ is a *strictly proper scoring rule* if equality in (2.38) only holds for $G \equiv F$ (i.e. there is not a distribution that can match the performance of F). We will make use of scoring rules in Chapter 5, to compare predictive performance of portfolios built with different model.

Continuous ranked probability score Continuous ranked probability score (CRPS) (see for instance Gneiting and Raftery, 2007) is one of the strictly proper scoring rules chosen to directly compare forecasting performance of different methods on the same dataset.

Given an observation x and the associated forecast distribution F, CRPS is defined mathematically as:

$$CRPS(F,x) = \int_{\mathbb{R}} \left(F(y) - I(y \geq x)\right)^2 dy \tag{2.39}$$

where $I(\cdot)$ is the indicator function assuming value 1 when the argument $y \geq x$ is true, and 0 othetwise. CRPS is a measure of discrepancy between the forecast CDF, F, and the empirical CDF of the observation x.

Logarithmic score Logarithmic score (LogS, Good, 1952) is another example of strictly proper scoring rule. Given an observation x, a score is assigned equal to the logarithm of the corresponding density:

$$\mathrm{LogS}\,(F,x) = \log f\,(x)\,. \tag{2.40}$$

Dawid-Sebastiani score The Dawid-Sebastiani score (DSS, Dawid and Sebastiani, 1999) is the last example of strictly proper scoring rule.

Given the mean μ and the variance σ^2 of the predictive distribution of a variable x,

DSS is calculated as

$$\mathrm{DSS}(F,x) = -\log\sigma^2 - \frac{1}{\sigma^2}(x-\mu)^2 \tag{2.41}$$

2.6 Review of some relevant probability distributions

2.6.1 Normal distribution

A random variable X is said to follow a Normal (or Gaussian) distribution with mean μ and variance σ^2 if its probability density function can be written as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}. \tag{2.42}$$

When $\mu = 0$ and $\sigma^2 = 1$, X is said to follow a standard Normal distribution.

Properties.

- Let $X \sim N(\mu,\sigma^2)$, then $Y = \dfrac{X-\mu}{\sigma} \sim N(0,1)$;
- Let $X_1,\ldots,X_n$ follow independent $N(\mu_i,\sigma_i^2)$, $i = 1,\ldots,n$. Then $Y = \sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$. This can be extended to correlated random variables, by adding the covariance terms;

2.6.2 Multivariate Normal distribution

This is a generalisation of the Normal distribution to dimensions higher than 1. A d-variate vector $\boldsymbol{X}$ is said to follow a multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ if its probability density function can be written as:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{2.43}$$

A useful property of the multivariate Normal distribution regards linear combinations of the random vector $\boldsymbol{X}$. Let $\boldsymbol{X} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{a}$ be a k-variate constant vector. For any $\boldsymbol{a}$, it holds that:

$$\sum_{i=1}^{k} a_i X_i = \boldsymbol{a}^T \boldsymbol{X} \sim N\left(\boldsymbol{a}^T \boldsymbol{\mu}, \boldsymbol{a}^T \boldsymbol{\Sigma}, \boldsymbol{a}\right) \tag{2.44}$$

2.6.3 Gamma distribution

A random variable X defined in the support $(0, +\infty)$ is said to follow a Gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$ if its probability density function ca be written as:

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp\left\{ -\beta x \right\}, \tag{2.45}$$

where $\Gamma(\cdot)$ is the gamma function. We denote $X \sim Ga(\alpha, \beta)$.

Properties .

- Let $X \sim Ga(\alpha, \beta)$. The expected value of X is $\mathrm{E}[X] = \frac{\alpha}{\beta}$; the variance of X is $\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$;
- If $X \sim Ga(1, \beta)$, then X follows an exponential distribution with parameter β, $X \sim Exp(\beta)$;
- If $\alpha > 1$, then the probability density function of X has a unique mode at $\frac{\alpha - 1}{\beta}$. If $\alpha \leq 1$, the distribution does not have a mode.
- Let $X_1, \ldots, X_n$ follow respectively $Ga(\alpha_i, \beta)$, $i = 1, \ldots, n$. Then, $\sum_{i=1}^{n} X_i \sim Ga\left(\sum_{i=1}^{n} \alpha_i, \beta\right)$

2.6.4 Student-t distribution

A continuous random variable X is said to follow a Student-t distribution (or simply t distribution) with $\nu > 0$ degrees of freedom if its probability density function can be written as:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}. \tag{2.46}$$

Provided certain conditions on ν are met, we have:

$$\begin{aligned} \mathrm{E}[X] &= 0, \quad \text{if } \nu > 1; \\ \mathrm{Var}(X) &= \frac{\nu}{\nu-2}, \quad \text{if } \nu > 2 \end{aligned}$$

The random variable X can be shifted and rescaled to obtain a t distribution with mean μ scale parameter σ^2 and again degrees of freedom ν. In this case, the probability density function becomes:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{\sigma}{\sqrt{\nu}} \left(1+\frac{(x-\mu)^2}{\sigma^2\nu}\right)^{-(\nu+1)/2} \tag{2.47}$$

Mean and variance of X change accordingly. We have in fact:

$$\begin{aligned} \mathrm{E}[X] &= \mu, \quad \text{if } \nu > 1; \\ \mathrm{Var}(X) &= \sigma^2\frac{\nu}{\nu-2}, \quad \text{if } \nu > 2 \end{aligned}$$

Integral representation of the t distribution In Bayesian statistics, a shifted and re-scaled version of the t distribution arises as marginalisation, with respect to the variance, of a Normal distribution with uknown mean and variance. Let X follow the t distribution with mean μ, scale parameter σ^2 and degrees of freedom ν in (2.47). It can

be shown that the probability density function of X arises as solution to the integral:

$$f_X(x) = \int_0^\infty f_{X|Z}(x \mid z) f_Z(z) dZ \tag{2.48}$$

where $X \mid Z = z \sim N\left(\mu, \frac{\sigma^2}{z}\right)$ and $Z \sim Ga\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$.

Notice that by adjusting the distribution of Z to a $Ga\left(\frac{\nu}{2}, \frac{\nu-2}{2}\right)$, the variance of X becomes $\mathrm{Var}(X) = \sigma^2$. This adjustment will be important to simulate samples directly from σ^2 (or sometimes σ^{-2}, as we will do in Chapter 4)with no need to perform any transformation.

2.6.5 Multinomial distribution

The multinomial distribution is a discrete distribution used to model n realisations of an event with k possible outcomes (for instance the roll of a die with 6 faces). In the particular case $k = 2$, we have the so called binomial distribution. Moreover, if $k = 2$ and $n = 1$ we have the Bernoulli distribution.

Let $\pi_1, \ldots, \pi_k$ be probabilities associated with k possible outcomes of an event, and $x_1, \ldots, x_k$ be the total number of realisations of each outcome, such that $\sum_{i=1}^k x_i = n$, where n is the number of experiments. Then the probability mass function for this distribution can be written as:

$$f(x_1 \ldots, x_n) = \frac{n!}{x_1!\, x_2!\, \ldots\, x_n!} \pi_1^{x_1} \pi_2^{x_2} \ldots \pi_k^{x_k} \tag{2.49}$$

2.6.6 Dirichlet distribution

The Dirichlet distribution is a family of continuous multivariate distributions, characterised by a parameter vector of positive real values $\boldsymbol{a} = (a_1, \ldots, a_k)$, $a_i > 0$ for all i. It is a generalisation of the beta distribution. In Bayesian statistics, the Dirichlet

distribution is a natural conjugate prior distribution for mixing weights in a mixture model.

A random vector $\boldsymbol{X} = (X1, \ldots, X_k)$ is said to follow a Dirichlet distribution if its probability density function can be written as:

$$f(\boldsymbol{x}) = \frac{\prod_{i=1}^{k} x_i^{a_i - 1}}{B(\boldsymbol{a})} \tag{2.50}$$

Conditions on $\boldsymbol{x} = (x_1, \ldots, x_k)$ are that $x_i > 0$ for all i and $\sum_{i=1}^{k} x_i = 1$.

2.7 GARCH models

Generalised autoregressive conditional heteroskedasticity (GARCH) models are a class of statistical models for time series in which the variance of the current error term, or innovation, is a function of the previous error terms and their respective variances. GARCH models find their application in financial time series, which often present periods of high variability followed by periods of low variability, as well as significant correlation in the square of the series. In fact, GARCH modelss are built under the assumption that the variance, or volatility, of an observation is depends upon the squares of previous innovations, as well as on past variances.

The traditional $GARCH(p,q)$ model introduced by Bollerslev (1986) as a generalisation of the autoregressive conditional heteroskedasticity (ARCH, Engle, 1982) is defined as follows:

$$\begin{aligned} \varepsilon_t &= \sigma_t \eta_t, \qquad \eta_t \sim N(0,1) \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \end{aligned} \tag{2.51}$$

where ε_t is the innovation at time t, η_t is a standard Normal random variable (but may be generalised to any white noise distribution), σ_t^2 is the variance, or volatility, of ε_t. A necessary restriction to the model parameters is $\alpha_i \geq 0$, $i = 0, \ldots, q$ and $\beta_j \geq 0$, $j = 1, \ldots, p$. If $p = 0$ the model reduces to an $ARCH(q)$ model, in which the variance is a linear combination of past variances; if $p = q = 0$ the process is white noise.

2.7.1 Multivariate GARCH models and Dynamic Conditional Correlation

A natural extension of GARCH models is their multivariate version. The advantage of a multivariate model is that it accounts for cross-correlation between different time series of interest, a feature that can be very useful with financial data, for example to build a portfolio of assets.

Bollerslev et al. (1988) and Engle and Kroner (1995) pioneered in the attempt to model conditional covariance matrices of predictors for multivariate time series with multivariate GARCH models, using different parametrisations known respectively as VEC and BEKK. Engle (2002) extended the idea of multivariate GARCH to the so called Dynamic Conditional Correlation models, in which each element of the time-dependent covariance matrix of the data is modelled to follow a GARCH process. Such models have computational advantages over multivariate GARCH models in that the number of parameters to be estimated in the correlation process is independent of the number of series to be correlated, by use of common parameters across all correlations.

Since then, much work has been done to develop multivariate GARCH models, with various applications in finance and econometrics. Of particular interest, attempts have been made in combining GARCH and factor models, with the aim of dimensionality reduction when modelling large portfolios or panel data. These models rely on the assumption that financial returns are described by a small number of underlying

common variables, or factors, which can be used to model the data more parsimoniously. Although all equal in concept, different approaches used different assumptions on such factors, and different techniques are used to derive them. For instance, Alexander (2000) uses a principal components analysis in which factors are assumed to follow independent GARCH processes, whereas Van der Weide (2002) considers the case in which factors are not orthogonal. Finally, Santos and Moura (2014) introduced the dynamic factor GARCH model with time-varying factor loadings.

We focus here on the DCC model by Engle (2002), which directly compares to our MAR model and its multivariate extension.

A multivatiate GARCH (MGARCH) model for a m-variate time series process $\boldsymbol{y}_t$ can be written as:

$$\begin{aligned} \boldsymbol{y_t} &= E\left[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}\right] + \boldsymbol{\varepsilon}_t \\ \operatorname{Var}\left(\boldsymbol{\varepsilon}_t\right) &= H_t \end{aligned} \tag{2.52}$$

where H_t is a positive definite matrix for all t. H_t can be decomposed as $H_t = D_t R_t D_t$, where D_t is a diagonal matrix with elements $(\sigma_{1t}, \ldots, \sigma_{mt})$ and R_t is a time-dependent correlation matrix, with the same requirements as H_t and with diagonal elemements equal to 1. In the standard MGARCH model, it is instead assumed that R_t is constant at all times t.

These equations will produce a correlation matrix at each time point. Engle (2002) suggests to specify each element of the matrix R_t by a univariate GARCH model. Let $r_{i,j,t}$ be a generic element of the correlation matrix R_t. $r_{i,j,t}$ is assumed to follow a $GARCH(1,1)$ process:

$$r_{i,j,t} = \bar{\rho}_{i,j} + \alpha\left(\varepsilon_{i,t-1}\varepsilon_{j,t-1} - \bar{\rho}_{i,j}\right) + \beta\left(r_{i,j,t} - \bar{\rho}_{i,j}\right) \tag{2.53}$$

where $\bar{\rho}_{i,j}$ is the unconditional correlation between $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$. Notice that the average

of $r_{i,j,t}$ will be $\bar{\rho}_{i,j}$, and the average of the variances will be 1. The estimator of the conditional correlation is ultimately

$$\hat{\rho}_{i,j,t} = \frac{r_{i,j,t}}{\sqrt{r_{i,i,t} r_{j,j,t}}} \tag{2.54}$$

2.8 Modern portfolio theory and financial risk

2.8.1 Modern portfolio theory

Modern portfolio theory (MPT, Markowitz, 1952) is a theory on how to construct portfolios to maximise expected return on a given level of risk. Likewise, it can minimise the risk for a given expected return.

MPT assumes that investors are adverse to risk, meaning that, for a given return, they prefer a lower level of risk. This is achieved by investing on multiple assets or asset classes, rather than on a single asset.

The expected return on a portfolio is calculated as a weighted average of asset returns, where the weights are the proportion of invested capital placed on each assets. The risk associated with a portfolio is calculated as a function of the variances of the assets and pairwise correlations. Furthermore, MPT in general allows short selling, which is reflected in negative values for the portfolio weights.

Terminology:

- **Short selling:** short selling consists in the investor borrowing shares of an asset which they believe will decrease in value by a future date. In this case, the investor sells the borrowed shares, which they will purchase back in the future and return to the lender. If the price of the shares has decreased over this period of time, the investor will buy back for a lower price than what they have previously sold, hence making a profit.

- **Efficient portfolio:** for a given target return μ, the portfolio of assets that has the lowest risk among all portofolios of the same assets with same target return μ is called *efficient portfolio*. We denote quantities related to an efficient portfolio with the subscript EFF.

- **Minimum variance portfolio:** The portfolio of assets with the lowest variance of all portfolios built with those same assets is called *minimum variance portfolio*. We denote quantities related to a minimum variance portfolio with the subscript MVP.

Let $\boldsymbol{y}_t$ be a multivariate time series of m financial assets. For simplicity, we assume for now that $\boldsymbol{y}_t$ is second order stationary, such that $\mathrm{E}[\boldsymbol{y}_t] = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{y}_t) = \boldsymbol{\Omega}$.

Suppose now that we would like to build a portfolio of these assets, and predict the expected portfolio return and the risk associated with it at the next time point $t+1$. Let w denote portfolio weights, $R_{t+1} = w^T \boldsymbol{y}_{t+1}$ denote the portfolio return at $t+1$, and

$$A = \mathbb{1}\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} \quad , \quad B = \boldsymbol{\mu}\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} \quad , \quad C = \mathbb{1}\boldsymbol{\Omega}^{-1}\mathbb{1} \quad , \quad D = CB - A^2 \tag{2.55}$$

where $\mathbb{1}$ is a vector of 1s of the same length as μ.

It can be proved that optimal weights for an efficient portfolio of these assets and target return μ_{EFF} are

$$w_{\mathrm{EFF}} = \frac{1}{D}\left(B\boldsymbol{\Omega}^{-1}\mathbb{1} - A\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^*\left(C\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} - A\boldsymbol{\Omega}^{-1}\mathbb{1}\right)\right) \tag{2.56}$$

Consequently, the expected return of that efficient portfolio is $R_{t+1} = w_{\mathrm{EFF}}^T \hat{\boldsymbol{y}}_{t+1}$, $\hat{\boldsymbol{y}}_{t+1}$ being the prediction of asset prices at $t+1$. The variance of such portfolio is $w_{EFF}^T \boldsymbol{\Omega} w_{EFF}$. Weights of the minimum variance portfolio of same assets $\{\boldsymbol{y}_t\}$, and

corresponding return, are:

$$w_{\text{MVP}} = \frac{\boldsymbol{\Omega}^{-1}\mathbb{1}}{C} \qquad \boldsymbol{\mu}_{\text{MVP}} = \frac{A}{C} \tag{2.57}$$

Notice that this methodology allows for weights w to assume negative values. An asset with a negative weight associated to it indicates that short-selling is applied to that asset.

2.8.2 Financial risk measures

A financial risk measure gives the probability associated with an estimated loss. Formally, financial risk measures are a class ρ of random variables in the support $(0,1)$ which satisfy the following properties:

1. **Normalised property:** $\rho(0) = 0$; this property states that there is no risk if no assets are held.

2. **Translative property:** let c be a real constant. Then $\rho(X+c) = \rho(X) + c$.

3. **Monotone property:** $X \leq Y \Rightarrow \rho(X) \geq \rho(Y)$; this property states that if portfolio Y always has better returns than portfolio X, then the risk associated with portfolio Y is always lower than that associated with portfolio X.

Furthermore, a risk measure is called a *coherent risk measure* if it satisfies two additional properties:

4. **Positive homogeneity:** let c be a real constant. Then $\rho(cX) = c\rho(X)$.

5. **Sub-additivity:** $\rho(X+Y) = \rho(X) + \rho(Y)$

We here introduce two of the most popular measures to assess risk associated with a financial investment, *value at risk* (VaR) and *expected shortfall* (ES).

Value at risk. Value at risk estimates how much an investment might lose, with a given probability p, in normal market conditions. Informally, $\text{VaR}_p(x) = \theta$ means that an investment x is likely to result in a loss of at least θ units with probability p in a set time window. Let X be a random variable desciribing some financial returns, such that negative values of X represent a loss and positive values represent a profit. The value at risk at level p is:

$$\text{VaR}_p(X) = -\inf\{x \in \mathcal{R}, F_X(x) = 1-p\} = -F_X^{-1}(1-p) \ . \tag{2.58}$$

What (2.58) tells us is that the value at risk θ associated with X at level p is minus the $1-p$ quantile of the distribution of X.

When the distribution of X is known and can be inverted, VaR can be calculated analytically. Otherwise, nonparametric methods are available. We will see in Chapter 5that for MAR models the distribution of interest is available, and quantiles can be estimated with analytical solutions.

Value at risk is not a coherent risk measure, as it does not satisfy the sub-additive property. However it still satisfies properties 1-4, plus the following:

- $\text{VaR}_p(X) = \text{VaR}_{1-p}(-X)$
- $X \leq 0 \Rightarrow \text{VaR}_p(X) \geq 0$

Expected shortfall. Expected shortfall is strictly related to value at risk. Broadly speaking, ES estimates the expected loss on an investment, assuming that a loss larger than VaR will be recorded. Expected shortfall is sometimes preferred to value at risk since it is more sensitive about the tails of the distribution, and most importantly because is a coherent risk measure.

Let X be an absolutely continuous random variable describing some financial returns, such that negative values of X represent a loss and positive values represent a

profit. The expected shortfall at associated with X at level p is obtained by solving the integral:

$$\mathrm{ES}_p(X) = \frac{1}{p}\int_0^p \mathrm{VaR}_\alpha(X)\, d\alpha \tag{2.59}$$

Expected shortfall can hence be calculated analytically. However, analytical solutions to such integral may not be straightforward, as they depend on the distribution assumption. For this reason, numerical integration methods such as Monte Carlo are sometimes preferred.

Chapter 3

Bayesian analysis of mixture autoregressive models covering the complete parameter space

Mixture autoregressive (MAR) models (Wong and Li, 2000) provide a flexible way to model time series with predictive distributions which depend on the recent history of the process. Not only do the predictive distributions change over time, but they are also different for different horizons for predictions made at a fixed time point. As a result, they inherently accommodate asymmetry, multimodality and heteroskedasticity. For this reason, mixture autoregressive models have been considered a valuable alternative to other models for time series, such as the SETAR model (Tong, 1990), the Gaussian transition mixture distribution model (Le et al., 1996), or the widely used class of GARCH models (Nelson, 1991). Another useful feature of MAR models is that they model jointly the conditional mean and autocovariance. Moreover, the autocovariances are zero on a subspace of the parameters. So, if an uncorrelated (weak white noise) model is required, as is often the case for financial time series, the parameters can be restricted to that subspace.

MAR models can be thought of as random coefficient autoregressive models (Boshnakov, 2011). Similarly to the usual autoregressions, there is a stationarity region for the parameters, outside which the MAR models are explosive and thus not generally useful.

Wong and Li (2000) considered estimation of MAR models based on the EM algorithm (Dempster et al., 1977). That method is particularly well suited for mixture-type models and works well. On the other hand, a Bayesian approach can offer the advantage of incorporating the uncertainty in the estimated models into the predictions.

Sampietro (2006) presented the first Bayesian analysis of MAR models. In his work, reversible jump MCMC (Green, 1995) is used to select the autoregressive orders of the components in the mixture, and models with different number of components are compared using methods by Chib (1995) and Chib and Jeliazkov (2001), which exploit the marginal likelihood identity. In addition, he derives analytically posterior distributions for all parameters in the selected model.

The Bayesian updates of the autoregressive parameters are problematic, because the parameters need to be kept in the stationarity region, which is very complex, and so cannot really be updated independently of each other. In the case of autoregressive (AR) models, it is routine to use parametrisation in terms of partial autocorrelations (Jones, 1987), which are subject only to the restriction to be in the interval $(-1, 1)$. Sampietro (2006) adapted this neatly to MAR models by parameterising the autoregressive parameters of each component of the MAR model with the partial autocorrelations of an AR model with those parameters.

A major drawback of Sampietro's sampling algorithm for the autoregressive parameters, is that it restricts the parameters of each component to be in the stationarity region of an autoregressive model. While this guarantees that the MAR model is stationary, it excludes from consideration considerable part of the stationarity region of

the MAR model (Wong and Li, 2000, p. 98; Boshnakov, 2011). Depending on the mixture probabilities, the excluded part can be substantial. For example, most examples in Wong and Li (2000, p. 98) cannot be handled by Sampietro's approach, see also the examples in Section 3.3.

Lau and So (2008) proposed an infinite mixture of autoregressive models and used a semi-parametric approach based on a Dirichlet process (Ferguson, 1973) and the so called Gibbs version of the weighted Chinese restaurant process (Lo, 2005) to select the optimal number of mixture components and assign observations to those. However, they do not assess conditions for second order stationarity of the model. Wood et al. (2011) used data segmentation for estimation of a variant of the MAR models—they divide the data into segments and assign each segment to one mixture component. Their approach is aimed at time series which are piecewise autoregressions (for example as a result of structural changes), has a different field of applications, and is not directly comparable to the MAR model considered here.

Hossain (2012) developed a full analysis (model selection and sampling), which reduced the constraints of Sampietro's analysis. Using Metropolis-Hastings algorithm and a truncated Gaussian proposal distribution for the moves, he directly simulated the autoregressive parameters from their posterior distribution. This method still imposes a constraint on the autoregressive parameters through the choice of boundaries for the truncated Gaussian proposal. While the truncation is used to keep the parameters in the stationarity region, the choice of boundaries is arbitrary and can leave out a substantial part of the stationarity region of the model. In addition, his reversible jump move for the autoregressive order seems conservative, as it uses functions which always prefer jumps towards low autoregressive orders (this will be seen in Section 3.2.5).

A common problem associated with mixtures is label switching (see for instance Celeux, 2000), which derives from symmetry in the likelihood function. If no prior information is available to distinguish components in the mixture, then the posterior

distribution will also be symmetric. It is essential that label switching is detected and handled properly in order to obtain meaningful results. A common way to deal with this, also used by Sampietro (2006) and Hossain (2012), is to impose identifiability constraints. However, it is well known that such constraints may lead to bias and other problems. In the case of MAR models, Hossain (2012) showed that these constraints may affect convergence to the posterior distribution.

We develop a new procedure which resolves the above problems. We propose an alternative Metropolis-Hastings move to sample directly from the posterior distribution of the autoregressive components. Our method covers the complete parameter space. We also propose a way of selecting optimal autoregressive orders using reversible jump MCMC for choosing the autoregressive order of each component in the mixture, which is less conservative than that of Hossain. We propose the use of a relabelling algorithm to deal a posteriori with label switching.

We apply the new method to both simulated and real datasets, and discuss the accuracy and performance of our algorithm, as well as its advantages over previous studies. Finally, we briefly introduce the idea of density forecasting using MCMC output.

The structure of this chapter is as follows. In Section 3.1 we review the mixture autoregressive model and somer relevant notation. In Section 3.2 we give detailed description of our method for Bayesian analysis of MAR models, including model selection, full description of the sampling algorithm, and the relabelling algorithm to deal with label switching. Section 3.3 shows results from application of our method to simulated and real dataset. Section 3.4 introduces the idea of density forecast using MCMC output.

3.1 The mixture autoregressive model

Recall the conditional cumulative distribution function of the mixture autoregressive model in (2.1)

$$F(y_t|\mathcal{F}_{t-1},\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k F_k\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k}\phi_{ki}y_{t-i}}{\sigma_k}\right),$$

where $\boldsymbol{\theta}$ is the vector of model parameters.

We now introduce some notation. Let

$$\mu_{tk} = \phi_{k0} + \sum_{i=1}^{p_k}\phi_{ki}y_{t-i}.$$

The error term associated with the kth component at time t is defined by

$$e_{tk} = y_t - \phi_{k0} - \sum_{i=1}^{p_k}\phi_{ki}y_{t-i} = y_t - \mu_{tk}. \tag{3.1}$$

A useful alternative expression for μ_{tk} is the following mean corrected form:

$$\mu_{tk} = \mu_k + \sum_{i=1}^{p_k}\phi_{ki}\left(y_{t-i} - \mu_k\right).$$

Comparing the two representations we get

$$\phi_{k0} = \mu_k\big(1 - \sum_{i=1}^{p_k}\phi_{ki}\big).$$

If $\sum_{i=1}^{p_k}\phi_{ki} \neq 0$, we also have

$$\mu_k = \frac{\phi_{k0}}{1 - \sum_{i=1}^{p_k}\phi_{ki}}. \tag{3.2}$$

A nice feature of this model is that the one-step predictive distributions are given directly by the specification of the model with (2.1). The h-steps ahead predictive

distributions of y_{t+h} at time t can be obtained by simulation (Wong and Li, 2000) or, in the case of Gaussian and α-stable components, analytically (Boshnakov, 2009).

We focus here on mixtures of Gaussian components. In this case, using the standard notations $\mathbf{\Phi}$ and $\boldsymbol{\phi}$ for the CDF and PDF of the standard Normal distribution, we have $F_k \equiv \mathbf{\Phi}$ and $f_k \equiv \boldsymbol{\phi}$, for $k = 1, \ldots, g$. The model in (2.1) can hence be written as

$$F(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \mathbf{\Phi}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k}\right) \tag{3.3}$$

or, alternatively, in terms of the conditional pdf

$$f(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \frac{\pi_k}{\sigma_k} \boldsymbol{\phi}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k}\right) \tag{3.4}$$

Conditional mean and variance of Y_t, respectively top and bottom of (3.5), are:

$$\begin{aligned} \mathrm{E}[y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}] &= \sum_{k=1}^{g} \pi_k \left(\phi_{k0} + \sum_{i=1}^{p} \phi_{ki} y_{t-i}\right) = \sum_{k=1}^{g} \pi_k \mu_{tk} \\ \mathrm{Var}(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) &= \sum_{k=1}^{g} \pi_k \sigma_k^2 + \sum_{k=1}^{g} \pi_k \mu_{tk}^2 - \sum_{k=1}^{g} (\pi_k \mu_{tk})^2 \end{aligned} \tag{3.5}$$

Notice that $\sum_{k=1}^{g} \pi_k \mu_{tk}^2 - \sum_{k=1}^{g} (\pi_k \mu_{tk})^2 \geq 0$, with equality for $\mu_{t1} = \mu_{t2} = \cdots = \mu_{tg}$. Consequently, $\mathrm{Var}(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) \geq \sum_{k=1}^{g} \pi_k \sigma_k^2$.

The correlation structure of a stable MAR process with maximum order p is similar to that of an $AR(p)$ process. At lag h we have:

$$\rho_h = \sum_{k=1}^{g} \pi_k \sum_{i=1}^{p} \phi_{ki} \rho_{|h-i|} = \sum_{i=1}^{p} \left(\sum_{k=1}^{g} \pi_k \phi_{ki}\right) \rho_{|h-i|} \qquad h \geq 1.$$

Setting $a_i = \left(\sum_{k=1}^{g} \pi_k \phi_{ki}\right)$ for $i = 1, \ldots, p$, we see that these are analogous to the Yule-Walker equations for an $AR(p)$ model. See Wong (1998) for more details. Notice that if the stability condition is not satisfied, meaning some of the roots of the matrix A in

Section 2.1.2 lie outside the unit circle, then the solution of this recurrence equation is not the autocorrelation function, which contradicts the assumption of stationarity.

The conditional mean in (3.5) may be written in similar fashion to resemble that of an $AR(p)$. Setting $\sum_{k=1}^{g} \pi_k \phi_{k0} = c$, we have:

$$\mathrm{E}[y_t \mid \mathcal{F}_{t-1}] = \sum_{k=1}^{g} \pi_k \phi_{k0} + \sum_{i=1}^{p} \left(\sum_{k=1}^{g} \pi_k \phi_{ki} \right) y_{t-i} = c + \sum_{i=1}^{p} a_i y_{t-i}. \tag{3.6}$$

While it is true that the linear predictor of y_t of the MAR model, calculated as the conditional expectation, is analogous to that of an AR(p) model with corresponding autoregressive parameters $a_1, \ldots, a_g$, one would be mistaken in assuming that the two models are analogous. In fact, the conditional variance, in general, is different between the two models, so that the variance on the prediction, if we assumed an AR(p) model, would be underestimated, as it does not account for component-specific variabilities.

3.2 Bayesian analysis of mixture autoregressive models

3.2.1 Likelihood function and missing data formulation

Given data $y_1, \ldots, y_n$, the likelihood function for the MAR model in the case of Gaussian mixture components takes the form of (3.4)

$$L(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\pi} | \boldsymbol{y}) = \prod_{t=p+1}^{n} \sum_{k=1}^{g} \frac{\pi_k}{\sigma_k} \boldsymbol{\phi} \left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k} \right).$$

The likelihood function is not very tractable and a standard approach is to resort to a missing data formulation (Dempster et al., 1977).

Let $\boldsymbol{Z}_t = (Z_{t1}, \ldots, Z_{tg})$ be a latent allocation random variable, where $\boldsymbol{z}_t$ is a g-dimensional vector with entry k equal to 1 if y_t comes from the k^{th} component of the mixture, and 0 otherwise. We assume that the $\boldsymbol{Z_t}$s are discrete random variables,

independently drawn from the discrete distribution:

$$P(z_{tk} = 1|g, \boldsymbol{\pi}) = \pi_k, \qquad k = 1, \ldots, g. \tag{3.7}$$

This setup, widely exploited in the literature (see, for instance Dempster et al., 1977; Diebolt and Robert, 1994) allows to rewrite the likelihood function in a much more tractable way as follows:

$$L(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\pi}|\boldsymbol{y}) = \prod_{t=p+1}^{n} \prod_{k=1}^{g} \left(\frac{\pi_k}{\sigma_k} \phi \left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k} \right) \right)^{z_{tk}} \tag{3.8}$$

In practice, the $\boldsymbol{z}_t$s are not available. We adopt a Bayesian approach to deal with this. We set suitable prior distributions on the latent variables and the parameters of the model and develop a methodology for obtaining posterior distributions of the parameters and dealing with other issues arising in the model building process.

3.2.2 Priors setup and choice of hyperparameters

The setup of prior distributions is based on Sampietro (2006) and Hossain (2012). In the absence of any relevant prior information it is natural to assume a priori that each data point is equally likely to be generated from any component, i.e. $\pi_1 = \cdots = \pi_g = 1/g$. This is a discrete uniform distribution, which is a particular case of the multinomial distribution. The conjugate prior of the latter is the Dirichlet distribution. We therefore set the prior for the mixing weigths vector, $\boldsymbol{\pi}$, to

$$\boldsymbol{\pi} \sim D(w_1, \ldots, w_g), \qquad w_1 = \cdots = w_g = 1. \tag{3.9}$$

The prior distribution on the component means is a normal distribution with common fixed hyperparameters ζ for the mean and κ for the precision, i.e.

$$\mu_k \sim N(\zeta, \kappa^{-1}), \qquad k = 1, \ldots, g. \tag{3.10}$$

For the component precisions, τ_k, a hierarchical approach is adopted, as suggested in Richardson and Green (1997). Here, for a generic k^{th} component the prior is a Gamma distribution with hyperparameters c (fixed) and λ, which itself follows a gamma distribution with fixed hyperparameters a and b. We have therefore

$$\begin{aligned} c &- \text{fixed} \\ \lambda &\sim Ga(a,b) \\ \tau_k \mid \lambda &\sim Ga(c,\lambda), \qquad k = 1, \ldots, g. \end{aligned} \tag{3.11}$$

The main difference between our approach and that of Sampietro (2006) and Hossain (2012) is in the treatment of the autoregressive parameters.

Sampietro (2006) exploits the one-to-one relationship between partial autocorrelations and autoregressive parameters for autoregressive models described in Jones (1987). Namely, he parameterises each MAR component with partial autocorrelations, draws samples from the posterior distribution of the partial autocorrelations via Gibbs-type moves and converts them to autoregressive parameters using the functional relationship between partial autocorrelations and autoregressive parameters. Of course, the term "partial autocorrelations" does not refer to the actual partial autocorrellations of the MAR process, they are simply transformed parameters. The advantage of this procedure is that the stability region for the partial autocorrelation parameters is just a hyper-cube with marginals in the interval $(-1, 1)$, while for the AR parameters it is a body whose boundary involves non-linear relationships between the parameters.

A drawback of the partial autocorrelations approach in the MAR case is that it covers only a subset of the stability region of the model. Depending on the other parameters, the loss may be substantial.

Hossain (2012) overcomes the above drawbacks by simulating the AR parameters directly. He uses Random Walk Metropolis, while applying a constraint to the proposal distribution (a truncated Normal). The truncation is chosen as a compromise that ensures that an arbitrarily large part of the stability region is covered, while keeping a reasonable acceptance rate. Although effective with "well behaved" data, there are scenarios, especially concerning financial examples, in which the loss of information due to a pre-set truncation becomes significant, as will be shown later on.

If, by use of a truncation, the true values of the parameters are excluded, the Markov Chain may converge towards the boundary of the constrained parameter space, and the resulting posterior distributions of such parameters would be misleading. Suppose for instance that the true value of a generic autoregressive parameter of a MAR model, ϕ_{ki}, was 1.2. If we decided to constrain that parameter in the interval $(-1,1)$, to ensure stability of the model, then the posterior distribution of ϕ_{ki} would be pushed towards the upper boundary of the interval, with its peak very close to 1. In conlcusion, we cannot safely choose the stability region of a MAR model beforehand, unless we know or assume that the specified model is correct.In this paper, we choose Random Walk Metropolis for simulation from the posterior distribution of autoregressive parameters, while exploiting the stability condition to avoid restraining the parameter space a priori.

With the above considerations, for the autoregressive parameters we choose a multivariate uniform distribution with range in the stability region of the model, and independence between parameters is assumed. Hence, for the parameter vector $\boldsymbol{\phi}$ prior distribution is such that:

$$p(\boldsymbol{\phi} \mid \boldsymbol{\pi}) \propto I\{Stable\}, \quad k = 1, \ldots, g.$$

where I denotes the indicator function assuming value 1 if the condition is satisfied and 0 otherwise (see section 2.1.2 for details on stability of MAR models). In other words, what we propose is a flat (uniform) prior over the stability region of the model. This uniform prior allows for better exploration of the parameter space than a Normal prior and doesn't mask multimodality.

Choice of hyperparameters. Here we discuss the settings for the hyperparameters ζ, κ, a, b, and c. We have already discussed that the hyperparameters for the Dirichlet prior distribution on the mixing weights (all equal to 1). Also, λ is a hyperparameter but it is a random variable with distribution which will be fully specified once a and b are.

Following Richardson and Green (1997), let $\mathcal{R}_y = \max(y) - \min(y)$ be the length of the interval variation of the dataset. Also fix the two hyperparameters $a = 0.2$ and $c = 2$. The remaining hyperparameters are set as follows:

$$\zeta = \min(y) + \frac{\mathcal{R}_y}{2} \qquad \kappa = \frac{1}{\mathcal{R}_y} \qquad b = \frac{100a}{c\mathcal{R}_y^2} = \frac{10}{\mathcal{R}_y^2}$$

3.2.3 Posterior distributions and acceptance probability for RWM

Following Sampietro (2006) and Hossain (2012), we derive here posterior distributions for all but the autoregressive parameters. For $k = 1, \ldots, g$, define the useful quantities:

$$e_{tk} = y_t - \nu_{tk}, \qquad n_k = \sum_{t=p+1}^{n} z_{tk}, \qquad b_k = 1 - \sum_{i=1}^{p_k} \phi_{ki}, \qquad \bar{e}_k = \frac{1}{n_k} \sum_{t=p+1}^{n} e_{tk} z_{tk}.$$

Simulation of the latent variables The latent variables Z_t are updated using Bayes Theorem. We can derive the posterior density of z_t to be

$$\begin{aligned} p(z_t|y_t,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\tau},\lambda) &\propto L(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\tau},\lambda|y_t,z_t)p(z_t|y_t,\boldsymbol{\pi}) \\ &\propto \sum_{k=1}^{g} \pi_k \, \boldsymbol{\phi}\left(\frac{e_{tk}}{\sigma_k}\right) \mathbb{I}\{z_t = k\} \end{aligned} \tag{3.12}$$

It follows that posterior probability of an observation at time t to be generated by the k^{th} component is

$$P(z_t = k|y_t,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\tau},\lambda) = \frac{\pi_k \, \boldsymbol{\phi}\left(\frac{e_{tk}}{\sigma_k}\right)}{\sum_{l=1}^{g} \pi_l \, \boldsymbol{\phi}\left(\frac{e_{tl}}{\sigma_l}\right)} \tag{3.13}$$

Posterior distribution of $\boldsymbol{\pi}$ Prior distribution for the mixing weight was chosen to be $\boldsymbol{\pi} \sim D(w_1,\ldots,w_g)$,

$w_1 = \cdots = w_g = 1$. The posterior distribution is consequently

$$\begin{aligned} p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\mu},\boldsymbol{\tau},\lambda) &\propto p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\pi})\, p(\boldsymbol{\pi}) \\ &\propto p(\boldsymbol{z}|\boldsymbol{\pi})\, p(\boldsymbol{\pi}) \\ &\propto \prod_{k=1}^{g} \pi_k^{n_k} \prod_{k=1}^{g} \pi_k^{1-1} \\ &\propto \prod_{k=1}^{g} \pi_k^{(n_k+1)-1} \end{aligned} \tag{3.14}$$

where $n_k = \sum_{t=p+1}^{n} z_{tk}$.

Thus

$$\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\mu},\boldsymbol{\tau} \sim D(1+n_1,\ldots,1+n_g) \tag{3.15}$$

Posterior distribution of $\boldsymbol{\mu}$ Prior distribution is identical for every component mean

$$\mu_k \sim N(\zeta,\kappa^{-1}),\ k = 1,\ldots,g$$

Choice of hyperparameters is discussed in Section 3.1.

Posterior distribution for the k^{th} component becomes

$$
\begin{aligned}
p(\mu_k|\boldsymbol{y},\boldsymbol{z},\boldsymbol{\mu}_{-k},\boldsymbol{\phi},\boldsymbol{\pi},\boldsymbol{\tau},\lambda) &\propto \prod_{\substack{t=p+1\\ z_t=k}}^{n} p\left(y_t,z_t|\mu_k\right) p\left(\mu_k\right) \\
&\propto \exp\left[-\frac{\tau_k}{2}\sum_{\substack{t=p+1\\ z_t=k}}^{n}\left(y_t-\mu_k-\sum_{i=1}^{p_k}\phi_{ki}(y_{t-1}-\mu_k)\right)^2\right]\times\exp\left[-\frac{\kappa}{2}\left(\mu_k-\zeta\right)^2\right] \\
&= \exp\left(-\frac{\tau_k}{2}\sum_{\substack{t=p+1\\ z_t=k}}^{n}\left[\left(y_t-\sum_{i=1}^{p_k}\phi_{ki}y_{t-1}\right)-\mu_k\left(1-\sum_{i=1}^{p_k}\phi_{ki}y_{t-1}\right)\right]^2\right) \\
&\times\exp\left[-\frac{\kappa}{2}\left(\mu_k-\zeta\right)^2\right] \\
&= \exp\left(-\frac{\tau_k}{2}\sum_{\substack{t=p+1\\ z_t=k}}^{n}\left(e_{kt}-\mu_k b_k\right)^2\right)\times\exp\left[-\frac{\kappa}{2}\left(\mu_k-\zeta\right)^2\right] \\
&= \exp\left(-\frac{\tau_k}{2}\left[\sum_{\substack{t=p+1\\ z_t=k}}^{n}\left(e_k t-\bar{e}_k\right)^2+n_k\left(\bar{e}_k-\mu_k b_k\right)^2\right]\right)\times\exp\left[-\frac{\kappa}{2}\left(\mu_k-\zeta\right)^2\right] \\
&\qquad\qquad \text{where } \bar{e}_k=\frac{1}{n_k}\sum_{\substack{t=p+1\\ z_t=k}}^{n} e_{kt} \\
&\propto \exp\left[-\frac{n_k\tau_k}{2}\left(\bar{e}_k-\mu_k b_k\right)^2-\frac{\kappa}{2}\left(\mu_k-\zeta\right)^2\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\tau_k n_k b_k^2+\kappa\right)\mu_k^2+\left(\tau_k n_k \bar{e}_k b_k+\kappa\zeta\right)\mu_k\right]
\end{aligned}
\tag{3.16}
$$

[]

Hence, posterior distribution for the mean of the k^{th} component is

$$\mu_k|\boldsymbol{y},\boldsymbol{z},\boldsymbol{\mu}_{-k},\boldsymbol{\phi},\boldsymbol{\pi},\boldsymbol{\tau},\lambda \sim N\left(\frac{\tau_k n_k \bar{e}_k b_k+\kappa\zeta}{\tau_k n_k b_k^2+\kappa},\frac{1}{\tau_k n_k b_k^2+\kappa}\right)$$

using the fact that $\exp\left(A\mu^2+2B\mu_k\right)\propto N\left(\mu;\frac{B}{A},\frac{1}{A}\right)$

Posterior distribution of λ and $\boldsymbol{\tau}$

We start by deriving the posterior distribution of λ:

$$\begin{aligned} p(\lambda|\boldsymbol{y},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\tau}) &\propto p(\boldsymbol{\tau}|\lambda)\, p(\lambda) \\ &\propto \prod_{k=1}^{g} \lambda^{c} \exp(-\lambda\tau_k) \times \lambda^{a-1} \exp(-b\lambda) \\ &\propto \lambda^{a+gc-1} \exp\left(-(b+\sum_{k=1}^{g}\tau_k)\lambda\right) \end{aligned} \tag{3.17}$$

And thus

$$\lambda|\boldsymbol{y},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\tau} \sim Ga\left(a+gc, b+\sum_{k=1}^{g}\tau_k\right)$$

Secondly, we can derive the posterior distribution for the precision of the k^{th} component

$$\begin{aligned} p(\tau_k|\boldsymbol{y},\boldsymbol{z},\boldsymbol{\tau}_{-k},\boldsymbol{\phi},\boldsymbol{\pi},\boldsymbol{\mu},\lambda,) &\propto p(\boldsymbol{y},\boldsymbol{z}|\tau_k)p(\tau_k) \\ &\propto \tau_k^{\frac{n_k}{2}} \exp\left(-\frac{\tau_k}{2}\sum_{\substack{t=p+1 \\ z_t=k}}^{n} e_{kt}^2\right) \times \tau_k^{c-1}\exp(-\lambda\tau_k) \\ &= \tau_k^{\left(c+\frac{n_k}{2}\right)-1} \exp\left[-\left(\lambda+\frac{1}{2}\sum_{\substack{t=p+1 \\ z_t=k}}^{n} e_{kt}^2\right)\tau_k\right] \end{aligned} \tag{3.18}$$

Thus

$$\tau_k|\boldsymbol{y},\boldsymbol{z},\boldsymbol{\tau}_{-k},\boldsymbol{\phi},\boldsymbol{\pi},\boldsymbol{\mu},\lambda \sim Ga\left(c+\frac{n_k}{2}, \lambda+\frac{1}{2}\sum_{\substack{t=p+1 \\ z_t=k}}^{n} e_{kt}^2\right)$$

Exploiting the fact that $\tau^{A-1}\exp(-B\tau) \propto Ga(\tau;A,B)$. All these parameters are updated via a Gibbs-type move.

Similarly, $\boldsymbol{z}_t$s are simulated from a multinomial distribution with associated posterior probabilities.

Update of $\boldsymbol{\phi}$

To update autoregressive parameters, let $\boldsymbol{\phi}_k$, $k = 1, \ldots, g$, be the set of current states of the autoregressive parameters, i.e. a set of observations from the posterior distribution of $\boldsymbol{\phi}_k$. We can simulate $\boldsymbol{\phi}_k^*$ from a proposal $MVN(\boldsymbol{\phi}_k, \Gamma_k^{-1})$ distribution, denoted by $q(\boldsymbol{\phi}_k^*, \boldsymbol{\phi}_k)$, with $\Gamma_k = \gamma_k I_{p_k}$, where I_{p_k} is the identity matrix of size p_k.

Here γ_k, $k = 1, \ldots, g$ is a tuning parameter, chosen in such way that the acceptance rate of RWM is optimal $(20 - 25\%)$ for component k. We allow γ_k to change between components, but to be constant within the same component. Notice the difference between our proposal and the two-step approach by Sampietro (2006), or the truncated Normal proposal chosen by Hossain (2012). The probability of accepting a move to the proposed $\boldsymbol{\phi}_k^*$ is

$$\alpha(\boldsymbol{\phi}_k, \boldsymbol{\phi}_k^*) = \min\left\{1, \frac{f(\boldsymbol{y} \mid \boldsymbol{\phi}_k^*)\, p(\boldsymbol{\phi}_k^*) q(\boldsymbol{\phi}_k, \boldsymbol{\phi}_k^*)}{f(\boldsymbol{y} \mid \boldsymbol{\phi}_k)\, p(\boldsymbol{\phi}_k) q(\boldsymbol{\phi}_k^*, \boldsymbol{\phi}_k)}\right\}, \tag{3.19}$$

where $q(\boldsymbol{\phi}_k, \boldsymbol{\phi}_k^*) = q(\boldsymbol{\phi}_k^*, \boldsymbol{\phi}_k)$, due to the symmetry in the Normal proposal. Therefore, the acceptance probability will only depend on the likelihood ratio of the new set of parameters over the current set of parameters, i.e.

$$\alpha(\boldsymbol{\phi}_k, \boldsymbol{\phi}_k^*) = \min\left\{1, \frac{f(\boldsymbol{y} \mid \boldsymbol{\phi}_k^*)}{f(\boldsymbol{y} \mid \boldsymbol{\phi}_k)}\right\} \tag{3.20}$$

where

$$\frac{f(\boldsymbol{y} \mid \boldsymbol{\phi}_k^*)}{f(\boldsymbol{y} \mid \boldsymbol{\phi}_k)} = \frac{\prod\limits_{\substack{t=p+1 \\ z_{tk}=1}}^{n} \exp\left\{-\frac{1}{2\sigma_k^2}\left(y_t - \phi_{k0} - \sum\limits_{i=1}^{p_k} \phi_{ki}^* y_{t-i}\right)^2\right\}}{\prod\limits_{\substack{t=p+1 \\ z_{tk}=1}}^{n} \exp\left\{-\frac{1}{2\sigma_k^2}\left(y_t - \phi_{k0} - \sum\limits_{i=1}^{p_k} \phi_{ki} y_{t-i}\right)^2\right\}}$$

The priors are absent from the above formula, since their ratio is 1,as a flat prior on the autoregressive parameters was assumed.

This means that the likelihood ratio for the k^{th} component is independent of current values of parameters for the remaining components, which enables to calculate likelihood ratios separately for each component.

The procedure described builds a candidate model with updated mixing weights, shift, scale and autoregressive parameters. However, because stability of such model does not only depend on the autoregressive parameters, we must ensure that the stability condition of Section 2.1.2 is satisfied. If this is not the case, the candidate model and all its parameters are rejected, and the current state of the chain is set to be the same as at the previous iteration.

3.2.4 The label switching problem

Once the samples have been drawn, label switching is dealt with using a k-means clustering algorithm proposed by Celeux (2000). It is common to use the identifiability constraint $\pi_1 > \pi_2 > \cdots > \pi_g$ but it is well known that it is problematic. Examples are given in the discussion to the paper by Richardson and Green (1997). It was shown in fact by Hossain (2012) that applying an identifiability constraint such as $\pi_1 > \pi_2 > \cdots > \pi_g$ may in some cases affect convergence of the chain. With our approach instead, we do not interfere with the chain during the simulation, and hence convergence is not affected.

Our algorithm works by first choosing the first m simulated values of the output after convergence. The value m shall be chosen small enough for label switching to not have occurred yet, and large enough to be able to calculate reliable initial values of cluster centres and their respective variances.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_g)$ be a subset of model parameters of size g, and N the size of the converged sample. The requirement on subsetting is that corresponding paramters of the different mixture components must be chosen, for instance $\boldsymbol{\theta} \equiv (\pi_1, \ldots, \pi_g)$ or

$\boldsymbol{\theta} \equiv (\mu_1, \ldots, \mu_g)$ among other choices. For any centre coordinate θ_i, $i = 1, \ldots, g$ we calculate the mean and variance, based on the first m simulated values, respectively as:

$$\bar{\theta}_i = \frac{1}{m}\sum_{j=1}^{m} \theta_i^{(j)} \qquad \bar{s_i^2} = \frac{1}{m}\sum_{j=1}^{m} \left(\theta_i^{(j)} - \bar{\theta}_i\right)^2$$

We set this to be the "true" permutation of the components, i.e. we now have an initial center $\bar{\boldsymbol{\theta}}^{(0)}$ with variances $\bar{s}_i^{(0)^2}$, $i = 1, \ldots, q$. The remaining $g! - 1$ permutations can be obtained by simply permuting these centres.

From these initial estimates, the r^{th} iteration ($r = 1, \ldots, N - m$) of the procedure consists of two steps:

- the parameter vector $\boldsymbol{\theta}^{(m+r)}$ is assigned to the cluster such that the normalised squared distance

$$\sum_{i=1}^{g} \frac{\left(\theta_i^{(m+r)} - \bar{\theta}_i^{(m+r-1)}\right)^2}{\left(s_i^{(m+r-1)}\right)^2} \tag{3.21}$$

  is minimised, where $\bar{\theta}_i^{(m+r-1)}$ is the i^{th} centre coordinate and $s_i^{(m+r-1)}$ its standard deviation, at the latest update $m + r - 1$.

- Centre coordinates and their variances are respectively updated as follows:

$$\bar{\theta}_i^{(m+r)} = \frac{m+r-1}{m+r}\bar{\theta}_i^{(m+r-1)} + \frac{1}{m+r}\theta_i^{(m+r)} \tag{3.22}$$

  and

$$\begin{aligned}(s_i^{(m+r)})^2 = \frac{m+r-1}{m+r}(s_i^{(m+r-1)})^2 + \frac{m+r-1}{m+r}\left(\bar{\theta}_i^{(m+r-1)} - \bar{\theta}_i^{(m+r)}\right)^2 \\ + \frac{1}{m+r}\left(\theta_i^{(m+r)} - \bar{\theta}_i^{(m+r)}\right)^2\end{aligned} \tag{3.23}$$

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

For the mixture autoregressive case, it is not always clear which subset of the parameters should be used. In fact, group separation might seem clearer in the mixing weights at times, as well as in the scale or shift parameters. Therefore this method requires graphical assistance, i.e. checking the raw output looking for clear group separation. However, it is advisable not to use the autoregressive parameters, especially when the orders are different.

Once the selected subset has been relabelled, labels for the remaining parameters can be switched accordingly.

3.2.5 Reversible Jump MCMC for choosing autoregressive orders

For this step, we use Reversible Jump MCMC (Green, 1995). At each iteration, one component k is randomly chosen from the model. Let p_k be the current autoregressive order of this component, and set p_{max} to be the largest possible value p_k may assume. For the selected component, we propose to increase or decrease its autoregressive order by 1 with probabilities

$$p_k^* = \begin{cases} p_k - 1 & \text{with probability } d(p_k) \\ p_k + 1 & \text{with probability } b(p_k) \end{cases}$$

where $b(p_k) = 1 - d(p_k)$, and such that $d(1) = 0$ and $b(p_{max}) = 0$. Notice that $d(p_k)$ (or equivalently $b(p_k)$) may be any function defined in the interval $[0,1]$ satisfying such condition. For instance, Hossain (2012) introduced two parametric functions for this step. However, in absence of relevant prior information, we choose $b(p_k) = d(p_k) = 0.5$ in our analysis, while presenting the method in the general case.

Finally, it is necessary to point out that in both scenarios we have a 1-1 mapping between current and proposed model, so that the resulting Jacobian is always equal to 1.

Given a proposed move, we proceed as follows:

- If the proposal is to move from p_k to $p_k^* = p_k - 1$, we simply drop ϕ_{kp_k}, and calculate the acceptance probability by multiplying the likelihood ratio and the proposal ratio, i.e.

$$\alpha\left(\mathcal{M}_{p_k}, \mathcal{M}_{p_k^*}\right) = \min\left\{1, \frac{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k^*}\right) p(\boldsymbol{\phi}_k^{p_k^*})}{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k}\right) p(\boldsymbol{\phi}_k^{p_k})} \times \left[\frac{b\left(p_k^*\right)}{d\left(p_k\right)} \times \boldsymbol{\phi}\left(\frac{\phi_{kp_k} - \phi_{kp_k}}{1/\sqrt{\gamma_k}}\right)\right]\right\} \quad (3.24)$$

  where $\boldsymbol{\phi}\left(\frac{\phi_{kp_k} - \phi_{kp_k}}{1/\sqrt{\gamma_k}}\right)$ is the density of the parameter dropped out of the model, according to its proposal distribution.

  If the candidate model is not stable, then it is automatically rejected, i.e. $\alpha\left(\mathcal{M}_{p_k}, \mathcal{M}_{p_k^*}\right) = 0$.

- If the proposed move is from p_k to $p_k^* = p_k + 1$, we proceed by simulating the additional parameter from a suitable distribution. In absence of relevant prior information, the choice is to simulate a value from a uniform distribution centred in 0 and with appropriate range, so that values both close and far apart from 0, both positive and negative, are taken into consideration.

  These considerations lead to draw $\phi_{kp_k^*} \sim \mathcal{U}(-1.5, 1.5)$

  The acceptance probability in this case is

$$\alpha\left(\mathcal{M}_{p_k}, \mathcal{M}_{p_k^*}\right) = \min\left\{1, \frac{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k^*}\right) p(\boldsymbol{\phi}_k^{p_k^*})}{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k}\right) p(\boldsymbol{\phi}_k^{p_k})} \times \left[\frac{d\left(p_k\right)}{b\left(p_k^*\right)} \times 3\right]\right\} \quad (3.25)$$

  where 3 is the inverse of the $\mathcal{U}(-1.5, 1.5)$ density.

  Once again, if the candidate model is not stable, $\alpha\left(\mathcal{M}_{p_k}, \mathcal{M}_{p_k^*}\right) = 0$ and the current model is retained.

Notice that, similarly to the sampler for autoregressive parameters, the prior ratio in both cases is equal to 1 and therefore omitted.

3.2.6 Choosing the number of components

To select the appropriate number of autoregressive components in the mixture, we apply the methods proposed by Chib (1995) and Chib and Jeliazkov (2001), respectively, for use of output from Gibbs and Metropolis-Hastings sampling. Both make use of the marginal likelihood identity.

From Bayes' theorem, we know that

$$p(g|\boldsymbol{y}) \propto f(\boldsymbol{y} \mid g)p(g), \tag{3.26}$$

where $p(g)$ is the prior distribution on g, and $f(\boldsymbol{y} \mid g)$ is the marginal likelihood function, defined as

$$f(\boldsymbol{y} \mid g) = \sum_{p} \int f(\boldsymbol{y} \mid \boldsymbol{\theta}, p, g)p(\boldsymbol{\theta}, p \mid g)d\boldsymbol{\theta} \tag{3.27}$$

with $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau})$ being the parameter vector of the model.

For any values $\boldsymbol{\theta}^*$, p^*, number of components g and observed data $\boldsymbol{y}$, we can use the marginal likelihood identity to decompose the marginal likelihood into parts that are know or can be estimated

$$\begin{aligned} f(\boldsymbol{y}|g) &= \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}^*, p^*, g)p\left(\boldsymbol{\theta}^*, p^* \mid g\right)}{p\left(\boldsymbol{\theta}^*, p^* \mid \boldsymbol{y}, g\right)} \\ &= \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}^*, p^*, g)p\left(\boldsymbol{\theta}^* \mid p^*, g\right)p(p^* \mid g)}{p\left(\boldsymbol{\theta}^* \mid p^*, \boldsymbol{y}, g\right)p(p^* \mid \boldsymbol{y}, g)} \end{aligned} \tag{3.28}$$

Notice that the only quantity not readily available in the above equation is $p\left(\boldsymbol{\theta}^* \mid p^*, \boldsymbol{y}, g\right)$. However, this can be estimated by running reduced MCMC simulations for fixed p^*

(which can be obtained by the RJMCMC method described in Section 2.2.4), as follows:

$$\begin{aligned}\hat{p}(\boldsymbol{\theta}^* \mid p^*, \boldsymbol{y}, g) = \hat{p}(\boldsymbol{\phi}^* \mid \boldsymbol{y}, p^*, g) \\ \hat{p}(\boldsymbol{\mu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g) \\ \hat{p}(\boldsymbol{\tau}^* \mid \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g) \\ \hat{p}(\boldsymbol{\pi}^* \mid \boldsymbol{\tau}^*, \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g)\end{aligned} \tag{3.29}$$

Once these quantities are estimated (see 3.31, 3.32, 3.33, 3.34), plug them in Equation (3.28), together with the other known quantities, to obtain the marginal likelihood for the model with fixed number of components g.

For higher accuracy of results, it is suggested to compare marginal likelihood with different g at points of high density in the posterior distribution of $\boldsymbol{\theta}^*$. We will use the estimated highest posterior density values.

Estimation of $\hat{p}(\boldsymbol{\phi}^* \mid \boldsymbol{y}, p^*, g)$

Suppose we want to estimate $\hat{p}(\boldsymbol{\phi}_{\boldsymbol{k}}^* \mid p^*, \boldsymbol{y}, g)$, for $k = 1, \ldots, g$. We partition the parameter space into two subsets, namely $\Psi_{k-1} = \left(p, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{k-1}, g\right)$ and $\Psi_{k+1} = \left(\boldsymbol{\phi}_{k+1}, \ldots, \boldsymbol{\phi}_g, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}\right)$, where parameters belonging to Ψ_{k-1} are fixed (known or already selected high density values).

First, produce a reduced chain of length N_j to obtain $\boldsymbol{\phi}_{\boldsymbol{k}}^*$, the highest density value for $\boldsymbol{\phi}_{\boldsymbol{k}}$, using the sampling algorithm in Section 4.3, applied to the non-fixed set of parameters only. Define Ψ_{k^*}, the set of known (fixed) parameters with the addition of $\boldsymbol{\phi}_{\boldsymbol{k}}^*$. From a second reduced chain of length N_i, simulate $\{\tilde{\Psi}_{k+1}^{(i)}, \tilde{z}^{(i)} \mid \Psi_{k^*}, \boldsymbol{y}\}$, as well as new observations $\tilde{\boldsymbol{\phi}}_k^{(i)}$ from the proposal density in Equation 10, centred in $\boldsymbol{\phi}_{\boldsymbol{k}}^*$.

Now, let $\alpha(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_{\boldsymbol{k}}^*)$ and $\alpha(\boldsymbol{\phi}_{\boldsymbol{k}}^*, \tilde{\boldsymbol{\phi}}_k^{(i)})$ denote acceptance probabilities respectively of the first and second chain. We can finally estimate the value of the posterior density at

$\boldsymbol{\phi}_k^*$ as

$$\hat{p}\left(\boldsymbol{\phi}_k^* \mid p^\star, \boldsymbol{\phi}_1^*, \ldots, \boldsymbol{\phi}_{k-1}^*, g\right) = \frac{\frac{1}{N_j}\sum_{j=1}^{N_j} \alpha(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_{\boldsymbol{k}}^*) q\left(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_k^*\right)}{\frac{1}{N_i}\sum_{i=1}^{N_i} \alpha(\boldsymbol{\phi}_{\boldsymbol{k}}^*, \tilde{\boldsymbol{\phi}}_k^{(i)})} \tag{3.30}$$

Repeat this procedure for all $k = 1, \ldots, g$ and multiply the single densities to obtain

$$\hat{p}\left(\boldsymbol{\phi}^* \mid \boldsymbol{y}, p^*, g\right) = \prod_{k=1}^{g} \hat{p}\left(\boldsymbol{\phi}_k^* \mid p^\star, \boldsymbol{\phi}_1^*, \ldots, \boldsymbol{\phi}_{k-1}^*, g\right). \tag{3.31}$$

Note that there are no requirements on what N_i and N_j should be, granted the first chain is long enough to have reached the stationary distribution.

Estimation of $\hat{p}\left(\boldsymbol{\mu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right)$

Run a reduced chain of length N. At each iteration, draw observations $\boldsymbol{z}^{(i)}$, $\boldsymbol{\pi}^{(i)}$, $\boldsymbol{\tau}^{(i)}$, $\boldsymbol{\mu}^{(i)}$. Set $\boldsymbol{\mu}^* = (\mu_1, \ldots, \mu_g)$, the parameter vector of highest posterior density. The posterior density at $\boldsymbol{\mu}^*$ can be estimated as

$$\hat{p}\left(\boldsymbol{\mu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{g} p\left(\mu_k^* \mid \boldsymbol{\phi}^*, \boldsymbol{\tau}^{(i)}, \boldsymbol{\pi}^{(i)}, \boldsymbol{y}, \boldsymbol{z}^{(i)}, p^*, g\right). \tag{3.32}$$

Estimation of $\hat{p}\left(\boldsymbol{\tau}^* \mid \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right)$

Run a reduced chain of length N. At each iteration, draw observations $\boldsymbol{z}^{(i)}$, $\boldsymbol{\pi}^{(i)}$, $\boldsymbol{\tau}^{(i)}$. Set $\boldsymbol{\tau}^* = (\tau_1, \ldots, \tau_g)$, the parameter vector of highest posterior density. Posterior density at $\boldsymbol{\tau}^*$ can be estimated as

$$\hat{p}\left(\boldsymbol{\tau}^* \mid \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{g} p\left(\tau_k^* \mid \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{\pi}^{(i)}, \boldsymbol{y}, \boldsymbol{z}^{(i)}, p^*, g\right). \tag{3.33}$$

Estimation of $\hat{p}\left(\boldsymbol{\pi}^* \mid \boldsymbol{\tau}^*, \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right)$

Run a reduced chain of length N. At each iteration, draw observations $\boldsymbol{z}^{(i)}, \boldsymbol{\pi}^{(i)}$. Set $\boldsymbol{\pi}^* = (\pi_1, \ldots, \pi_g)$, the parameter vector of highest posterior density. Posterior density at $\boldsymbol{\pi}^*$ can be estimated as

$$\hat{p}\left(\boldsymbol{\pi}^* \mid \boldsymbol{\tau}^*, \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right) = \frac{1}{N} \sum_{i=1}^{N} \prod_{k=1}^{g} p\left(\pi_k^* \mid \boldsymbol{y}, \boldsymbol{z}^{(i)}, p^*, g\right). \tag{3.34}$$

3.3 Application

3.3.1 Simulation examples

For comparative and demonstrative purposes, we show applications of our method using two simulated datasets from

(A): the MAR$(2;1,1)$ model

$$y_t = \begin{cases} -0.5y_{t-1} + \varepsilon_{t1} & \text{with probability } \pi_1 = 0.5, \\ y_{t-1} + \varepsilon_{t2} & \text{with probability } \pi_2 = 0.5, \end{cases}$$

where $\varepsilon_{t1} \sim N(0,1)$ and $\varepsilon_{t2} \sim N(0,2^2)$ for all t;

(B): the MAR$(3;2,1,1)$ model

$$y_t = \begin{cases} -0.5y_{t-1} + 0.5y_{t-2} + \varepsilon_{t1} & \text{with probability } \pi_1 = 0.5, \\ -0.4y_{t-1} + \varepsilon_{t2} & \text{with probability } \pi_2 = 0.3, \\ y_{t-1} + \varepsilon_{t3} & \text{with probability } \pi_3 = 0.2, \end{cases}$$

where $\varepsilon_{t1} \sim N(0,1)$, $\varepsilon_{t2} \sim N(0,2^2)$, $\varepsilon_{t3} \sim N(0,4^2)$ for all t.

The two time series include respectively 300 and 600 simulated observations. Process **(A)** is similar to the one considered by Hossain (2012) and Wong and Li (2000),

while **(B)** was chosen to illustrate in practice how label switching is dealt with. The issue of label switching for **(B)** can be seen in Figure 3.3, where we show the raw MCMC output with signs of label switching between components 2 and 3 (green and red lines), and the relabelled output after applying the algorithm.

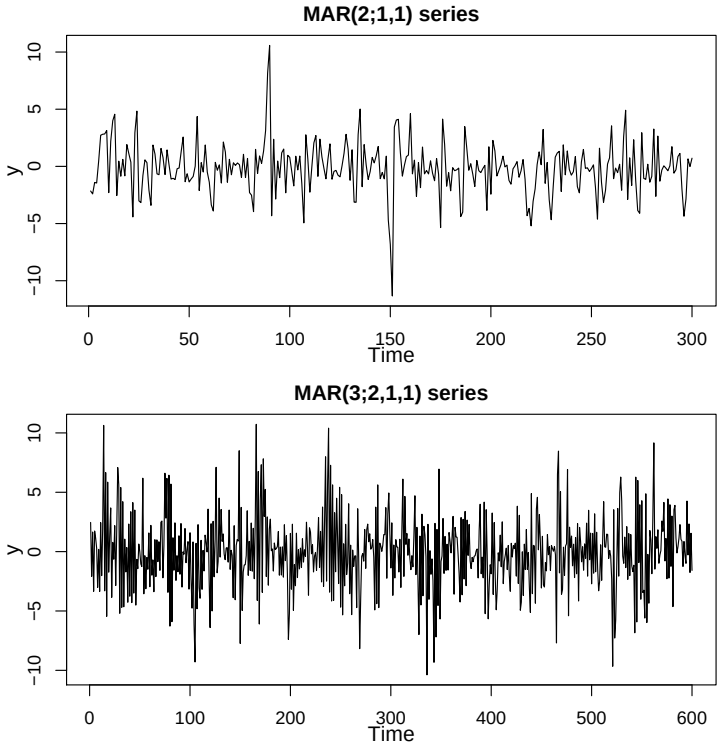

Figure 3.1: Simulated series from **(A)** (top) and **(B)** (bottom).

The algorithm then proceeds as described in Algorithm 1 below:

Algorithm 1

1: **for** $g \leftarrow 2, \ldots, g_{max}$ **do**
2: *RJMCMC and determine* $p_1^*, \ldots, p_k^*$
3: *Calculate* $f(\boldsymbol{y} \mid g)$
4: *Select* $g^* = \max f(\boldsymbol{y} \mid g)$, $g = 2, \ldots, g_{max}$
5: *Simulate* $f(\boldsymbol{\theta} \mid y, g^*, \boldsymbol{p}^*)$

As we can see from Tables, 3.1, 3.2 and 3.3, and Figures 3.2 and 3.4, the "true" model is chosen in both cases, as it has the largest marginal log-likelihood. In addition, true values of the parameters are found in high density regions of their respective

Model (A)	Preference	Marg. log-lik
MAR$(2;1,1)$	0.7399	-611.8113
MAR$(3;1,1,1)$	0.1819	-613.0888
MAR$(4;1,1,1,4)$	0.0382	-923.1585

Model (B)	Preference	Marg. log-lik
MAR$(2;2,1)$	0.6258	-1468.628
MAR$(3;2,1,1)$	0.2937	-1383.061
MAR$(4;2,1,2,1)$	0.0491	-1470.543

Table 3.1: Results from simulation studies. "Preference" is the proportion of times the model was retained against all models with same number of components.

posterior distributions.

Model A	True Value	Posterior Mean	Standard Error	90% HPDR
ϕ_{10}	0	0.011	0.0268	(-0.032, 0.055)
ϕ_{20}	0	-0.183	3.273	(-5.672, 5.206)
ϕ_{11}	-0.5	-0.449	0.037	(-0.511, -0.389)
ϕ_{21}	1	0.994	0.079	(0.869, 1.136)
σ_1	1	0.992	0.079	(0.862, 1.119)
σ_2	2	2.069	0.149	(1.825, 2.311)
π	0.5	0.571	0.046	(0.494, 0.647)

Table 3.2: Results of simulation from posterior distribution of the parameters under model (A).

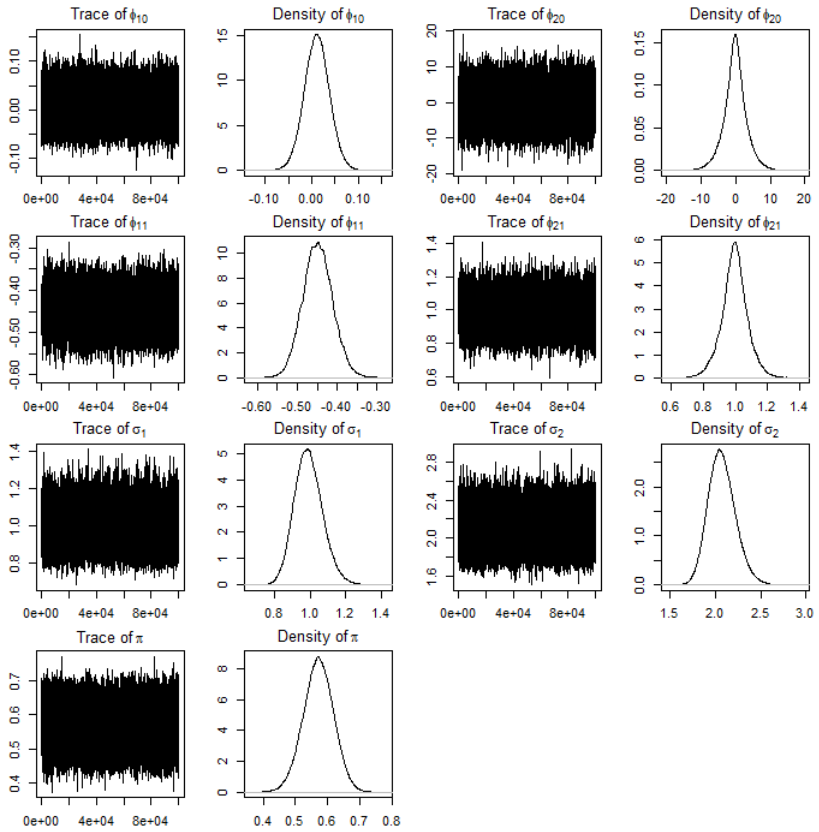

Figure 3.2: Trace and density plots of selected model from (A). Sample size is 100000, after discarding 50000 draws as burn-in period.

Model B	True Value	Posterior Mean	Standard Error	90% HPDR
ϕ_{10}	0	0.001	0.018	(-0.009, 0.007)
ϕ_{20}	0	0.005	0.253	(-0.078, 0.091)
ϕ_{30}	0	0.102	2.133	(-3.145, 3.405)
ϕ_{11}	-0.5	-0.483	0.038	(-0.536, -0.427)
ϕ_{12}	0.5	0.498	0.034	(0.450, 0.547)
ϕ_{21}	-0.4	-0.461	0.105	(-0.596, -0.327)
ϕ_{31}	1	0.731	0.264	(0.432, 1.058)
σ_1	1	1.035	0.246	(0.804, 1.156)
σ_2	2	2.035	0.439	(1.625, 2.522)
σ_3	4	4.074	0.341	(3.559, 4.573)
π_1	0.5	0.495	0.056	(0.411, 0.568)
π_2	0.3	0.293	0.064	(0.207, 0.395)
π_3	0.2	0.212	0.041	(0.148, 0.275)

Table 3.3: Results of simulation from posterior distribution of the parameters under model (B).

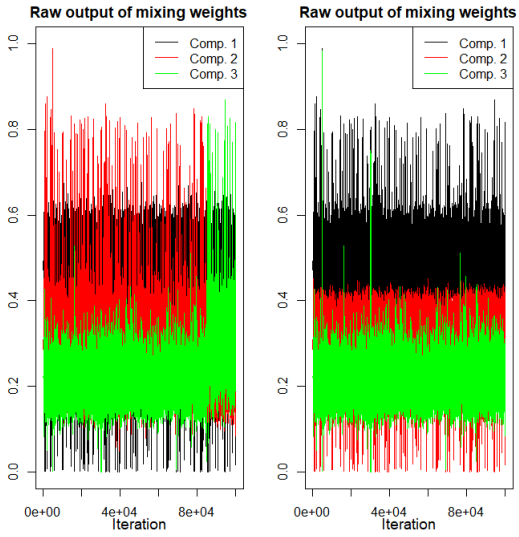

Figure 3.3: Comparison of raw output (left) and output adjusted for label switching of mixing weights from **(B)**. We notice the effectiveness of the relabelling algorithm applied to our MCMC.

3.3.2 The IBM common stock closing prices

The IBM common stock closing prices (Box and Jenkins, 1976) is a financial time series widely explored several times in the literature (see, for instance Wong and Li, 2000). It contains 369 observations from May 17^{th} 1961 to November 2^{nd} 1962. Original and difference series can be seen in Figure 3.5.

Following previous studies, we consider the series of first order differences. To allow direct comparison with Wong and Li (2000) and Hossain (2012), we set $\phi_{k0} = 0$, $k = 1, \ldots, g$.

With the procedure outlined in Algorithm 1 our method chooses a MAR$(3;4,1,1)$ to best fit the data, amongst all 2, 3, and 4 component models of maximum order $p_k = 5$, $k = 1, \ldots, g$. The RJMCMC algorithm selects this model roughly 25% of the time, ahead of MAR$(3;3,1,1)$ with 13%. The marginal log-likelihood for this model is -1245.51, which is larger than that of the best 2 and 4 component models, a

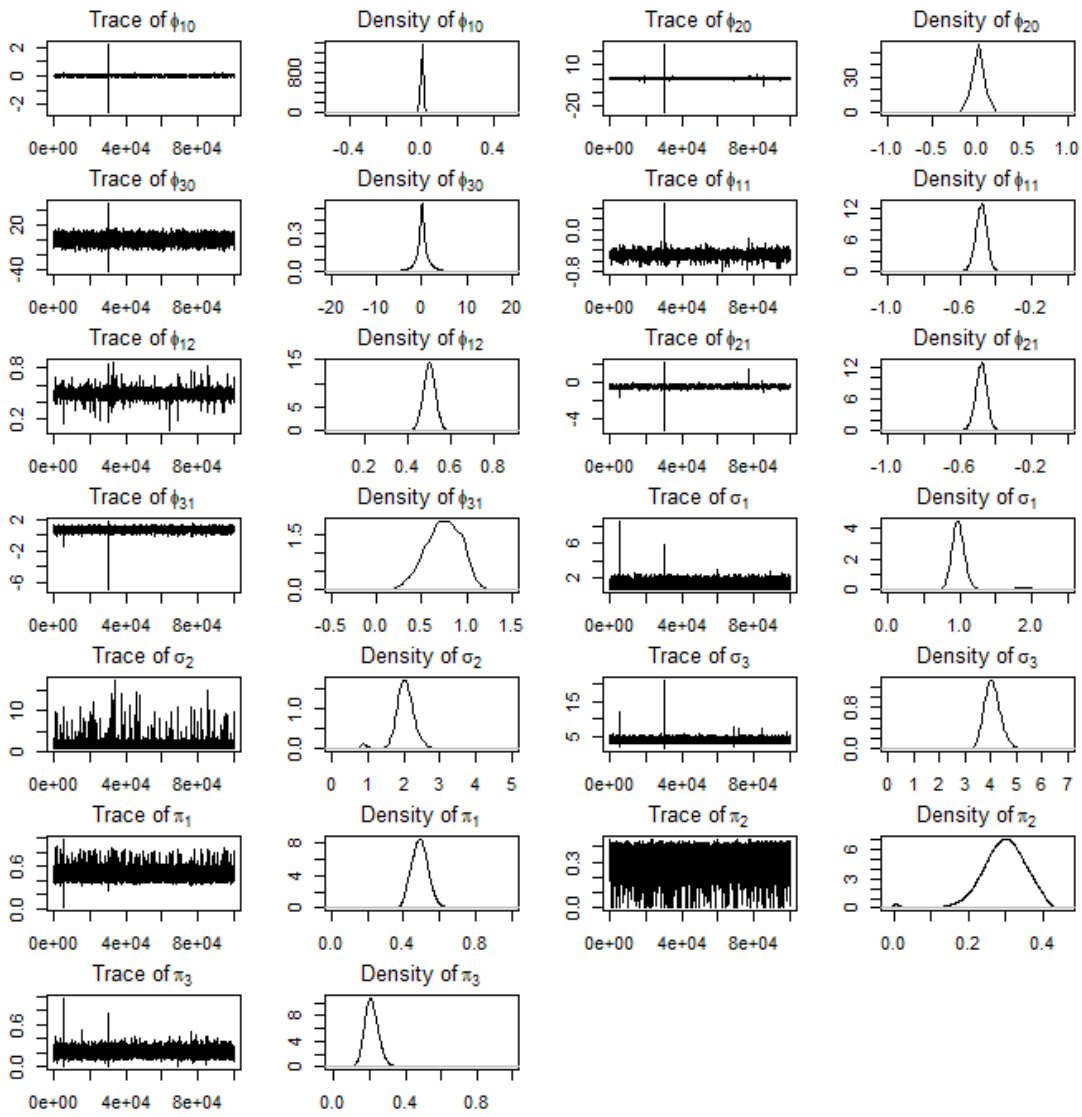

Figure 3.4: Trace and density plots of parameters from (B). Sample size is 100000, after discarding 50000 draws as burn-in period.

MAR$(2;1,1)$ and a MAR$(4;1,1,1,1)$, which respectively have a value of marginal log-likelihood equal to -1248.921 and -1252.381. We immediately notice that this is different from the selected model in Wong and Li (2000), who selected a MAR$(3;1,1,0)$ as best model. Such difference may occur as the frequentist approach fails to capture the multimodality in the distribution of certain parameters, which we can clearly see from Figure 3.6. In fact, by attempting to fit a MAR$(3;4,1,1)$ model by EM-Algorithm from several different starting points, we concluded that this would actually provide a

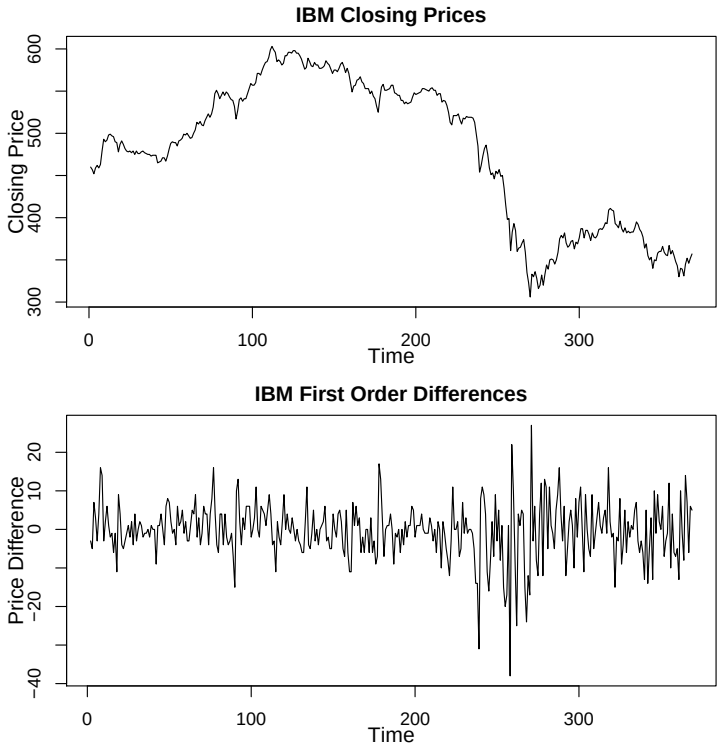

Figure 3.5: Times series of IBM closing prices (top) and series of the first order differences (bottom)

better fit than the MAR$(3;1,1,0)$ chosen by Wong and Li. Furthermore, different starting points to the EM-Algorithm result in convergence to different parameter values for the autoregressive components, which approximately correspond to the modes shown in Figure 3.6.

Figure 3.7 shows once again the time series of first order differences of IBM closing prices, with the addition of two lines representing prediction intervals. Specifically, the red lines delimit the 95% highest density region of the average one step prediction densities, calculated using the sample from the parameter posterior distributions (see Section 3.4) for each y_t for $t > 4$. The blue lines denote instead the 95% prediction interval, calculated as the average one step point predictor $\pm$ twice the average conditional standard error recorded for the predictor, as defined in (**??**). It appears from the

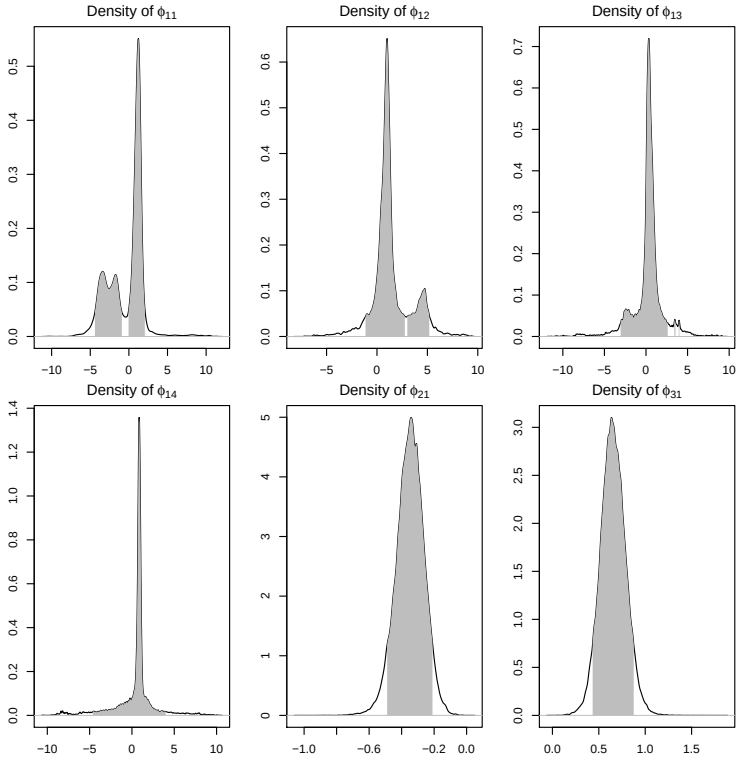

Figure 3.6: Posterior distributions of autoregressive parameters from selected model MAR$(3;4,1,1)$, with 90% HPDR highlighted. We can clearly see multimodality occurring for certain parameters. Sample of 300000 simulated values post burn-in.

picture that there is indeed an advantage in using prediction density over point prediction. While there is not a substantial difference between the two predictors in periods of relatively low volatility, as the very start of the series shows, the interval calculated using density prediction seem to provide more certainty in periods of higher volatility. This can be seen around observations $250-280$, a period of high volatility for the series, where we can see several spikes, and therefore a large prediction interval, for the blue lines, while density prediction seems to accommodate well the sudden jumps in the series. Overall, we may say that, using the highest density region of density forecasts, a MAR model is able to account for the time-dependent volatility and its persistence in the IBM difference series. Furthermore, if we decided for a narrower prediction interval, the density forecast method would allow us to detect presence of multiple modes, so that the highest density region may no longer be continuous. This

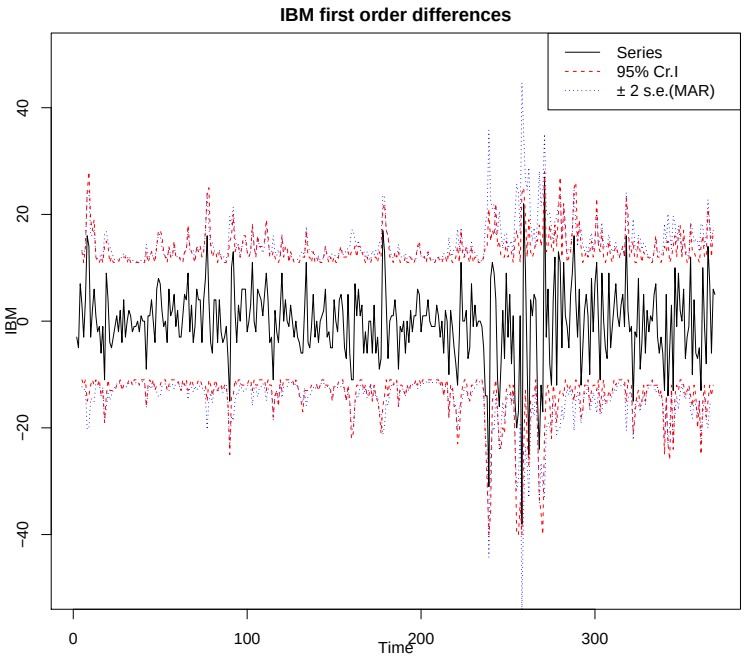

Figure 3.7: IBM first order differences with 95% prediction interval from (mean) density forecast (red) and point prediction $\pm$ twice the (mean) standard error with fitted MAR$(3;4,1,1)$ model.

feature will be seen in Section 3.4.

3.3.3 The Canadian lynx data

Another dataset widely explored in time series literature, amongst which by Wong and Li (2000), is the annual record of Canadian lynx trapped in the Mackenzie River district in Canada between 1821 and 1934. This dataset, listed by Elton and Nicholson (1942), includes 111 observations.

Following previous studies, we consider the natural logarithm of the data, which presents a typical autoregressive correlation structure with 10 years cycles. We notice the presence of multimodality in the log-data, with two local maxima (see Figure 3.8). This suggest that the series may be in fact generated by a mixture of two components.

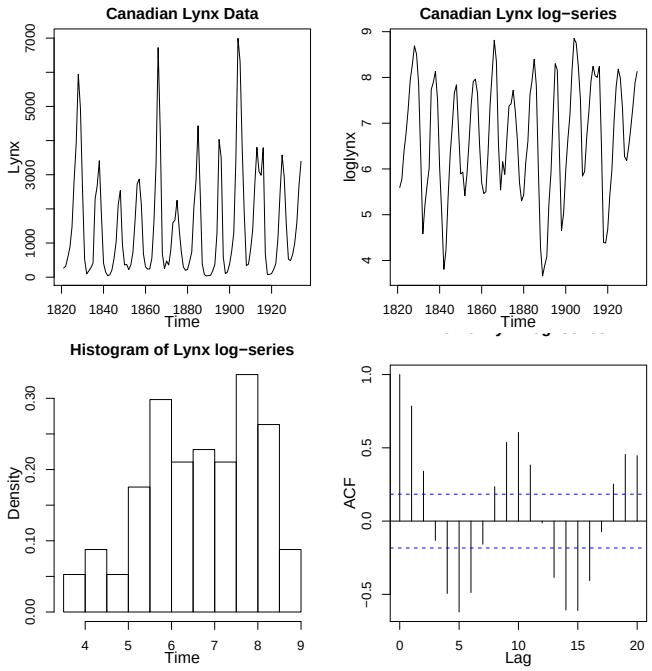

Figure 3.8: Original time series of Canadian lynx (top left), series of natural logarithms (top right), histogram of log-data (bottom left) and autocorrelation plot of log-data (bottom right). The data presents a typical autoregressive correlation structure, as well as multimodality.

In their analysis, Wong and Li (2000) choose a MAR$(2;2,2)$ as best model to fit the data. However, their choice was based on the minimum *BIC* criterion, which in their paper does not always seem reliable for MAR models, particularly with small

datasets.

Aiming to have a better insight about the data, we apply our Bayesian method. The selected model is in this case a MAR$(2;1,2)$, preferred over a MAR$(2;2,2)$ by the algorithm, and to all 2, 3 and 4 component models with autoregressive order $p = 1,2,3,4$. In particular, RJMCMC selects MAR$(2;1,2)$ about 38% of the time, against 20% for *MAR*$(2;2,2)$. The latter is also the model selected by Wong and Li (2000) by EM-Algorithm. Hence, the Bayesian model selection suggests that one fewer autoregressive parameter may be required.

The marginal log-likelihood of this model is -131.0381, which is larger than that of other candidate models MAR$(3;1,2,2)$ with -176.4684 and MAR$(4;1,2,2,1)$ with -154.9989.

We generated a sample of size 100000 from the posterior distribution of the parameters of the selected MAR$(2;1,2)$ model. It is noticed that, for most parameters, the 90% credibility region includes the MLEs obtained by Wong and Li (2000). The only exception stands for the scale parameters, which seem to be slightly larger than such MLEs. However, this may be due to our model containing one fewer AR parameter. On the other hand, these results are in line with the estimates obtained by fitting a MAR$(2;1,2)$ using the EM algorithm, since all estimates are well within the corresponding 90% highest posterior density region.

Parameter	MLE	HD value	Standard Error	90% HPDR
ϕ_{10}	0.4957	0.4962	1.6897	(-1.2599, 3.4341)
ϕ_{20}	2.5728	1.6945	1.2663	(-0.0138, 3.8897)
ϕ_{11}	0.9901	1.0779	0.0667	(0.9893 1.1320)
ϕ_{21}	1.5042	1.7205	0.1594	(1.4717, 1.9866)
ϕ_{22}	-0.8984	-0.7966	0.1528	(-1.0578, -0.5604)
σ_1	0.2313	0.3553	0.1846	(0.2162, 0.6451)
σ_2	0.4828	0.6010	0.1006	(0.4933, 0.7478)
π	0.2358	0.3280	0.1247	(0.1536, 0.5555)

Table 3.4: Summary statistics of sample of size 100000 from posterior distributions of the parameters of the selected model for the log-lynx data.

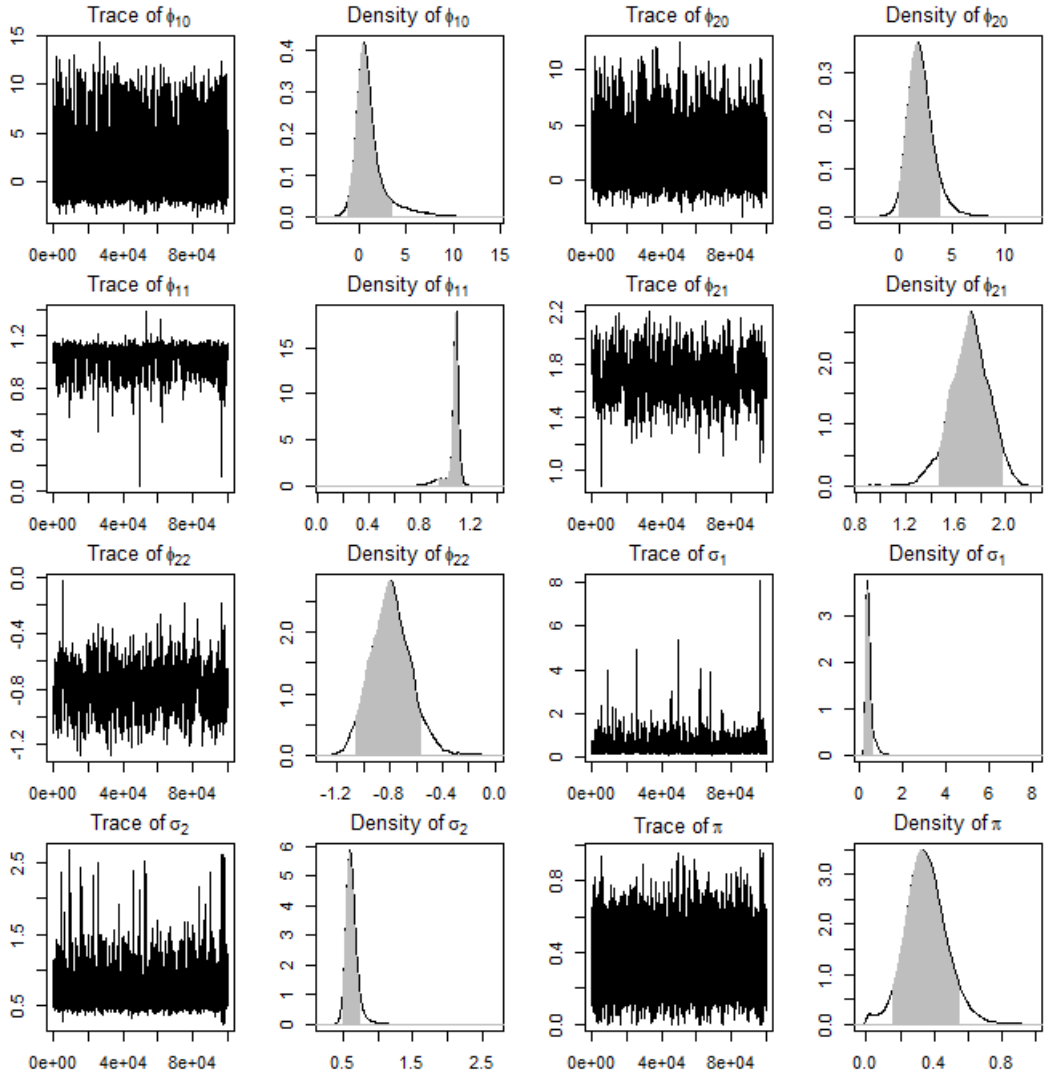

Figure 3.9: Posterior trace plots and density of selected MAR$(2;1,2)$ model for the natural logarithm of Canadian lynx data. For all parameters, the credibility region contains the estimated values from Wong and Li (2000). Sample size is 100000, after 50000 burn-in iterations.

3.4 Bayesian density forecasts with mixture autoregressive models

Once a sample from the posterior is obtained, it is useful to use it to make predictions on future (or off-set) observations.

Wong and Li (2000) and Boshnakov (2009) respectively introduced a simulation based and an analytical method for density forecasts assuming a MAR model. The first method relies on Monte Carlo simulations, while the second derives exact h-step ahead predictive distributions of a given observation.

On one hand, we could estimate density forecasts using the highest posterior density values (i.e. the peak of the posterior distribution). However, it is better in this case to exploit the entire simulated sample as follows:

1. Label each simulation from 1 to N, e.g. $\boldsymbol{\theta}^{(i)}$, $i = 1,\ldots,N$.

2. Arbitrarily define a grid of points which the density shall be evaluated at. With reference to the IBM density forecast example in Figure 3.10, we selected 1000 equally spaced points between 300 and 450 on the $x-$axis. We denote a generic grid point as s.

3. Derive the density forecast $f^{(i)}_{y_{t+h}}\left(y_{t+h} \mid \mathcal{F}_t, \boldsymbol{\theta}^{(i)}\right)$, evaluate it at each grid point, and repeat for $i = 1,\ldots,N$. In this way, we have a sample of N evaluations of the density forecast at each grid point s.

4. Estimate the mean density forecast at each grid point as

$$\hat{f}_{y_{t+h}}(s \mid \mathcal{F}_t) = \frac{1}{N}\sum_{i=1}^{N} f^{(i)}_{y_{t+h}}\left(s \mid \mathcal{F}_t, \boldsymbol{\theta}^{(i)}\right)$$

In this way, we obtain a sample from the h-step ahead density forecast of an observation

of interest. We then average the density at each point over its sample size, to obtain an estimate of the mean density forecast.

We estimate the 1-step and 2-step predictive distributions of the IBM data at $t = 258$ using the analytical method by Boshnakov (2009), and compare them to the ones obtained by EM algorithm (see Figure 3.10). The solid red lines represent the density obtained by Boshnakov (2009) using EM estimates and the exact method. Results of our method are represented by the solid black lines, with the dashed lines as 90% credibility region. The figure also shows how quickly the uncertainty on the predictions grows as we move further in the future, with the 2-step predictive density looking much flatter.

We can see that there are no substantial differences in the shape of these predictive distributions. However, we notice that, particularly for the 2-step predictor, averaging seems to ”stabilise” the density line.

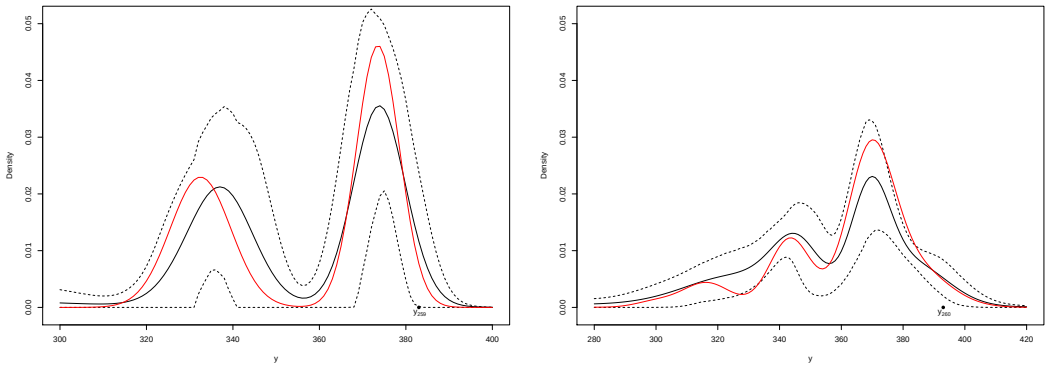

Figure 3.10: Mean density of 1 and 2 steps ahead predictor at $t = 258$ for the IBM data. The solid black line represents our Bayesian method, with the 90% credibility interval identified by the dashed lines. The solid red line represents the predicted density using parameter values from EM estimation by Wong and Li.

We notice from the plots that, clearly for the 1-step predictor and slighlty for the 2-step predictor, the density obtained by MCMC attaches higher density to the observations of interest y_{259} and y_{260}.

3.5 Discussion

We presented an innovative fully Bayesian analysis of mixture autoregressive models with Gaussian components, in particular a new methodology for simulation from the posterior distribution of the autoregressive parameters, which covers the whole stationarity region, compared to previous approaches that constrained it in one way or another. Our approach allowed us to better capture presence of multimodality in the posterior distribution of model parameters. We also introduced a way of dealing with label switching that does not interfere with convergence to the posterior distribution of the model parameters. This consisted in using a relabelling algorithm a posteriori.

Simulations indicate that the method works well. We presented results for two simulated data sets. In both cases the "true" model was selected, and posterior distributions showed high densities regions around the "true" values of the parameters.

The ability of our method to explore the complete stationarity region of the autoregressive parameters allows it to capture better multimodality of distributions. This was illustrated with the IBM and the Canadian Lynx datasets. In the former (Figure 3.6) we saw how multimodality in the posterior distribution of autoregressive parameters was captured, aspects which were missed in the analyses of Hossain (see for instance Figures 3.10 and 3.11 in Hossain, 2012). For this example, it was also noticed that modes of posterior distributions of the autoregressive parameters roughly correspond to point estimates obtained by EM estimation. In the latter (Figure 3.9), we found the mode of ϕ_{21} to be quite distant from 0, with values close to 2 lying in the credibility interval. In this case, the risk with Hossain's method would be to truncate the Normal proposal at points such that a significant part of the stationarity region of the model is not covered. Sampietro's method would have failed to detect such a mode, since it is outside the interval $[-1,1]$.

In conclusion, we may say that our algorithm provides accurate and informative

estimation and a more thorough and comprehensive estimation of model parameters and their distributions, and therefore result in more accurate predictions.

Chapter 4

Bayesian mixture autoregressive model with Student-t innovations

4.1 Introduction

We have already discussed several times how mixture autoregressive models were introduced as a flexible tool to model time series data which presents asymmetry, multimodality and heteroskedasticity. For this reason, MAR models have proven valid to deal with financial returns, which often present one or more of such features.

In their paper, Wong and Li (2000) describe a MAR model with Gaussian innovations, in which the condtitional distribution of each component in the mixture is assumed to be Normal, and use the EM-Algorithm Dempster et al. (1977) for paramter estimation. Since this, examples of Bayesian estimation for MAR models with Gaussian innovations have been presented (see for instance Sampietro, 2006).

Wong et al. (2009) introduced the mixture autoregressive model with Student-t innovations, in which the mixture components are now assumed, conditionally on the past history of the process, to follow a Student-t distribution. The reason behind this different hypothesis for the innovations is that the Student-t distribution, having heavier

tails than the Normal distribution, could be more suitable to model financial returns. In addition, it was argued by the authors that, because the tails of the distribution can be adjusted, a higher level of flexibility is achieved compared to the Gaussian MAR model.

We present in this chapter a fully Bayesian approach to estimating paramters of a mixture autoregressive model with Student-t innovations. Conditional to the past history of the process, each mixture component is assumed to follow a standardised Student-t distribution as formulated in Wong et al. (2009). In addition, exploiting the so called integral representation of the Student-t distribution, component variances do not depend on the degrees of freedom, so that they can be estimated directly. The proposed method is able to identify the best model to fit a time series, as well as estimate parameter posterior distributions, by adapting the MCMC methods seen in Chapter 3.

The degrees of freedom of each mixture component are treated as random variables in the model. In the Bayesian framework, Geweke (1993) proposes a suitable prior distribution for such parameters in the case of a linear regression model with Student-t errors. However, results admittedly may be highly affected by the choice of prior distribution, and therefore one must be careful incorporating their prior belief or knowledge about the data. Geweke (1994) also used a similar approach to time series data with the assumption of Student-t innovations. In both cases, the choice was an exponential prior for the degrees of freedom, which is conservative towards low values regardless of the choice of the hyperparameter. We propose the use of a more informative prior distribution on the degrees of freedom, in order to try to better incorporate prior beliefs on the model. The chosen distribution, unlike the exponential, will favor values that are considered ”more likely” a priori.

In general, it is conventient for the Student-t distribution to constrain the degrees of freedom parameters to be larger than 2, as this ensures existence of both mean and

variance of the distribution. Geweke (1993) and Geweke (1994), as well as different apporches to the problem such as Fonseca et al. (2008), do not seem to take this into account in their analysis. On the contrary, our prior distribution for the degrees of freedom will ensure existence of the first and second moment. Notice that, would we require existence of third and fourth moment, the parameter space would need to be further restricted to ensure they are larger than 4, which our analysis can easily be adapted to.

The chapter is structured as follows: Section 4.2 reviews the mixture autoregressive model with Student-t innovations, its properties, the missing data formulation and the first and second order stationarity (stability) condition. Section 4.3 presents a fully Bayesian analysis of the MAR model with Student-t innovations, including model selection and estimation of posterior distributions of the parameters. Section 4.4 shows a simulation study to present how the methodology works in practice, and finally Section 4.5 presents an example with real time series data.

4.2 Mixture autoregressive model with Student-t innovations

A process $\{y_t\}$ is said to follow a mixture autoregressive (MAR) process with Student-t innovations (Wong et al., 2009) if its conditional CDF can be written as:

$$F\left(y_t \mid \mathcal{F}_{t-1}\right) = \sum_{k=1}^{g} \pi_k F_{\nu_k}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k}\right) \tag{4.1}$$

where:

- $\mathcal{F}_t$ is the sigma-field generated by the process up to, and including (t-1).

- g is the number of mixture components.
- $0 < \pi_k < 1, k = 1, \ldots, g$ are the mixing weights, specifying a discrete probability distribution in $[1, g]$ such that $\sum_{k=1}^{g} \pi_k = 1$ and $\pi_g = 1 - \sum_{k=1}^{g-1} \pi_k$.
- $F_{\nu_k}(\cdot)$, $k = 1, \ldots, g$ denotes the conditional CDF of a standardised Student-t distribution for component k of the mixture, with corresponding degrees of freedom ν_k. Formally, we denote a standardised t distribution with mean μ, variance σ^2 and degrees of freedom ν as $\mathcal{S}(\mu, \sigma^2, \nu)$.
- $\phi_k = \left(\phi_{k1}, \ldots, \phi_{kp_k}\right)$ is the vector of autoregressive parameters for the k^{th} mixture component, with ϕ_{k0} being shift parameter. p_k is the autoregressive order, and we $p = \max(p_k)$ to be the largest autoregressive order in the model. A useful convention is to set $\phi_{kj} = 0$ for $p_k < j \leq p$.
- $\sigma_k, k = 1 \ldots, g$ is the scale parameter, and we define $\tau_k = 1/\sigma_k^2$, the corresponding "precision" parameter.
- If the process starts at $t = 1$, then (4.1) holds for $t > p$.
- The MAR model described in (4.1) is formally denoted as $\text{tMAR}\,(g; p_1, \ldots, p_g)$, where g is the number of mixture components, $p_1, \ldots, p_g$ are the autoregressive orders corresponding to the mixture components, and t implies that the mixture components follow distinct Student-t distributions.

The pdf of the Student-t distribution can be expressed using the so called *integral representation*. Suppose a random variable X follows a Student-t distribution with mean μ, variance σ^2 and degrees of freedom ν. Then the marginal pdf of X can be written as:

$$f_X(x) = \int_0^\infty f_{X|\xi}\,(x \mid \xi)\, f_\xi(\xi) d\xi \tag{4.2}$$

where $X \mid \xi \sim N\left(\mu, \frac{\sigma^2}{\xi}\right)$ and $\xi \sim Ga\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$. This setup is valid for the non-standardised Student-t distribution, for which the variance is equal to $\sigma^2 \frac{\nu}{\nu-2}$.

For the standardised Student-t, it is necessary to adjust the distribution of ξ to a $Ga\left(\frac{\nu}{2}, \frac{\nu-2}{2}\right)$. With this adjustment, the variance of the distribution becomes σ^2, so it no longer depends on the degrees of freedom. At the same time, the degrees of freedom play a part in determining the shape of the distribution, including the tails, and it is therefore important to estimate them accurately.

Given (4.2) and the subsequent considerations, the pdf of the model can be written as

$$\begin{aligned} f(y_t \mid \mathcal{F}_{t-1}) = \sum_{k=1}^{g} \pi_k \sqrt{\frac{\tau_k \xi_t}{2\pi}} \exp\left\{ -\frac{\tau_k \xi_t}{2} \left(y_t - \phi_{k0} - \sum_{i=1}^{p+k} \phi_{ki} y_{t-i} \right)^2 \right\} \\ \times \frac{\frac{\nu_k - 2}{2}^{\nu_k/2}}{\Gamma\left(\frac{\nu_k}{2}\right)} \xi_t^{\nu_k/2-1} \exp\left\{ -\frac{\nu_k - 2}{2} \xi_t \right\} \end{aligned} \tag{4.3}$$

Wong et al. (2009) showed that conditional expectation, conditional variance and autocorrelation functions are identical to the Gaussian MAR model. Respectively:

$$\begin{aligned} \mathrm{E}[y_t \mid \mathcal{F}_{t-1}] &= \sum_{k=1}^{g} \pi_k \mu_{tk} \\ \mathrm{Var}(y_t \mid \mathcal{F}_{t-1}) &= \sum_{k=1}^{g} \pi_k \sigma_k^2 + \sum_{k=1}^{g} \pi_k \mu_{tk}^2 - \sum_{k=1}^{g} (\pi_k \mu_{tk})^2 \\ \rho_h &= \sum_{k=1}^{g} \pi_k \sum_{i=1}^{p} \phi_{ki} \rho_{|h-i|} = \sum_{i=1}^{p} \left(\sum_{k=1}^{g} \pi_k \phi_{ki} \right) \rho_{|h-i|} \qquad h \geq 1 \end{aligned} \tag{4.4}$$

where $\mu_{tk} = \phi_{k0} + \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}$ and ρ_h is the autocorrelation at lag h.

4.3 Bayesian analysis of Student-t MAR model

Given a time series $y_1,\ldots,y_n$, the likelihood function for the Student-t MAR model using (4.3) is:

$$L(\boldsymbol{\phi},\boldsymbol{\sigma},\boldsymbol{\pi}\mid \boldsymbol{y},\boldsymbol{\xi}) = \prod_{t=p+1}^{n}\sum_{k=1}^{g}\pi_k\sqrt{\frac{\tau_k\xi_t}{2\pi}}\exp\left\{-\frac{\tau_k\xi_t}{2}\left(y_t-\phi_{k0}-\sum_{i=1}^{p+k}\phi_{ki}y_{t-i}\right)^2\right\} \times \frac{\frac{\nu_k-2}{2}^{\nu_k/2}}{\Gamma\left(\frac{\nu_k}{2}\right)}\xi_t^{\nu_k/2-1}\exp\left\{-\frac{\nu_k-2}{2}\xi_t\right\} \tag{4.5}$$

The likelihood function is not very tractable and the standard approach is to resort to the missing data formulation (Dempster et al., 1977). Let $\boldsymbol{Z}_t = (Z_{t1},\ldots,Z_{tg})$ be an allocation random variable, where $\boldsymbol{z}_t$ is a g-dimensional vector with entry k equal to 1 if y_t was generated from the k^{th} component in the mixture, and 0 otherwise. We assume that the $\boldsymbol{z}_t$s are discrete random variables, independently drawn from the discrete distribution:

$$P(z_{tk}=1\mid g,\boldsymbol{\pi}) = \pi_k$$

This setup, widely exploited in the literature of finite mixture models (see, for instance Diebolt and Robert, 1994) allows to rewrite the likelihood function in a much more tractable way as follows:

$$L(\boldsymbol{\phi},\boldsymbol{\sigma},\boldsymbol{\pi}\mid \boldsymbol{y},\boldsymbol{\xi},\boldsymbol{z}) = \prod_{t=p+1}^{n}\sum_{k=1}^{g}\left(\pi_k\sqrt{\frac{\tau_k\xi_t}{2\pi}}\exp\left\{-\frac{\tau_k\xi_t}{2}\left(y_t-\phi_{k0}-\sum_{i=1}^{p+k}\phi_{ki}y_{t-i}\right)^2\right\} \times \frac{\frac{\nu_k-2}{2}^{\nu_k/2}}{\Gamma\left(\frac{\nu_k}{2}\right)}\xi_t^{\nu_k/2-1}\exp\left\{-\frac{\nu_k-2}{2}\xi_t\right\}\right)^{z_{tk}} \tag{4.6}$$

Notice that, because exactly one $z_{tk} = 1$ at each time t, the augmented likelihood is a product, and therefore easier to handle.

In practice, both the z_ts and the ξ_ts are not available. We refer to them as latent variables of the model, and we use a Bayesian approach to deal with this.

4.3.1 Priors setup and hyperparameters

The setup of prior distributions mostly exploits and adapts the existing literature (for examples see, for instance, Diebolt and Robert, 1994; Geweke, 1993; Sampietro, 2006).

In absence of relevant prior information, it is reasonable to assume that each observation is equally likely to be generated from any of the mixture components, i.e. $\pi_1 = \ldots, \pi_g = 1/g$. This implies a discrete uniform distribution for the z_ts, which is a particular case of the multinomial distribution. The natural conjugate prior for it is a Dirichlet distribution for $\boldsymbol{\pi}$, and therefore we set:

$$\boldsymbol{\pi} \sim \mathcal{D}(w_1, \ldots, w_g), \qquad w_1 = \cdots = w_g = 1$$

The prior distribution of each ξ_t directly depends upon the corresponding z_t, i.e. which of the mixture component has generated the observation y_t. By model specification, for a generic $z_{tk} = 1$, prior distribution on ξ_t is

$$\xi_t \mid \boldsymbol{z}_t \sim Ga(\frac{\nu_k}{2}, \frac{\nu_k - 2}{2})$$

The prior distribution on the component means is a Normal distribution with common hyperparamters ζ for the mean and κ for the precision

$$\mu_k \sim N\left(\zeta, \kappa^{-1}\right), \qquad k = 1, \ldots, g$$

where μ_k is defined in the same way as in (3.2).

For the precision τ_k, a hierarchical approach is adopted, as suggested by Richardson and Green (1997). Specifically, we set

$$\tau_k \sim Ga(c,\lambda), \qquad k = 1,\ldots,g$$
$$\lambda \sim Ga(a,b)$$

To account for potential multimodality in the distribution, we choose a multivariate uniform prior distribution for the autoregressive parameters, limited in the stability region of the model. Hence, for the parameter vector $\boldsymbol{\phi}$ we have:

$$p(\boldsymbol{\phi} \mid \boldsymbol{\pi}) \propto I\{Stable\}$$

where $I\{\cdot\}$ is the indicator function assuming value 1 if the model is stable and 0 otherwise.

For prior distributions on the degrees of freedom ν_k, $k = 1,\ldots,g$, Geweke (1993) suggests an exponential distribution. However, the author aknowledges that the posterior distribution could potentially be highly influenced by the choice of prior. The exponential distribution naturally favours low degrees of freedom, so it may not always be suitable. We opt instead for $Ga(\alpha_k,\beta_k)$, $k = 1,\ldots,g$ prior distributions, which are more flexible, and allow to better incorporate prior information or belief.

Two more considerations have to be made: degrees of freedom must be larger than 2, too guarantee existence of first and second moments of the Student-t distribution; for degrees of freedom larger than 30, it is reasonable to use a Normal approximation. Therefore, we opted for truncating the prior distribution so that only values in the interval $(2,30)$ belong to the parameter space.

Choice of hyperparameters We require specification for hyperparameters ζ, κ, c, a and b. Although λ is also a hyperparameter, it is treated as a random variable, fully

specified once a and b are chosen.

Following standard setup of mixture models (Richardson and Green, 1997, e.g.), let $\mathcal{R}_y = \max(y) - min(y)$ be the length variation of the dataset. Hyperparameters are then set as follows:

$$a = 0.2 \qquad c = 2 \qquad b = \frac{100a}{c\mathcal{R}_y^2} = \frac{10}{\mathcal{R}_y^2}$$

$$\zeta = \min(y) + \frac{R_y}{2} \qquad \kappa = R_y^{-1}$$

The choice of α_k and β_k for prior distributions of degrees of freedom parameters are the result of three considerations: in general, choosing $\alpha_k > 1$ ensures a peak in the gamma distribution, which could drive the posterior distribution towards such peak. In addition, the mode of a gamma distribution is equal to $\frac{\alpha_k - 1}{\beta_k}$, because of the inevitable subjectivity of this prior, it is reasonable to choose a distribution that sees its peak around the point of maximum likelihood. Denoting $\hat{\nu}_k^{EM}$ the estimate of degrees of freedom using the EM-algorithm approach (Wong et al., 2009), we set a condition that

$$\frac{\alpha_k - 1}{\beta_k} = \hat{\nu}_k^{EM}$$

Finally, we may want to assume that degrees of freedom for all components have a priori the same variance (at least approximately, given the truncated nature of the prior). Given a target variance s^2, this can be done by setting:

$$\frac{\alpha_k}{\beta_k^2} = s^2$$

Thus, each α_k and β_k are carefully chosen so that such conditions are satisfied.

4.3.2 Simulation of latent variables and posterior distributions

We here give formulas for simulation of the latent variables in the model, $\boldsymbol{z}$ and $\boldsymbol{\xi}$, and posterior distributions of model parameters. The methodology is analogous to that of 3.2.3, adjusted for the different distribution assumption on the innovations.

Let $\boldsymbol{\phi}(\cdot)$ denote the pdf of the standard Normal distribution. In addition, let $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\tau}, \lambda, \boldsymbol{\nu})$, and $\boldsymbol{\theta}_{-x}$ the parameter vector excluded x. We introduce the following notation:

$$e_{tk} = y_t - \phi_{k0} - \sum_{i=1}^{p} \phi_{ki} y_{t-i}, \qquad k = 1, \ldots, g; \qquad t = (p+1), \ldots, n$$

$$n_k = \sum_{t=p+1}^{n} z_{tk} \qquad \bar{e}_k = \frac{1}{n_k} \sum_{t:z_{tk}=1} e_{tk} \qquad b_k = 1 - \sum_{i=1}^{p} \phi_{ki}$$

$$c_k = \sum_{t:z_{tk}=1} \xi_t \left(e_{tk} - \bar{e}_k\right) \qquad d_k = \sum_{t:z_{tk}=1} \xi_t$$

Simulation of the latent variables The posterior probability of an observation y_t being generated from component k is:

$$P(z_{tk} = 1 \mid y_t, \xi_t, \boldsymbol{\theta}) = \frac{\frac{\pi_k}{\sigma_k} f_{\nu_k}\left(\frac{e_{tk}}{\sigma_k / \sqrt{\xi_t}}\right)}{\sum_{l=1}^{g} \frac{\pi_l}{\sigma_l} f_{\nu_l}\left(\frac{e_{tl}}{\sigma_l / \sqrt{\xi_t}}\right)} \tag{4.7}$$

Realisations of $\boldsymbol{z}$ are then drawn via a multinomial distribution with the corresponding probabilities.

Posterior distribution of $\boldsymbol{\xi}_t$

As we used the integral representation of the Student-t distribution, we assumed a

priori that $\xi_t \mid \boldsymbol{z}_t = k \sim Ga\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right)$. The posterior distribution is:

$$\begin{aligned} p(\xi_t \mid y_t, \boldsymbol{z}_t, \boldsymbol{\theta}) &\propto \xi_t^{1/2} \exp\left\{ -\frac{\tau_k \xi_t}{2} \left(y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i} \right)^2 \right\} \times \xi_t^{\frac{\nu_k}{2}-1} \exp\left\{ -\frac{\nu_k}{2} \xi_t \right\} \\ &= \xi_t^{\frac{\nu_k-1}{2}} \exp\left\{ -\xi_t \left[\frac{\tau_k}{2} \left(y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i} \right)^2 + \frac{\nu_k}{2} \right] \right\} \end{aligned} \tag{4.8}$$

And hence $\xi_t \mid y_t, \boldsymbol{z}_t, \boldsymbol{\theta} \sim Ga\left(\frac{\nu_k + 1}{2}, \frac{\tau_k e_{tk}^2}{2} + \frac{\nu_k}{2}\right)$.

Posterior distribution of $\boldsymbol{\pi}$

The prior distribution is $p(\boldsymbol{\pi}) \sim D(1, \ldots, 1)$. The posterior distribution is:

$$p(\boldsymbol{\pi} \mid \boldsymbol{y}, \boldsymbol{z}) \propto \prod_{k=1}^{g} \pi_k^{n_k} \tag{4.9}$$

Therefore $\boldsymbol{\pi} \mid \boldsymbol{y}, \boldsymbol{z} \sim D(n_1 + 1, \ldots, n_g + 1)$.

Posterior distribution of $\boldsymbol{\mu}_k$

The prior distribution is $\mu_k \sim N(\zeta, \kappa^{-1})$. The posterior distribution can be derived

as follows:

$$
\begin{aligned}
p(\mu_k \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta}_{-\mu_k}) &\propto \exp\left\{ -\frac{\tau_k}{2} \sum_{t:z_t=k} \xi_t \left[(y_t - \mu_k) - \sum_{i=1}^{p_k} \phi_{ki} (y_{t-i} - \mu_k) \right]^2 \right\} \times \exp\left\{ -\frac{\kappa}{2} (\mu_k - \zeta)^2 \right\} \\
&= \exp\left\{ -\frac{\tau_k}{2} \sum_{t:z_t=k} \xi_t \left[\left(\underbrace{y_t - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}_{e_{tk}} \right) - \left(\underbrace{1 - \sum_{i=1}^{p_k} \phi_{ki}}_{b_k} \right) \mu_k \right]^2 \right\} \times \exp\left\{ -\frac{\kappa}{2} (\mu_k - \zeta)^2 \right\} \\
&= \exp\left\{ -\frac{\tau_k}{2} \sum_{t:z_t=k} \xi_t (e_{tk} - b_k \mu_k)^2 \right\} \times \exp\left\{ -\frac{\kappa}{2} (\mu_k - \zeta)^2 \right\} \\
&= \exp\left\{ -\frac{\tau_k}{2} \sum_{t:z_t=k} \xi_t (e_{tk} - \bar{e}_k + \bar{e}_k - b_k \mu_k)^2 \right\} \times \exp\left\{ -\frac{\kappa}{2} (\mu_k - \zeta)^2 \right\} \\
&= \exp\left\{ -\frac{\tau_k}{2} \sum_{t:z_t=k} \xi_t \left[(e_{tk} - \bar{e}_k)^2 + + (\bar{e}_k - b_k \mu_k)^2 + 2 (e_{tk} - \bar{e}_k)(\bar{e}_k - b_k \mu_k) \right] \right\} \\
&\times \exp\left\{ -\frac{\kappa}{2} (\mu_k - \zeta)^2 \right\} \\
&\propto \exp\left\{ -\tau_k (\bar{e}_k - b_k \mu_k) \underbrace{\sum_{t:z_t=k} (e_{tk} - \bar{e}_k)}_{c_k} - \frac{\tau_k}{2} (\bar{e}_k - b_k \mu_k)^2 \underbrace{\sum_{t:z_t=k} \xi_t}_{d_k} \right\} \\
&\times \exp\left\{ -\frac{\kappa}{2} (\mu_k - \zeta)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \left(\tau_k b_k^2 d_k + \kappa \right) \mu_k^2 + \left[\tau_k b_k (c_k + d_k \bar{e}_k) + \kappa \zeta \right] \mu_k \right\}
\end{aligned}
\tag{4.10}
$$

And therefore, we conclude that $\mu_k \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta}_{-\mu_k} \sim N\left(\dfrac{\tau_k b_k (c_k + d_k \bar{e}_k) + \kappa \zeta}{\tau_k b_k^2 d_k + \kappa}, \dfrac{1}{\tau_k b_k^2 d_k + \kappa} \right)$.

Posterior distribution of $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}_k$

The prior distribution for τ_k was hierarchical, with the hyperparameter λ itself being random. We first derive posterior distribution of λ.

$$
\begin{aligned}
p(\lambda \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta}_{-\lambda}) &\propto \left(\prod_{k=1}^{g} \lambda^c \exp\left\{ -\lambda \tau_k \right\} \right) \times \lambda^{a-1} \exp\{-b\lambda\} \Bigg\} \\
&= \lambda^{cg+a-1} \exp\left\{ -\left(b + \sum_{k=1}^{g} \tau_k \right) \lambda \right\}
\end{aligned}
\tag{4.11}
$$

Therefore, we have $\lambda \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta}_{-\lambda} \sim Ga\left(cg + a, b + \sum_{k=1}^{g} \tau_k\right)$. Posterior distribution of τ_k conditional on λ is then

$$\begin{aligned} p\left(\tau_k \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta}_{-\tau_k}\right) &\propto \prod_{t:z_t=k} \tau^{1/2} \exp\left\{ -\frac{\tau_k}{2} \sum_{t:z_t=k} \xi_t \left(y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i} \right)^2 \right\} \\ &\quad \times \tau_k^{c-1} \exp\left\{ -\lambda \tau_k \right\} \\ &= \tau_k^{c-1+n_k/2} \exp\left\{ -\left[\frac{1}{2} \sum_{t:z_t=k} \xi_t \left(\underbrace{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}_{e_{tk}} \right)^2 + \lambda \right] \tau_k \right\} \end{aligned} \tag{4.12}$$

and hence we conclude $\tau_k \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta}_{-\tau_k} \sim Ga\left(c + \frac{n_k}{2}, \frac{1}{2} \sum_{t:z_t=k} \xi_t e_{tk}^2 + \lambda\right)$.

Update of $\boldsymbol{\phi}$ and $\boldsymbol{\nu}$

Posterior distributions of $\boldsymbol{\phi}_k$ and ν_k do not have the form of a standard distribution, therefore we resort to Metropolis-Hastings methods for simulation.

For the autoregressive parameters, $\boldsymbol{\phi}_k$, $k = 1, \ldots, g$, we make us of random walk metropolis. Let $\boldsymbol{\phi}_k$ be the current state of the chain. We simulate a candidate value $\boldsymbol{\phi}_k^*$ from the proposal distribution $MVN\left(\boldsymbol{\phi}_k, \gamma_k I_{p_k}\right)$, where γ_k is a tuning parameter and I_{pk} is the $p_k \times p_k$ identity matrix. A move to the candidate value $\boldsymbol{\phi}_k^*$ is then accepted with probability

$$\alpha\left(\boldsymbol{\phi}_k, \boldsymbol{\phi}_k^*\right) = \min\left(1, \frac{\exp\left\{ -\frac{\tau_k}{2} \sum_{t:z_{tk}=1} \left(y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki}^* y_{t-i} \right)^2 \right\}}{\exp\left\{ -\frac{\tau_k}{2} \sum_{t:z_{tk}=1} \left(y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i} \right)^2 \right\}} \right) \tag{4.13}$$

Before the candidate value is accepted, stability of the updated model, as defined

in Section 2.1.2, is assessed. If the model is not stable, then we set $\alpha\left(\boldsymbol{\phi}_k^{\cdot}\boldsymbol{\phi}_k^*\right) = 0$, and the candidate value is automatically rejected.

The posterior distribution of a generic ν_k can be written as:

$$p\left(\nu_k \mid \boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{z}, \boldsymbol{\theta}_{-\nu_k}\right) \propto \frac{\left(\frac{\nu_k - 2}{2}\right)^{n_k \nu_k/2}}{\Gamma\left(\frac{\nu_k}{2}\right)} \prod_{t:z_{tk}=1} \xi_t^{\nu_k/2-1} \exp\left\{\frac{\nu_k - 2}{2} \sum_{t:z_{tk}=1} \xi_t\right\} \times$$

$$\nu_k \exp\{-\beta\nu_k\} \quad (4.14)$$

which indeed is not a standard distribution. We propose an independent sampler. Regardless of the current state of the chain, say ν_k, we simulate a candidate value ν_k^* from its prior distribution. In this way, the acceptance probability reduces to the likelihood ratio between the candidate value and the current value, i.e.

$$\alpha(\nu_k, \nu_k^*) = \min\left(1, \frac{\left(\frac{\nu_k^* - 2}{2}\right)^{n_k \nu_k^*/2}}{\left(\frac{\nu_k - 2}{2}\right)^{n_k \nu_k/2}} \frac{\Gamma\left(\frac{\nu_k}{2}\right)}{\Gamma\left(\frac{\nu_k^*}{2}\right)} \frac{\prod_{t:z_{tk}=1} \xi_t^{\nu_k^*/2-1}}{\prod_{t:z_{tk}=1} \xi_t^{\nu_k/2-1}} \frac{\nu_k^*}{\nu_k} \exp\left\{-\beta(\nu_k - \nu_k^*)\right\}\right) \quad (4.15)$$

4.3.3 Choosing autoregressive orders

For this step, we resort to reversible jump MCMC (Green, 1995), updating the equations of Section 3.2.5 to account for the new model assumptions. At each iteration, one of the g mixture components, say k, is chosen at random. Let p_k be the current autoregressive order of such component. In addition, set p_{max} as the largest possible autoregressive order, chosen arbitrarily. The proposal is to increase the autoregressive order to $p_k^* = p_k + 1$ with probability $b(p_k)$, or decrease it to $p_k^* = p_k - 1$ with probability $d(p_k)$. $b(\cdot)$ may be any function defined in $[0,1]$ satisfying $b(p_{max}) = 0$, and $d(p_k) = 1 - b(p_k)$.

Both scenarios have a $1-1$ mapping between current and candidate model, since the only difference between the two is the addition or subtraction of the largest order autoregressive parameter. Therefore, the Jacobian is always equal to 1.

Given a proposed move, we proceed as follows:

- If the proposed move is to $p_k^* = p_k - 1$, the autoregressive parameter ϕ_{kp_k} is dropped from the model, and the acceptance probability is the product of the likelihood and the proposal ratio, i.e.

$$\alpha(p_k, p_k^*) = \min\left\{1, \frac{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k^*}\right)}{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k}\right)} \times \frac{b(p_k^*)}{d(p_k)} \times \boldsymbol{\phi}\left(\frac{\phi_{pk} - \phi_{pk}}{1/\sqrt{\gamma_k}}\right)\right\} \tag{4.16}$$

- If the proposal is to move to $p_k^* = p_k + 1$, we simulate the additional parameter $\phi_{kp_k^*}$ from a $\mathcal{U}(-1.5, 1.5)$ distribution. This choice ensures that values close to 0 are equally as likely to be taken into consideration as values far from zero, while trying to maintain the algorithm as efficient as possible in terms of drawing values within the stability region of the model.

  In this case, the acceptance probability is the ratio between the likelihood and the proposal, i.e.

$$\alpha(p_k, p_k^*) = \min\left\{1, \frac{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k^*}\right)}{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k}\right)} \times \frac{d(p_k^*)}{b(p_k)} \times 3\right\} \tag{4.17}$$

  where 3 is the inverse of the density of any $\phi_{kp_k^*}$ under a $\mathcal{U}(-1.5, 1.5)$ proposal distribution.

Notice that, in both scenarios, if the candidate model does not satisfy the stability condition of Section 2.1.2, then it is automatically rejected.

Ultimately, the model which is selected the most number of times over a predetrmined number of iterations is retained to be the best fit for the data (for a certain

fixed g).

4.3.4 Choosing the number of mixture components

The analysis presented so far works under the assumption of correct specification of the number of mixture components g. We now need a way to select a suitable number of mixture components.

Recall the marginal likelihood identity. The marginal likelihood function, which is only conditional on the number of mixture components g, is defined as:

$$f(\boldsymbol{y} \mid g) = \sum_{p} \int f(\boldsymbol{y} \mid \boldsymbol{\theta}, p, g) p(\boldsymbol{\theta}, p \mid g) d\theta \tag{4.18}$$

where $\boldsymbol{\theta}$ is the vector of model parameters. In our case, $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\nu})$.

For any values $\boldsymbol{\theta}^*$, p^*, g and observed data y, the marginal likelihood identity can be decomposed into products of quantities that can be estimated:

$$f(\boldsymbol{y} \mid g) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}^*, p^*, g) p(\boldsymbol{\theta}^* \mid p^*, g) p(p^* \mid g)}{p(\boldsymbol{\theta}^* \mid \boldsymbol{y}, p^*, g) p(p^* \mid y, g)} \tag{4.19}$$

Notice that most quantities in (4.19) are ready available. In fact, $f(\boldsymbol{y} \mid \boldsymbol{\theta}^*, p^*, g)$ is the conditional pdf of the data, which is known under the model specification; $p(\boldsymbol{\theta}^* \mid p^*, g)$ is the set of prior densities on the model parameters (see Section 4.3.1); $p(p^* \mid g)$ is the prior on the maximum autoregressive order, which is discrete uniform in $[1, p_{max}]$ a priori (see Section 4.3.3); $p(p^* \mid \boldsymbol{y}, g)$ is the posterior distribution of the selected autoregressive orders, which we approximate by the proportion of times the RJMCMC algorithm in Section 4.3.3 retains such model; finally, $p(\boldsymbol{\theta}^* \mid \boldsymbol{y}, p^*, g)$ is the set of posterior densities on the model parameters (see Section 4.3.2), which needs to be estimated.

To estimate $p(\boldsymbol{\theta}^* \mid \boldsymbol{y}, p^*, g)$ we recur to the the methods by Chib (1995) and Chib

and Jeliazkov (2001), respectively for use of output from Gibbs sampling and Metropolis-Hastings sampling. The method is analogous to that used in Sampietro (2006), taking into account the different model specification, and the additional model parameters introduced for the degrees of freedom of each mixture component.

Notice that $p(\boldsymbol{\theta}^* \mid \boldsymbol{y}, p^*, g)$ can be further decomposed into a product:

$$\begin{aligned} p(\boldsymbol{\theta}^* \mid \boldsymbol{y}, p^*, g) = & p(\boldsymbol{\phi}^* \mid \boldsymbol{y}, p^*, g) \\ & p(\boldsymbol{\nu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g) \\ & p(\boldsymbol{\mu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{y}, p^*, g) \\ & p(\boldsymbol{\tau}^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{y}, p^*, g) \\ & p(\boldsymbol{\pi}^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \boldsymbol{y}, p^*, g) \end{aligned} \tag{4.20}$$

Once all quantities have been estimated, they are plugged into (4.19) to estimate the marginal loglikelihood.

To compare models with different g, the algorithm must be run separately for each individual g_1, g_2, and so on. In addition, for better efficiency it is recommended that models with different number of mixture components are compared on the basis of high density values of the parameters according to their distributions in (4.20).

Estimation of $p(\boldsymbol{\phi}^* \mid \boldsymbol{y}, p^*, g)$

Posterior distributions of autoregressive paramters are estimated by a Metropolis-Hastings algorithm. Here we describe how to estimate the probability of interest.

For a generic mixture component k, we partition the parameter space into two subsets, namely $\Psi_{k-1} = (p, \boldsymbol{\phi}_1^*, \ldots, \boldsymbol{\phi}_{k-1}^*, g)$ and $\Psi_{k+1} = (\boldsymbol{\phi}_{k+1}, \ldots, \boldsymbol{\phi}_g, \boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$, where parameters in Ψ_{k-1} are fixed.

First, produce a reduced chain of length N_j for the non-fixed parameters, and fix $\boldsymbol{\phi}_k^*$

to be the highest density value. Define now $\Psi_k = (\Psi_{k-1}, \boldsymbol{\phi}_k^*)$

Run a second reduced chain of length N_i (N_i and N_j may be equal) for Ψ_{k+1}, as well as a sample $\tilde{\boldsymbol{\phi}}_k$ from the proposal distribution $MVN(\boldsymbol{\phi}_k^*, \gamma_k I_{p_k})$.

Finally, let $\alpha(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_k^*)$ and $\alpha(\boldsymbol{\phi}_k^*, \tilde{\boldsymbol{\phi}}_k^{(i)})$ be the acceptance probabilities of the Metropolis-Hastings algorithm, respectively for the first and the second chain. The conditional density at $\boldsymbol{\phi}_k^*$ can then be estimated as

$$p(\boldsymbol{\phi}_k^* \mid \Psi_{k-1}, \boldsymbol{y}, p^*, g) = \frac{\frac{1}{N_j}\sum_{j=1}^{N_j} \alpha(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_k^*) q_{\phi_k}\left(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_k^*\right)}{\frac{1}{N_i}\sum_{i=1}^{N_i} \alpha(\boldsymbol{\phi}_k^*, \tilde{\boldsymbol{\phi}}_k^{(i)})} \tag{4.21}$$

where $q\left(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_k^*\right)$ denotes the density of $\boldsymbol{\phi}_k^{(j)}$ under the proposal $MVN(\boldsymbol{\phi}_k^*, \gamma_k I_{p_k})$.

Estimation of $p(\nu^* \mid \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g)$

Degrees of freedom are also estimated via Metropolis-Hastings, therefore we proceed in a similar way.

For a generic component k, partition the parameter space into $\Omega_{k-1} = (p, \boldsymbol{\phi}^*, \nu_1, \ldots, \nu_{k-1}, g)$ and $\Omega_{k+1} = (\nu_{k+1}, \ldots, \nu_g, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$.

Produce a reduced chain of length N_j for the non-fixed parameters and fix ν_k^* to be the highest density value, and define $\Omega_k = (\Omega_{k-1}, \nu_k^*)$.

Run as second chain of length N_i for Ω_{k+1}, as well as second sample $\tilde{\nu}_k$ from the proposal disitribution. Let $\alpha(\nu_k^{(j)}, \nu_k^*)$ and $\alpha(\nu_k^*, \tilde{\nu}_k^{(i)})$ be acceptance probabilities respectively of the first and second chain. The conditional density at ν_k^* can be estimated as

$$p(\nu_k^* \mid \Omega_{k-1}, \boldsymbol{y}, p^*, g) = \frac{\frac{1}{N_j}\sum_{j=1}^{N_j} \alpha(\nu_k^{(j)}, \nu_k^*) q_{\nu_k}\left(\nu_k^{(j)}, \nu_k^*\right)}{\frac{1}{N_i}\sum_{i=1}^{N_i} \alpha(\nu_k^*, \tilde{\nu}_k^{(i)})} \tag{4.22}$$

where $q_{\nu_k}\left(\nu_k^{(j)}, \nu_k^*\right)$ denotes the density of $\nu_k^{(j)}$ under the prior (proposal) distribution $Ga(\alpha, \beta)$.

Estimation of $p(\boldsymbol{\mu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{y}, p^*, g)$

Run a reduced chain of length N_j for the non-fixed parameters. Set $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_g^*)$ to be the highest density value. The posterior density of $\boldsymbol{\mu}^*$ can be estimated as:

$$p(\boldsymbol{\mu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{y}, p^*, g) = \frac{1}{N} \sum_{j=1}^{N_j} \prod_{k=1}^{g} p\left(\mu_k^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{\tau}^{(i)}, \boldsymbol{\pi}^{(i)}, y, \boldsymbol{z}^{(i)}, p^*, g\right) \tag{4.23}$$

Estimation of $p(\boldsymbol{\tau}^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{y}, \boldsymbol{p}^*, \boldsymbol{g})$

Run a reduced chain of length N_j for the non-fixed parameters. Set $\boldsymbol{\tau}^* = (\tau_1^*, \ldots, \tau_g^*)$ to be the highest density value. The posterior density of $\boldsymbol{\tau}^*$ can be estimated as:

$$p(\boldsymbol{\tau}^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{y}, p^*, g) = \frac{1}{N_j} \sum_{j=1}^{N_j} \prod_{k=1}^{g} p\left(\tau_k^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\pi}^{(i)}, \boldsymbol{y}, \boldsymbol{z}^{(i)}, p^*, g\right) \tag{4.24}$$

Estimation of $p(\boldsymbol{\pi}^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \boldsymbol{y}, \boldsymbol{p}^*, \boldsymbol{g})$

Run a reduced chain of length N_j for the mixing weights, which are now the only non-fixed parameters. Set $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_g^*)$ to be the highest density value. The posterior density of $\boldsymbol{\pi}^*$ can be estimated as:

$$p(\boldsymbol{\pi}^* \mid \boldsymbol{\nu}^*, \boldsymbol{\phi}^*, \boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \boldsymbol{y}, p^*, g) = \frac{1}{N_j} \sum_{j=1}^{N_j} p\left(\boldsymbol{\pi}^* \mid \boldsymbol{\phi}^*, \boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \boldsymbol{y}, \boldsymbol{z}^{(i)}, p^*, g\right) \tag{4.25}$$

4.4 Example

To illustrate performance of our method, we simulated a time series of length $n = 500$ from the process:

$$y_t = \begin{cases} -0.5y_{t-1} + 0.5y_{t-2} + \varepsilon_{t1} & \text{with probability } \pi_1 = 0.4 \\ 1.1y_{t-1} + \varepsilon_{t2} & \text{with probability } \pi_2 = 0.4 \\ -0.4y_{t-1} + \varepsilon_{t3} & \text{with probability } \pi_3 = 0.2 \end{cases}$$

where $\varepsilon_{t1} \sim \mathcal{S}(0, 5^2, 4)$, $\varepsilon_{t2} \sim \mathcal{S}(0, 3^2, 14)$ and $\varepsilon_{t3} \sim \mathcal{S}(0, 1, 10)$, and $\mathcal{S}(\mu, \sigma^2, \nu)$ is the Student-t distribution with mean μ, variance σ^2 and degrees of freedom ν. We denote this model as $\text{tMAR}(3; 2, 1, 1)$.

The series can be seen in Figure 4.1, and it represents what in practice one should be looking for to assume a MAR model. The series looks in fact heteroskedastic, amd the plot of the sample autocorrelation shows that data are slightly correlated at lag 2. Both these features may indicate that the underlying generating process is mixture autoregressive.

For the analysis, we compared all possible models with 2 and 3 mixture components, and maximum autoregressive order equal to 4.

For what regards the optimal autoregressive orders, the RJMCMC algorithm chooses a $\text{tMAR}(3; 2, 1, 1)$ among all 3-component models with a preference of 0.8054, which means the model was retained as "best" 3-component model for roughly 81% of the iterations. The competitor 2-component models is a a $\text{tMAR}(2; 2, 1)$, with a preference of 0.8149. When compared with each other in terms of marginal log-likelihood, the best model is $\text{tMAR}(3; 2, 1, 1)$ with marginal log-likelihood of -1502.77 against -1519.166 for $\text{tMAR}(2; 2, 1)$.

We then simulated a sample of length 100000 from the posterior distribution of the

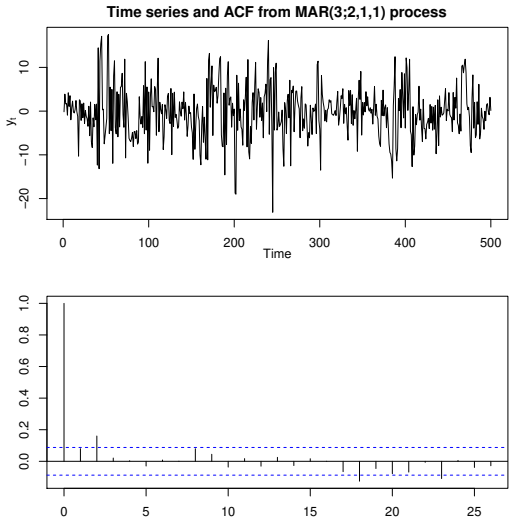

Figure 4.1: Simulated time series from tMAR$(3;2,1,1)$ process (top) and sample autocorrelation.

paramters, after allowing 10000 burn-in iterations. Results are displayed in Figure 4.2 and Figure 4.3.

We can see from Figure 4.2 that almost all "true" parameters are included within the 95% posterior density region of their respective distribution. The only exception is found in μ_1, for which such region is $[-1.449, -0.0177]$. However, it must be taken into account that component 1 has the largest variance and the largest autoregressive order, and is therefore more subject to sampling variability. For what regards the degrees of freedom parameter, all three components have their peak near the true values of the paramaeters: respectively, peaks are found between $[4,7]$, $[11,13]$ and $[8,11]$ (true values are 4, 14 and 10).

Overall, we may be satisfied with performance of the algorithm.

4.5 The IBM common stock closing prices

We propose once again an analysis of the IBM common stock closing prices seen in Section 3.3.2. This way, we will have a term of comparison between the respective

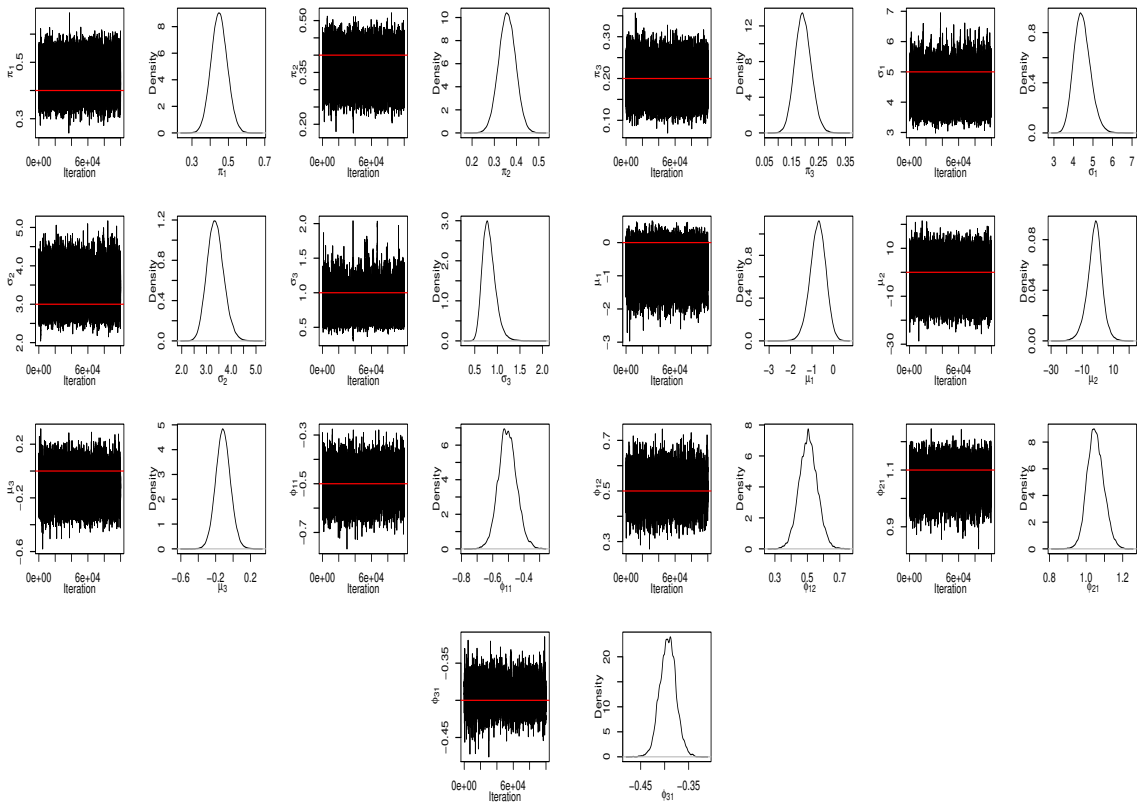

Figure 4.2: Trace and density plots of full conditional posterior distributions of model parameters under selected tMAR$(3;2,1,1)$ model. Red lines highlight true values.

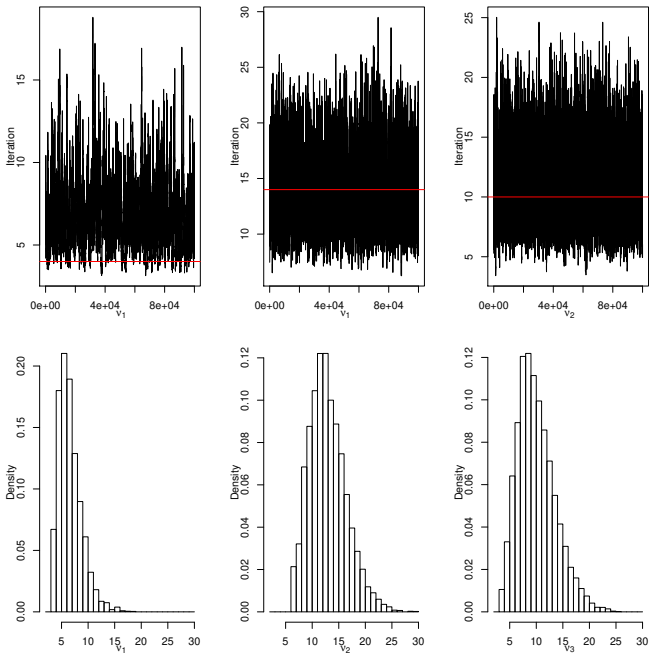

Figure 4.3: Trace plots and histograms of full conditional posterior distributions of degrees of freedom parameters under selected tMAR$(3;2,1,1)$ model, with unit bin-width. Red lines highlight true values.

assumptions in each chapter of Gaussian innovations and Student-t innovations.

We consider the series of first order differences, which can be seen in Figure 4.4. The series presents clear signs of heteroskedasticity, therefore a tMAR model may be a reasonable choice to model the data.

For comparison with previous studies, shifts ϕ_{k0}, $k = 1,\ldots,g$ are fixed to 0, hence are not paramters in the model. This taken into account, our method chooses a tMAR$(2;1,1)$ as best fit among all tMAR models with 2 and 3 mixture components and maximum autoregressive order equal to 4. More specifically, the model was retained about half of the iterations (5067 times over 10000 iterations) by RJMCMC, meaning it is preferred to models with 2 mixing components and larger autoregressive orders. Furthermore, the marginal loglikelihood for this model is -1232.678, which is larger than that of the competing tMAR$(3;2,1,1)$, -1258.073, which was selected as best 3-component model.

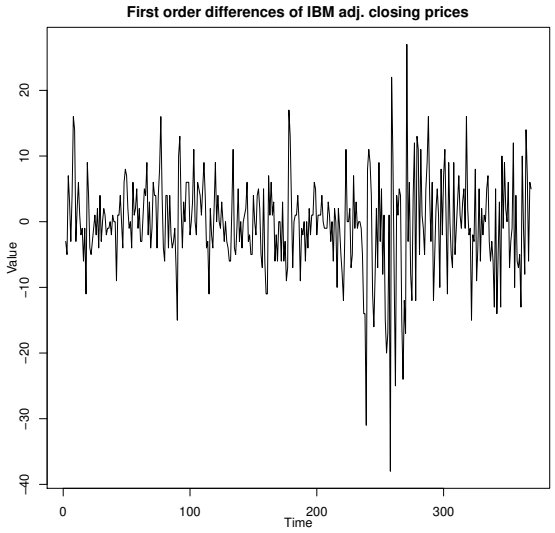

Figure 4.4: Series of first order differences for IBM adjusted closing prices.

Once again, we simulated a sample of size 100000 from the posterior distribution of the parameters, after 10000 burn-in iterations, which can be seen in Figure 4.5.

In Chapter 3, we selected a Gaussian MAR$(3;1,1,4)$ as best fit for the same dataset. In such model, one of the mixture components was "specialised" to model very few observations with large variability. However, the tMAR model, thanks to its flexibility in the tails of the distribution, only requires 2 components to account for such noise, returning a model which is simpler, in that it has fewer parameters, and most importantly has a more straightforward interpretation. This may well result in more accurate estimates of posterior distributions, as the Markov Chain will converge more quickly, and consequently in more accurate and reliable forecasts.

Figure 4.6 shows a comparison of the average density forecast, as described in Section 3.4, between the tMAR$(2;1,1)$ model and the MAR$(3;4,1,1)$ model fitted in Chapter 3. We can see that, regardless of having fewer parameters, the two predictive distributions do not change substantially for the fitted tMAR$(2;1,1)$ with respect to the Gaussian MAR. For the tMAR model, the one-step predictive distribution was

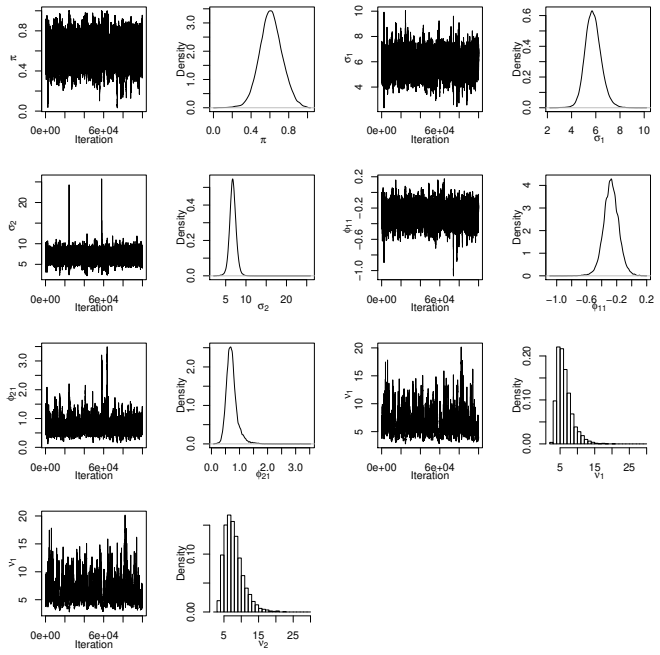

Figure 4.5: Trace and density plots of parameter posterior distributions under selected tMAR$(2;1,1)$ model for the IBM data.

estimated analytically using (4.3), while the two-step predictive distribution was estimated by Monte Carlo simulations.

4.6 Discussion

We have seen a fully Bayesian analysis of mixture autoregressive models with standardised Student-t innovations. In a simulation example, it was shown how the method can correctly find the best model to fit a given dataset. In addition, we saw that the proposed MCMC for simulation from parameter posterior distributions quickly converges to stationarity, and that true values of those parameters are found in high density region.

Later, we showed the analysis performed on the IBM common stock closing prices, a dataset widely exploited in the literature of heteroskedastic models. In particular, we focused on comparison with the analysis of Gaussian MAR models seen in Chapter 3. Results tell that, thanks to the flexibility of the Student-t distribution in its tails,

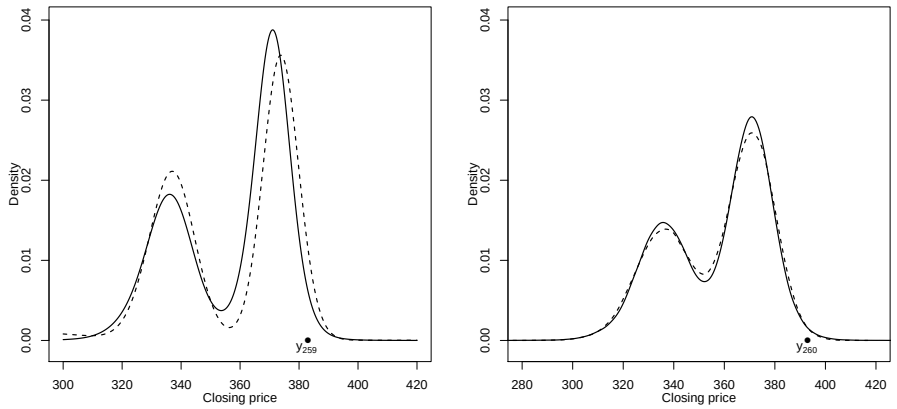

Figure 4.6: One and two step ahead density forecasts at $t = 258$ for the tMAR$(2;1,1)$ model (solid line) and MAR$(3;4,1,1)$ (dashed line) for the IBM closing prices.

we are able now to fit the data with a considerably more parsimonious model, which also has an easier interpretation. The advantage of having a simpler model is that the Markov Chain will converge to its stationary distribution more quickly, meaning that we obtain a more accurate estimate of posterior distributions of the model parameters with a smaller sample (hence also saving computational time). Forecasts will also benefit from this, as they will, in general, be more accurate and reliable.

Chapter 5

Portfolio optimisation with mixture vector autoregressive models

When it comes to multivariate time series, heteroskedasticity implies that the covariance matrix of an observation at a given time point depends upon the recent history of the process. This may be due to changes in the volatility of a single series, as well as in the cross-correlations between any two series of interest. As a result, one cannot trust sample estimates of the (unconditional) covariance matrix, or linear time series models to build reliable predictions about the future. Therefore, obtaining reliable estimates of covariance matrices remains an important challenge in multivariate financial time series for the purpose of portfolio optimisation and financial risk management which use, for instance, modern portfolio theory (Markowitz, 1952).

Bollerslev et al. (1988) and Engle and Kroner (1995) pioneered in the attempt of modelling conditional covariance matrices of predictors for multivariate time series with multivariate GARCH models, using different parametrisations known respectively as VEC and BEKK. Engle (2002) extended the idea of multivariate GARCH to the so called Dynamic Conditional Correlation models, in which each element of

the time-dependent covariance matrix of the data is modelled to follow a GARCH process. Such models have computational advantages over multivariate GARCH models in that the number of parameters to be estimated in the correlation process is independent of the number of series to be correlated, by use of common parameters across involved in the estimation.

Since then, much work has been done to develop multivariate GARCH models, with various applications in finance and econometrics. Of particular interest to us are those attempts attempts which combine multivariate GARCH and factor models, with the aim of dimensionality reduction when modelling large portfolios or panel data. These models rely on the assumption that financial returns are described by a small number of underlying common variables, or factors, which can be used to model the data more parsimoniously. Although all equal in concept, different approaches used different assumptions on the factors, and different techniques are used to derive them. For instance, Alexander (2000) uses a principal components analysis in which factors are assumed to follow independent GARCH processes, whereas Van der Weide (2002) considers the case in which factors are not orthogonal. Finally, Santos and Moura (2014) introduced the dynamic factor GARCH model with time-varying factor loadings.

We propose using a mixture vector autoregressive (MVAR) model (Fong et al., 2007) for portfolio optimisation. MVAR models are the multivariate extension of the mixture autoregressive (MAR) model by Wong and Li (2000). Combining predictive distributions which depend on the recent history of the process, MVAR models can accommodate asymmetry, multimodality, heteroskedasticity and cross-correlation in multivariate time series data. Theoretical properties of MVAR were explored for the case of a multivariate Gaussian mixture in Fong et al. (2007) and Kalliovirta et al. (2016).

Financial returns are typically assumed to be uncorrelated or weakly correlated.

The stationary region of the parameters of MAR and MVAR models contains the uncorrelated case, which allows these properties to be achieved smoothly as part of the estimation process.

Using the Gaussian MVAR model assumption, we are able to fully specify conditional predictive distributions of future observations. We will show how it is possible to combine modern portfolio theory (Markowitz, 1952) and the assumption of Gaussian mixture vector autoregressive model for portfolio optimisation. Under this model assumption, we will also estimate the risk associated with the forecast. Finally, we will compare the performance of our method with that of the dynamic conditional correlation model by Engle (2002) and the vector autoregressive model (VAR).

5.1 The mixture vector autoregressive model

Mixture vector autoregressive models or MVAR (Fong et al., 2007) are the multivariate extension of Mixture Autoregressive Models (Wong and Li, 2000).

The MVAR model with g Gaussian components, and an m dimensional observation vector $\boldsymbol{y}_t$ is defined as

$$F(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \sum_{k=1}^{g} \pi_k \Phi \left(\Omega_k^{-1/2} \left(\boldsymbol{y}_t - \Theta_{k0} - \sum_{i=1}^{p_k} \Theta_{ki} \boldsymbol{y}_{t-i} \right) \right) \tag{5.1}$$

where

- $\boldsymbol{y}_t$ is a $m \times 1$ data vector at time t.
- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ are the mixing weights, such that $0 < \pi_k < 1$ for $k = 1, \ldots, g$, and $\sum_{i=k}^{g} \pi_k = 1$.
- Ω_k is the covariance matrix of component k.

- p_k, $k = 1, \ldots, g$ is the autoregressive order of component k. We denote $p = \max(p_k)$.

- Θ_{k0} is a $m \times 1$ intercept vector for component k, and $\Theta_{k1}, \ldots, \Theta_{kp_k}$ are $m \times m$ matrices of autoregressive parameters. If $p_k < p$, then $\Theta_{kl} = 0_m$ for $p_k < l \leq p$, where 0_m is the zero-matrix of size $m \times m$.

- $\Phi(\cdot)$ is the CDF of the standard multivariate Normal distribution, and $\phi(\cdot)$ is the corresponding pdf.

- Assuming start at $t = 1$, (5.1) holds for $t > p$.

Regularity conditions and parameter estimation by EM algorithm are discussed in Fong et al. (2007) and Kalliovirta et al. (2016).

MVAR may be seen as an alternative to multivariate GARCH when the data presents heteroskedasticity and time-dependent correlation matrices, while also accounting for possible multimodality and asymmetry in the distribution.

For parameter estimation, we resort once again to the missing data formulation. Suppose that a m-variate time series $\{\boldsymbol{y}_t\}$ of length n follows a MVAR process. Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ be an unobserved allocation random variable, where $\boldsymbol{z}_t$ is a g-dimensional vector with component k equal to 1 if $\boldsymbol{y}_t$ comes from the k^{th} component, and 0 otherwise, and such that exactly one element of $\boldsymbol{z}_t$ is equal to 1.

Following notation from Fong et al. (2007), let $\tilde{\Theta}_k = \left[\Theta_{k0}, \Theta_{k1}, \ldots, \Theta_{kp_k}\right]$ and $X_{tk} = \left(1, \boldsymbol{y}_{t-1}^T, \ldots, \boldsymbol{y}_{t-p_k}^T\right)^T$. In addition, let ϑ denote the complete set of parameters. Parameter estimates are then obtained by EM-algorithm (Dempster et al., 1977) with the following steps:

- **E-step**

$$\tau_{tk} = \mathrm{E}\left[z_{tk} \mid y_t, \vartheta\right] = \frac{\pi_k \, \boldsymbol{\phi}\left(\Omega_k^{-1/2}\left(\boldsymbol{y}_t - \Theta_{k0} - \sum_{i=1}^{p_k} \Theta_{ki}\boldsymbol{y}_{t-i}\right)\right)}{\sum_{l=1}^{g} \pi_l \, \boldsymbol{\phi}\left(\Omega_l^{-1/2}\left(\boldsymbol{y}_t - \Theta_{l0} - \sum_{i=1}^{p_l} \Theta_{li}\boldsymbol{y}_{t-i}\right)\right)} \tag{5.2}$$

- **M- step**

$$\begin{aligned}
\hat{\pi}_k &= \frac{1}{n-p}\sum_{t=p+1}^{n} \tau_{tk} \\
\hat{\tilde{\Theta}}_k &= \left(\sum_{t=p+1}^{n} \tau_{tk} X_{tk} X_{tk}^T\right)^{-1}\left(\sum_{t=p+1}^{n} \tau_{tk} X_{tk} \boldsymbol{y}_t^T\right) \\
\hat{\Omega}_k &= \frac{\sum_{t=p+1}^{n} \tau_{tk} e_{tk} e_{tk}^T}{\sum_{t=p+1}^{n} \tau_{tk}}
\end{aligned} \tag{5.3}$$

where $e_{tk} = \boldsymbol{y}_t - \Theta_{k0} - \sum_{i=1}^{p_k} \Theta_{ki}\boldsymbol{y}_{t-i}$.

E-step and M-step are repeated recursively until convergence to maximum likelihood estimates of the parameters.

First and second order stationarity conditions are discussed by Saikkonen (2007) (see also Boshnakov, 2011, or Section 2.1.2 for the univariate case) . Let

$$A_k = \begin{bmatrix}
\Theta_{k1} & \Theta_{k2} & \dots & \Theta_{kp-1} & \Theta_{kp} \\
I_m & 0_m & \dots & 0_m & 0_m \\
0_m & I_m & \dots & 0_m & 0_m \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0_m & 0_m & \dots & I_m & 0_m
\end{bmatrix}, \qquad k = 1, \dots, g \tag{5.4}$$

where I_m and 0_m are respectively the identity matrix and the zero matrix of size $m \times m$.

A necessary and sufficient condition for the MVAR model to be stationary is that the eigenvalues of $\sum_{k=1}^{g} \pi_k A_k \otimes A_k$ are smaller than 1 in modulus. A MVAR model that satisfies this condition is said to be *Stable*. In practice, to assess stability of the fitted model, parameters are replaced by their estimates.

5.1.1 Prediction with mixture vector autoregressive models

In the context of mixture models, density forecasts are often more attractive than point predictors and prediction intervals. This is because the qualitative features of a predictive distribution, such as multiple modes or skewness, are more intuitive and useful than simply a forecast and the associated prediction interval. In addition, when the predictive distribution is available, prediction intervals can easily be obtained by extracting the quantiles of interest (Boshnakov, 2009; Lawless and Fredette, 2005). Therefore, we here present derivation of full predictive distributions for MVAR models, which will be used throughout the analysis.

By model assumption, the one step ahead conditional predictive distribution at time t is fully specified, and it is that of (5.1) where, for notational convenience, we replace t with $t+1$, i.e.

$$F(\boldsymbol{y}_{t+1} \mid \mathcal{F}_t) = \sum_{k=1}^{g} \pi_k \Phi \left(\Omega_k^{-1/2} \left(\boldsymbol{y}_{t+1} - \Theta_{k0} - \sum_{i=1}^{p_k} \Theta_{ki} \boldsymbol{y}_{t+1-i} \right) \right).$$

Thus, the conditional distribution of the one step ahead predictor is a mixture of g Gaussian components and it depends on previous observations. In particular, the conditional covariance matrix depends on previous values of the process, a defining property of heteroskedasticity. To obtain the conditional mean and the covariance matrix

let $\mu_{t+1,k} = \Theta_{k0} + \sum_{i=1}^{p_k} \Theta_{ki} \boldsymbol{y}_{t+1-i}$, for $k = 1, \ldots, g$. Then

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{y}_{t+1} \mid \mathcal{F}_t\right] &= \sum_{k=1}^{g} \pi_k \mu_{t+1,k} = \mu_{t+1} \\ \mathrm{Cov}\left(\boldsymbol{y}_{t+1} \mid \mathcal{F}_t\right) &= \sum_{k=1}^{g} \pi_k \Omega_k + \sum_{k=1}^{g} \pi_k \left(\mu_{t+1,k} - \mu_{t+1}\right)\left(\mu_{t+1,k} - \mu_{t+1}\right)^T \\ &= \sum_{k=1}^{g} \pi_k \Omega_k + \sum_{k=1}^{g} \pi_k \mu_{t+1,k} \mu_{t+1,k}^T - \mu_{t+1}\mu_{t+1}^T \end{aligned} \tag{5.5}$$

Using a method analogous to that of Boshnakov (2009), we can derive the conditional distribution for the two-step ahead predictor as a mixture of g^2 Gaussian components:

$$F\left(\boldsymbol{y}_{t+2} \mid \mathcal{F}_t\right) = \sum_{k=1}^{g} \sum_{l=1}^{g} \pi_k \pi_l \Phi\left(\Psi_{kl}^{-1/2}\left(\boldsymbol{y}_{t+2} - \mu_{kl}\right)\right) \tag{5.6}$$

where, for each $k, l = 1, \ldots, g$,

$$\begin{aligned} \mu_{kl} &= \Theta_{k0} + \Theta_{k1}\Theta_{l0} + \sum_{i=1}^{p-1} \left(\Theta_{k,i+1} + \Theta_{k1}\Theta_{li}\right) \boldsymbol{y}_{t-1-i} + \Theta_{k1}\Theta_{lp}\boldsymbol{y}_{t-1-p} \\ \Psi_{kl} &= \Omega_k + \Theta_{k1}\Omega_l\Theta_{k1}^T \end{aligned}$$

Note that, in general, $\mu_{kl} \neq \mu_{lk}$ and $\Psi_{kl} \neq \Psi_{lk}$. Expectation and covariance matrix of this predictor are:

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{y}_{t+2} \mid \mathcal{F}_t\right] &= \sum_{k=1}^{g} \sum_{l=1}^{g} \pi_k \pi_l \mu_{kl} = \mu_{t+2} \\ \mathrm{Cov}\left(\boldsymbol{y}_{t+2} \mid \mathcal{F}_t\right) &= \sum_{k=1}^{g} \pi_k \pi_l \Psi_{kl} + \sum_{k=1}^{g} \sum_{l=1}^{g} \pi_k \pi_l \mu_{kl} \mu_{kl}^T - \mu_{t+2}\mu_{t+2}^T \end{aligned} \tag{5.7}$$

Full derivation of (5.7), as well as proof of the conditional distribution of $\boldsymbol{y}_{t+2}$, is given below. By recursing this procedure, we could derive a full conditional predictive distribution for any horizon h. However, the number of components in the mixture increases to g^h as h increases, so that analytic expressions may no longer be attractive or meaningful for large g or h. Therefore simulation methods may be preferred for

approximate computation of predictive densities for larger horizons.

Proof of (5.7)**.**

Let $\boldsymbol{z}_t$ be the allocation random variable defined in Section 5.1, and assume $z_{t+2,k} = 1, z_{t+1,l} = 1$ at times $t+2$ and $t+1$ respectively. We have that

$$\begin{aligned}
\boldsymbol{y}_{t+2} &= \mu_{t+2,k} + \Omega^{1/2}\varepsilon_{t+2,k} \\
&= \mu_{t+2,k} - \Theta_{k,1}\boldsymbol{y}_{t+1} + \Theta_{k,1}\boldsymbol{y}_{t+1} + \Omega^{1/2}\varepsilon_{t+2,k} \\
&= \left(\mu_{t+2,k} - \Theta_{k,1}\boldsymbol{y}_{t+1} + \Theta_{k,1}\mu_{t+1,l}\right) + \Theta_{k,1}\Omega_l^{1/2}\varepsilon_{t+1,l} + \Omega^{1/2}\varepsilon_{t+2,k} \\
&= \mu_{t+2;k,l} + \Theta_{k,1}\Omega_l^{1/2}\varepsilon_{t+1,l} + \Omega^{1/2}\varepsilon_{t+2,k}
\end{aligned} \tag{5.8}$$

where $\varepsilon_{t+h,k}$ is the innovation term associated with the k^{th} component.

We wish to predict y_{t+2} using available the information at time t. In order to do this, we require an expression that does not contain y_{t+1}. Hence, we rewrite $\mu_{t+2;k,l}$ as

$$\begin{aligned}
\mu_{t+2;k,l} &= \mu_{t+2,k} - \Theta_{k,1}\boldsymbol{y}_{t+1} + \Theta_{k,1}\mu_{t+1,l} \\
&= \Theta_{k,0} + \sum_{i=1}^{p}\Theta_{k,i}y_{t+2-i} - \Theta_{k,1}\boldsymbol{y}_{t+1} + \Theta_{k,1}\left(\Theta_{l,0} + \sum_{i=1}^{p}\Theta_{l,i}\boldsymbol{y}_{t+1-i}\right) \\
&= \Theta_{k,0} + \Theta_{k,1}\Theta_{l,0} - \Theta_{k,1}\boldsymbol{y}_{t-1} + \Theta_{k,1}\boldsymbol{y}_{t+1} + \sum_{i=2}^{p}\Theta_{k,i}\boldsymbol{y}_{t+2-i} \\
&\quad + \Theta_{k,1}\sum_{i=1}^{p}\Theta_{l,i}\boldsymbol{y}_{t+1-i} \\
&= \Theta_{k,0} + \Theta_{k,1}\Theta_{l,0} + \sum_{i=1}^{p-1}\Theta_{k,i+1}\boldsymbol{y}_{t+1-i} + \Theta_{k,1}\sum_{i=1}^{p}\Theta_{l,i}\boldsymbol{y}_{t+1-i} \\
&= \Theta_{k,0} + \Theta_{k,1}\Theta_{l,0} + \sum_{i=1}^{p-1}\left(\Theta_{k,i+1} + \Theta_{k,1}\Theta_{l,i}\right)\boldsymbol{y}_{t+1-i} + \Theta_{k,1}\Theta_{l,p}\boldsymbol{y}_{t+1-p}
\end{aligned}$$

And therefore we have the expression for $\boldsymbol{y}_{t+2}$

$$\boldsymbol{y}_{t+2} = \Theta_{k,0} + \Theta_{k,1}\Theta_{l,0} + \sum_{i=1}^{p-1} \left(\Theta_{k,i+1} + \Theta_{k,1}\Theta_{l,i}\right) \boldsymbol{y}_{t+1-i} + \Theta_{k,1}\Theta_{l,p}\boldsymbol{y}_{t+1-p} + \Theta_{k,1}\Omega_l^{1/2}\varepsilon_{t+1,l} + \Omega_k^{1/2}\varepsilon_{t+2,k}$$

We deduce that, given observed $\boldsymbol{z}_{t+2}, \boldsymbol{z}_{t+1}$:

$$\begin{aligned}
&\mathrm{E}[\boldsymbol{y}_{t+2} \mid \boldsymbol{z}_{t+2}, \boldsymbol{z}_{t+1}, \mathcal{F}_t] \\
&\quad = \Theta_{k,0} + \Theta_{k,1}\Theta_{l,0} + \sum_{i=1}^{p-1} \left(\Theta_{k,i+1} + \Theta_{k,1}\Theta_{l,i}\right) \boldsymbol{y}_{t+1-i} + \Theta_{k,1}\Theta_{l,p}\boldsymbol{y}_{t+1-p} \\
&\mathrm{Cov}\left(\boldsymbol{y}_{t+2} \mid \boldsymbol{z}_{t+2}, \boldsymbol{z}_{t+1}, \mathcal{F}_t\right) = \Theta_{k,1}\Omega_l\Theta_{k,1} + \Omega_k
\end{aligned}$$

The conditional distribution of $\boldsymbol{y}_{t+2}$ can be then derived through its characteristic function. Recall the characteristic function for the multivariate normal distribution and $\boldsymbol{y}_{t+1}$ can be written as

$$\varphi_{t+1|t} \equiv \mathrm{E}\left[e^{is^T \boldsymbol{y}_{t+1}} \mid \mathcal{F}_{t-1}\right] = \mathrm{E}\left[\sum_{k=1}^{g} \pi_k e^{is^T \mu_{t+1;k}} \varphi_k(\Omega_K^{1/2} s)\right]$$

It follows that, for y_{t+2}, we have

$$\begin{aligned}
\varphi_{t+2|t}(s) &\equiv \mathrm{E}\left[e^{is^T y_{t+2}} \mid \mathcal{F}_t\right] = \mathrm{E}\left[\mathrm{E}\left(e^{is^T \boldsymbol{y}_{t+2}} \mid \boldsymbol{z}_{t+2}, \boldsymbol{z}_{t+1}, \mathcal{F}_t\right) \mid \mathcal{F}_t\right] \\
&= \mathrm{E}\left[is^T \mu_{t+2;k,l}\, \mathrm{E}\left(e^{\Theta_{k,1}\Omega_l^{1/2}\varepsilon_{t+1,l} + \Omega^{1/2}\varepsilon_{t+2,k}} \mid \boldsymbol{z}_{t+2}, \boldsymbol{z}_{t+1}, \mathcal{F}_t\right) \mid \mathcal{F}_t\right] \\
&= \sum_{k,l=1}^{g} \pi_k \pi_l e^{is^T \mu_{t+2;k,l}} \varphi_1(\Theta_{k,1}\Omega_l^{1/2} s) \varphi_2(\Omega_k^{1/2} s)
\end{aligned}$$

Thus, the conditional distribution of y_{t+2} given $\mathcal{F}_t$ is a mixture of g^2 components with mixing weights $\pi_k\pi_l$. For a normal mixture, we also have that:

$$\varphi_1(\Theta_{k,1}\Omega_l^{1/2} s)\varphi_2(\Omega_k^{1/2} s) = e^{\Theta_{k,1}\Omega_l\Theta_{k,1}^T} e^{\Omega_k} = e^{\Theta_{k,1}\Omega_l\Theta_{k,1}^T + \Omega_k}$$

which shows that the conditional distribution of the two-step predictor is a mixture of Normals with means $\mu_{t+2;k,l}$ and covariance matrices $\Theta_{k,1}\Omega_l\Theta_{k,1}^T + \Omega_k$.

5.2 Portfolio optimisation with MVAR models

Suppose that a multivariate time series $\{\boldsymbol{y}_t\}$ of asset returns is observed, and it is believed that the underlying generating process is MVAR. From Section 5.1.1, conditional distributions of the 1 and 2 step predictors are fully specified, and can be estimated by plugging parameter estimates into the relevant equations.

Now, let w denote the weights of a portfolio built with assets $\{y_t\}$ (allowing short selling), and let $R_{t+1} = w^T \boldsymbol{y}_{t+1}$ be the portfolio return at time $t+1$. Intuitively, because our model consists of a mixture of multivariate normal components, we can apply the property in (2.44) to conclude that the conditional distribution of R_{t+1} is also (univariate) mixture normal, with corresponding mixing weights $\boldsymbol{\pi}$ from the fitted multivariate model. By model assumption in fact, at each time $t+1$ an observation y_{t+1} is assumed to be generated from one of g components of the mixture. Consequently, R_{t+1} is obtained by applying (2.44) to the selected component. Recursing this for all g components the result is itself a mixture distribution for R_{t+1}.

In terms of MVAR model parameters we write:

$$F(R_{t+1} \mid \mathcal{F}_t) = \sum_{k=1}^{g} \pi_k \Phi\left(\frac{R_{t+1} - w^T \mu_{t+1,k}}{\sqrt{w^T \Omega_k w}}\right) \tag{5.9}$$

Conditional mean and variance of R_{t+1} are:

$$\begin{aligned} \mathrm{E}[R_{t+1} \mid \mathcal{F}_t] &= \sum_{k=1}^{g} \pi_k \left(w^T \mu_{t+1,k}\right) = \sum_{k=1}^{g} \pi_k \mu^*_{t+1,k} = \mu^* \\ \mathrm{Var}(R_{t+1} \mid \mathcal{F}_t) &= \sum_{k=1}^{g} \pi_k \left(w^T \Omega_k w\right) + \sum_{k=1}^{g} \pi_k \left(\mu^*_{t+1,k}\right)^2 - (\mu^*)^2 \end{aligned} \tag{5.10}$$

where $\mu_{t+1,k}$ is the same as in (5.5).

Modern portfolio theory (Markowitz, 1952) gives us a way to calculate weights w^* to construct the most efficient portfolio for a given return, and to calculate the efficient portfolio of assets with the minimum possible variance. A portfolio with target return μ is said to be an efficient portfolio when the variance associated with it is the lowest amongst all portfolios of the same assets having that same target return. The minimum variance portfolio is the efficient portfolio with the lowest possible variance of all efficient portfolios of the same assets (see also Section 2.8.1 for more details). For the remainder of the analysis, we will denote efficient portfolios with the subscript EFF, and minimum variance portfolios with the subscript MVP. We now see how modern portfolio theory can be used to predict future observations assuming a MVAR model.

For the MVAR case, let $\mathrm{E}[\boldsymbol{y}_{t+1} \mid \mathcal{F}_t] = \mu_{t+1}$ and $\mathrm{Cov}(\boldsymbol{y}_{t+1} \mid \mathcal{F}_t) = \Omega_{t+1}$. Recall expressions of the quantities (2.55), revisited for MVAR models:

$$A = \mathbb{1}\Omega_{t+1}^{-1}\mu_{t+1} \quad , \quad B = \mu_{t+1}\Omega_{t+1}^{-1}\mu_{t+1} \quad , \quad C = \mathbb{1}\Omega_{t+1}^{-1}\mathbb{1} \quad , \quad D = CB - A^2 \tag{5.11}$$

where $\mathbb{1}$ is a vector of 1s of the same length as μ_{t+1}.

It can be proved that optimal weights for an efficient portfolio of these assets and target return μ_{EFF} are

$$w_{\mathrm{EFF}} = \frac{1}{D}\left(B\Omega_{t+1}^{-1}\mathbb{1} - A\Omega_{t+1}^{-1}\mu_{t+1} + \mu^*\left(C\Omega_{t+1}^{-1}\mu_{t+1} - A\Omega_{t+1}^{-1}\mathbb{1}\right)\right) \tag{5.12}$$

and the variance of such portfolio can be calculated equivalently as $Var(R_t \mid \mathcal{F}_{t-1})$ (MVAR model assumption) or $w^T\Omega_t w$ (the variance of an efficient portfolio of assets)

since

$$w_{\text{EFF}}^T \Omega_{t+1} w_{\text{EFF}} = \sum_{k=1}^{g} \pi_k \left(w_{\text{EFF}}^T \Omega_k w_{\text{EFF}} \right) + \sum_{k=1}^{g} \pi_k \left(w_{\text{EFF}}^T \mu_{t+1,k} \right)^2 - \left[\sum_{k=1}^{g} \pi_k \left(w_{\text{EFF}}^T \mu_{t+1,k} \right) \right]^2 = Var(R_{t+1} \mid \mathcal{F}_t) \tag{5.13}$$

In practice, $\mu_{t+1,k}$ and Ω_{t+1} are replaced with their estimates $\hat{\mu}_{t+1,k}$ and $\hat{\Omega}_{t+1}$.

Weights of the minimum variance portfolio of same assets $\{\boldsymbol{y}_t\}$, and corresponding return, are:

$$w_{\text{MVP}} = \frac{\Omega_{t+1}^{-1} \mathbb{1}}{C} \qquad \mu_{\text{MVP}} = \frac{A}{C} \tag{5.14}$$

Conditional predictive distributions can also be calculated analytically for any $h \geq 2$. However, one must keep in mind that such predictive distribution would be a mixture of g^h components, and their computations be cumbersome, so that simulation methods may be preferred in some cases as h increases.

Consider the case $h = 2$. The conditional predictive distribution $F(\boldsymbol{y}_{t+2} \mid \mathcal{F}_t)$ for the MVAR model is a mixture of g^2 Gaussian components given in (5.7). Similarly to the case $h = 1$, we can derive the full conditional distribution of R_{t+2}, which is again a mixture of g^2 Gaussian components:

$$F(R_{t+2} \mid \mathcal{F}_t) = \sum_{k,l=1}^{g} \pi_k \pi_l \Phi \left(\frac{R_{t+2} - w^{(2)} \mu_{kl}}{w^{(2)^T} \Psi_{kl} w^{(2)}} \right) \tag{5.15}$$

where $w^{(2)}$ is the vector of optimal weights for this portfolio, indicating we are predicting 2 steps into the future. Similarly to the case $h = 1$, one can now calculate $\mathrm{E}[\boldsymbol{y}_{t+2} \mid \mathcal{F}_t] = \mu_{t+2}$ and $\mathrm{Cov}(\boldsymbol{y}_{t+2} \mid \mathcal{F}_t) = \Omega_{t+2}$ and adapt (2.55), (2.56), (5.13) and (2.57) to obtain an efficient or minimum variance portfolio.

5.3 Simulated data example

We simulate a series of size $n = 500$ of hypothetical stock returns from the $3-$variate MVAR$(2;1,1)$ process with CDF

$$F(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = 0.75\boldsymbol{\Phi}\left(\frac{\boldsymbol{y}_t - v_1}{\Omega_1}\right) + 0.25\boldsymbol{\Phi}\left(\frac{\boldsymbol{y}_t - v_2}{\Omega_2}\right)$$

where

$$v_1 = \Theta_{10} + \Theta_{11}\boldsymbol{y}_{t-1} \qquad v_2 = \Theta_{20} + \Theta_{21}\boldsymbol{y}_{t-1}$$

and

$$\Theta_{10} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad \Theta_{11} = \begin{bmatrix} 0.5 & 0 & 0.4 \\ -0.3 & 0 & 0.5 \\ -0.6 & 0.5 & -0.3 \end{bmatrix} \Omega_1 = \begin{bmatrix} 1 & 0.5 & -0.40 \\ 0.5 & 2 & 0.8 \\ -0.4 & 0.8 & 4 \end{bmatrix}$$

$$\Theta_{20} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad \Theta_{21} = \begin{bmatrix} -0.5 & 1 & -0.4 \\ 0.3 & 0 & -0.2 \\ 0 & -0.5 & 0.5 \end{bmatrix} \Omega_2 = \begin{bmatrix} 1 & 0.2 & 0 \\ 0.2 & 2 & -0.55 \\ 0 & -0.55 & 4 \end{bmatrix}$$

The three univariate series can be seen in Figure 5.1, with their autocorrelation and cross-correlation plots in Figure 5.2. The data is very representative of what we should be looking for, in a real case scenario, to assume an underlying MVAR process. We notice in fact signs of heteroskedasticity in each of the series, and autocorrelations and cross-correlations significantly different from 0 at lags larger than 0. The latter is what separates MVAR from multivariate GARCH models, which assume the original series to be uncorrelated.

Parameter estimates are calculated using the EM Algorithm with the formulas in (5.2) and (5.3). In order to perform out of sample prediction, data from $\boldsymbol{y}_1$ to $\boldsymbol{y}_{498}$ were used for estimation. This leaves $\boldsymbol{y}_{499}$ and $\boldsymbol{y}_{500}$ out as observations 1 and 2 time points

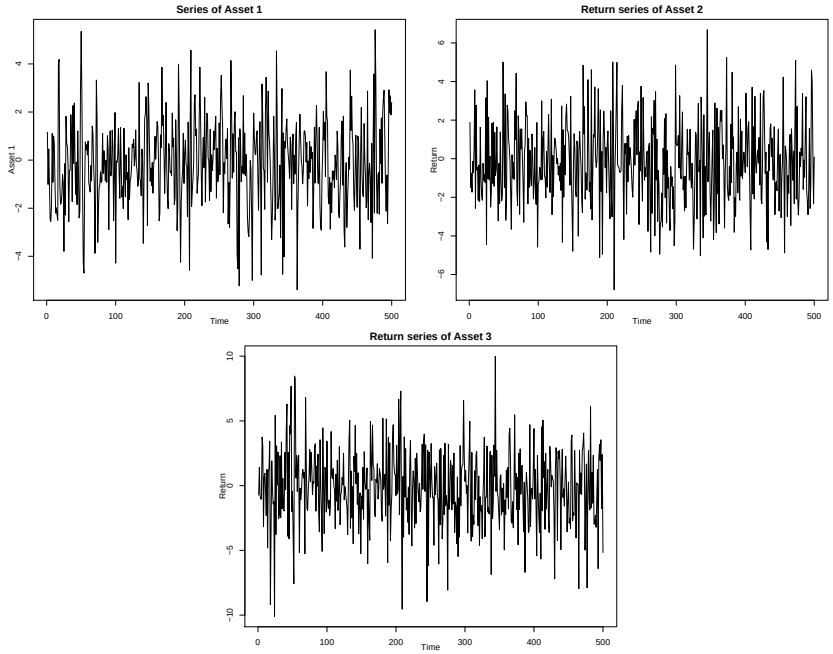

Figure 5.1: Simulated time series of stock returns Asset 1 (top left), Asset 2 (top right) and Asset 3 (bottom).

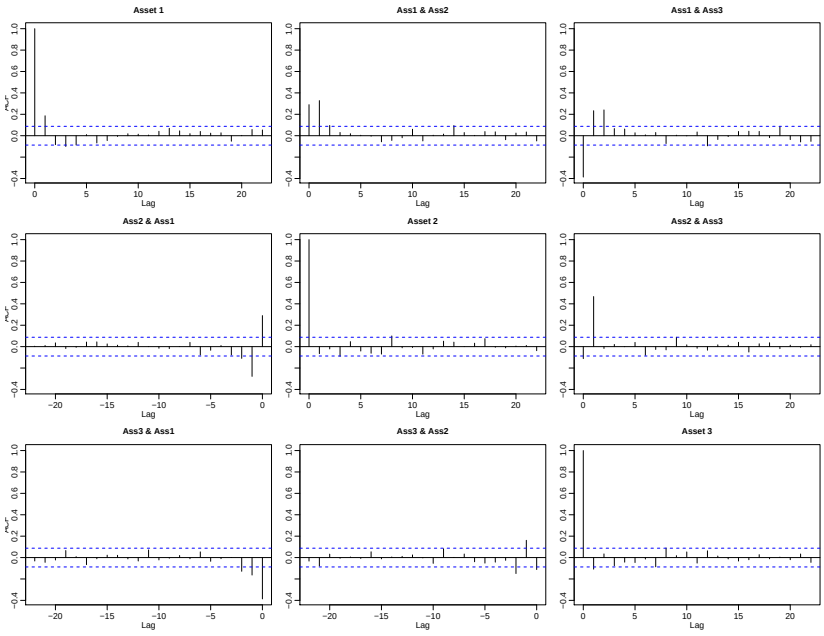

Figure 5.2: Autocorrelation and corss-correlation plots of the simulated time series data.

in the future, to be predicted:

$$\hat{\boldsymbol{\pi}} = (0.7242, 0.2758)$$

$$\Theta_{10} = \begin{bmatrix} -0.0022 \\ -0.0303 \\ 0.1276 \end{bmatrix} \quad \hat{\Theta}_{11} = \begin{bmatrix} 0.4931 & -0.0339 & 0.4169 \\ -0.3156 & -0.0012 & 0.5078 \\ -0.6141 & 0.6007 & -0.3844 \end{bmatrix} \hat{\Omega}_1 = \begin{bmatrix} 0.9551 & 0.4783 & -0.2776 \\ 0.4783 & 1.9123 & 0.9736 \\ -0.2776 & 0.9736 & 3.9455 \end{bmatrix}$$

$$\hat{\Theta}_{20} = \begin{bmatrix} 0.0338 \\ 0.5499 \\ -0.7580 \end{bmatrix} \quad \hat{\Theta}_{21} = \begin{bmatrix} -0.4595 & 1.0124 & -0.4004 \\ 0.3343 & -0.1423 & -0.1551 \\ -0.1273 & -0.2336 & 0.6509 \end{bmatrix} \hat{\Omega}_2 = \begin{bmatrix} 0.8767 & 0.4794 & -0.3627 \\ 0.4794 & 2.9148 & -0.6576 \\ -0.3627 & -0.6576 & 9.8135 \end{bmatrix}$$

We then calculate the one step ahead conditional mean and variance based on parameter estimates:

$$\mathrm{E}[\boldsymbol{y}_{499} \mid \mathcal{F}_{498}] = \hat{\mu}_{499} = \begin{bmatrix} -0.1750 \\ -0.9655 \\ -1.4361 \end{bmatrix}$$

$$\mathrm{Cov}(\boldsymbol{y}_{499} \mid \mathcal{F}_{498}) = \hat{\Omega}_{499} = \begin{bmatrix} 1.3109 & -0.6080 & -0.0768 \\ -0.6080 & 5.3174 & -0.5642 \\ -0.0768 & -0.5642 & 5.9420 \end{bmatrix}$$

Given $\hat{\Omega}_{499}$, we can calculate the minimum variance portfolio, which is obtained for weights $w_{\mathrm{MVP}} = (0.6434, 0.2228, 0.1338)^T$. The corresponding expected return on this portfolio at time 499 is $\mu_{\mathrm{MVP}} = -0.5198$, with standard deviation $\sigma_{\mathrm{MVP}} = 0.8475$.

Suppose now that we wish to increase our return to $\mu^* = 0$, i.e. no expected loss, at the cost of a larger variance. We can calculate weights to construct an efficient

portfolio of these assets as seen in Section 5.2. We obtain:

$$w_{\text{EFF}} = \begin{bmatrix} 1.1097 \\ 0.0781 \\ -0.1878 \end{bmatrix}$$

The interpretation of w_{EFF} is that the otpimal portfolio yielding expected return of 0 is constructed by short-selling a small amount of Asset 3, and investing 110.97% and 7.81% of the total capital (meanwhile increased by short selling) into Asset 1 and Asset 2 respectively. Notice that the target return is $w_{\text{EFF}}\mu_t = \mu_{\text{EFF}} = 0$ as desired.

We can now calculate the quantities we need for the conditional predictive distribution of R_{499}:

$$\mu_1^* = w^* \mu_{499,1} = 0.2642 \qquad \sigma_1^{*^2} = 1.4968 \approx (1.2235)^2$$
$$\mu_2^* = w^* \mu_{499,2} = -0.6939 \qquad \sigma_2^{*^2} = 1.6062 \approx (1.3025)^2$$

Therefore, the conditional distribution of $R_{499} = w^{*^T} \boldsymbol{y}_{499}$ is

$$F(R_{499} \mid \mathcal{F}_{498}) = 0.7242 \times \Phi\left(\frac{R_{499} - 0.2642}{1.2235}\right) + 0.2758 \times \Phi\left(\frac{R_{499} - 0.6939}{1.3025}\right)$$

The standard deviation associated to this portfolio is $\sigma_{\text{EFF}} = 1.3173$ which, as expected, is larger than σ_{MVP}. More importantly, we can use the distribution assumption on R_{499} to estimate risk measures. Figure 5.3 shows the conditional distribution of R_{499}. The dot on the left hand side of the figure, highlighted with a dashed line, is the value at risk at 95% level. We find that the value at risk at such level is 2.2039, with expected shortfall of 2.7912. This means that an investor could expect a loss on this portfolio larger than 2.2039 units with probability 0.05, and when this threshold is exceeded, the expected loss is of 2.7912 units. The observed return is also shown in

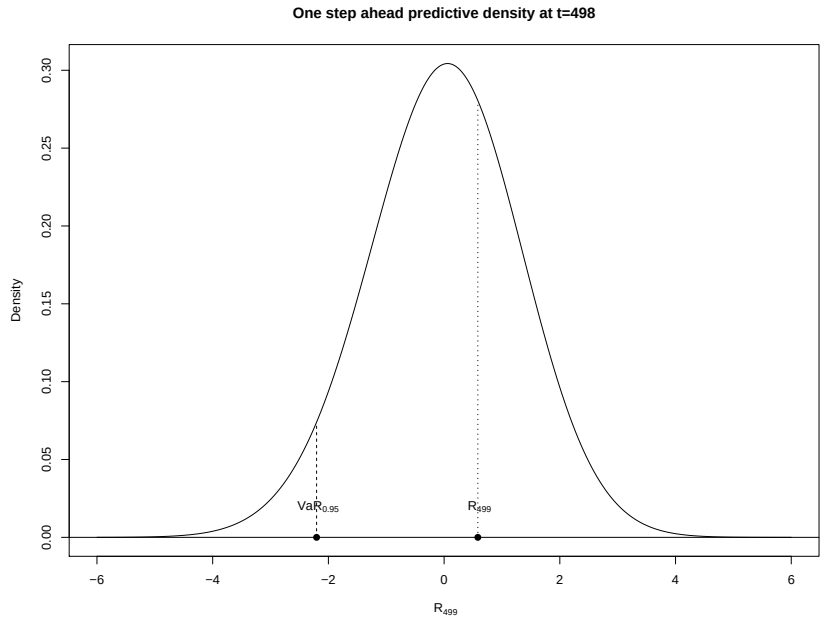

Figure 5.3: Conditional one-step predictive density of R_{499}, with VaR at 95% (dashed line) and observed return (dotted line) highlighted.

Figure 5.3 as a dot with dotted line. We notice that it lies on a region of high density of the predictive distribution.

We can also estimate the conditional distribution of the two-step ahead predictor at $t = 498$, $F(R_{500} \mid \mathcal{F}_{498})$. This is shown in Figure 5.4).

The minimum variance portfolio for a a two-step ahead portfolio of assets is calculated with weights $w^{(2)}_{\text{MVP}} = (0.4367, 0.2822, 0.2811)$, with an expected return $\mu^{(2)}_{\text{MVP}} = -0.3918$, with $\sigma^{(2)}_{\text{MVP}} = 1.1784$, showing the increasing uncertainty as we attempt to predict further into the future.

Once again we consider building a portfolio of assets yielding expected return $\mu^* = 0$. Optimal weights for this portfolio are

$$w^{(2)}_{\text{MVP}} = \begin{bmatrix} -0.9404 \\ 1.5193 \\ 0.4211 \end{bmatrix}$$

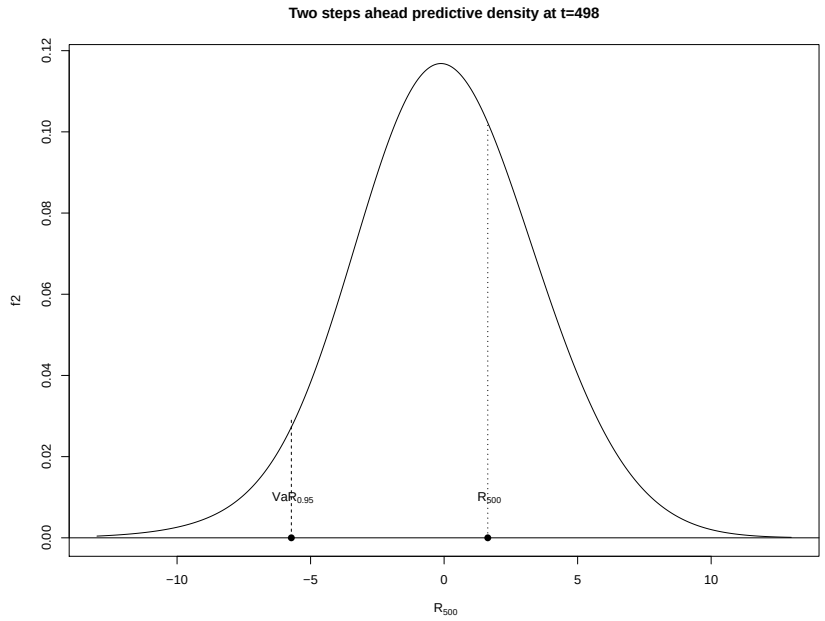

Figure 5.4: Conditional two-step predictive density of R_{500}, with VaR at 95% (dashed line) and observed return (dotted line) highlighted.

From Figure 5.4, we notice how the density is now flatter, which is sign of a larger variability. This is confirmed by an estimated standard deviation of 3.5056 for the two-step predictorR_{500}, which is a significant increase. This also results in much larger estimated VaR $= -5.0207$ (in absolute value) at the same 95% level, with expected shortfall equal to -7.4505. Once again, the observed return (dotted line) is in a high density region of the predictive distribution. This was all to be expected, since it is reasonable to think that forecasts will become less and less accurate as we try to predict further in the future.

Overall, we can be satisfied with the performance of our method in predicting portfolio returns.

5.4 Application to the US stock market

We consider a multivariate dataset of $m = 4$ stocks on the US stock market: Dell Technologies Inc. (`DELL`), Microsoft Corporation (`MSFT`), Intel Corporation (`INTC`), and International Business Machine Corporation (`IBM`). The data were obtained from Yahoo! Finance (`https://finance.yahoo.com`). The original time series include daily Adjusted Close Prices between January 2^{nd} 2016 and January 29^{th} 2020 (867 observations). For each series and $t = 2, \ldots, 867$, we calculated daily returns as $(\text{Price}_t - \text{Price}_{t-1})/\text{Price}_{t-1}$. The resulting series, displayed individually in Figure 5.5, includes 866 observations.

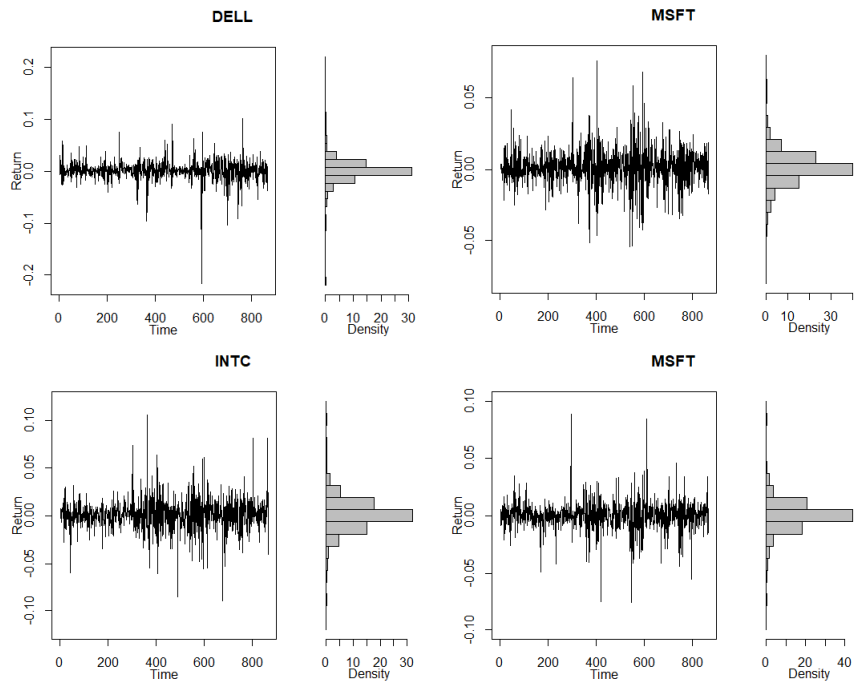

Figure 5.5: Time series of returns of DELL (top left), MSFT (top right), INTC (bottom left) and IBM (bottom right).

All four univariate series in Figure 5.5 show signs of heteroskedasticity. The histograms show signs of heavy tails, too. This second feature was confirmed by calculation of sample excess kurtosis (all significantly larger than 0). In addition, from a

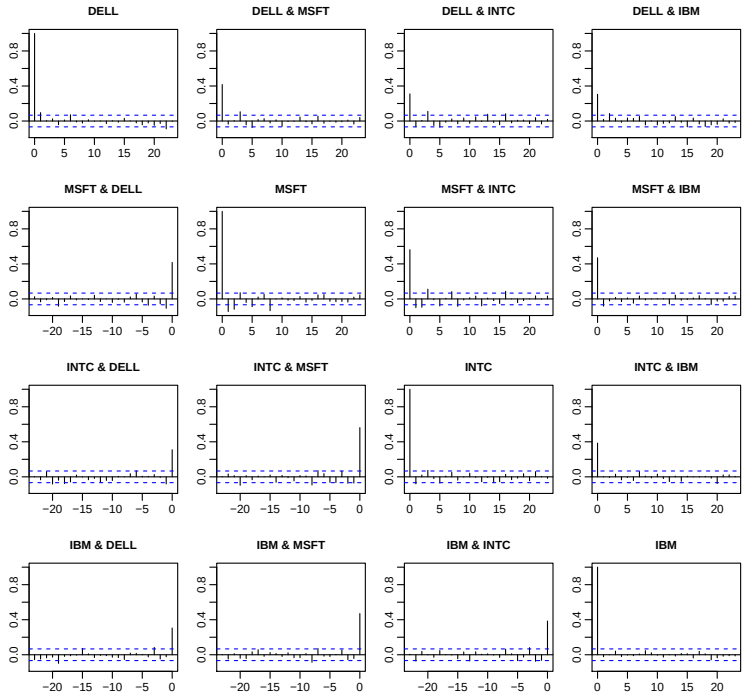

Figure 5.6: Autocorrelation and cross-correlation plots for the multivariate time series. Notice the presence of correlation and cross-correlation in the data.

preliminary analysis, it was noticed that the data presents autocorrelation at least at lags 1 and 2, and cross-correlations at lags 0, 1 and 2 (see Figure 5.6). Therefore, it is reasonable to consider a MVAR generating process for the data.

Several models were fitted. In terms of diagnostics, a MVAR$(3;3,2,1)$ was chosen as best fit. Estimation was carried out on the first 864 observations, omitting the last two for out-of-sample prediction.

Given parameter estimates and the one-step ahead predictive distribution at $t = 864$, we calculate weights for the minimum variance portfolio built with these assets, which yields a mean return of approximately $0.0024\,(0.24\%)$. The standard deviation associated with this portfolio is $\sigma_{\text{MVP}} = 0.0092$, with weights $w_{MVP} = (0.1319, 0.4597, -0.0136, 0.4220)$.

Now, assume we would like to increase our mean return to $0.007 = 0.7\%$. We can

calculate optimal weights

$$w_{\text{EFF}} = (-0.5832, 0.9538, 0.1085, 0.5209)^T .$$

Weights are interpreted as follows: an investor shall short-sell an amount of around 0.58 times their inital capital in DELL stocks, and reinvest the new total in the remaining three assets, with a major bet on MSFT and IBM. The idea behind this is that it is believed that DELL stocks will decrease in value between the present and the nearest future, and therefore one could short-sell to make a profit. On the other hand, it is believed that the remaining three assets will increase their value in the same time span, and in particular MSFT stocks. However, the standard deviation associated with this portfolio is $\sigma_{\text{EFF}} = 0.0139$, a slight increase compared to σ_{MVP}, considering the scale of the data.

For the latter portfolio, we calculate the one-step ahead conditional distribution of $R_{865} = \sum_{m=1}^{4} w_m^* y_{m,865}$ using parameter estimates from the MVAR model fitting:

$$\begin{aligned} F(R_{865} \mid \mathcal{F}_{864}) =& 0.1316\Phi\left(\frac{R_{865}+0.00052}{0.0266}\right) + 0.5627\Phi\left(\frac{R_{865}+0.00178}{0.0093}\right) \\ &+ 0.3057\Phi\left(\frac{R_{865}+0.01932}{0.0169}\right) \end{aligned}$$

The corresponding predictive density can be seen in Figure 5.7.

Value at risk at $\alpha = 95\%$ is estimated at 0.0174, with expected shortfall of 0.0299. The subsequently observed return is $R_{865} = -0.0062$, which we can see lies on a region of high density, and therefore is somewhat plausible.

We can also look at building a portfolio of the same assets looking two steps into the future, at $t = 866$. The minimum variance portfolio in this case yields an expected return $\mu_{\text{MVP}}^{(2)} = -0.011$, with associated standard deviation $\sigma_{\text{MVP}}^{(2)} = 0.0101$. The optimal weights for this minimum variance portfolio were estimated as $(0.1278, 0.2366, 0.1700, 0.4655)$.

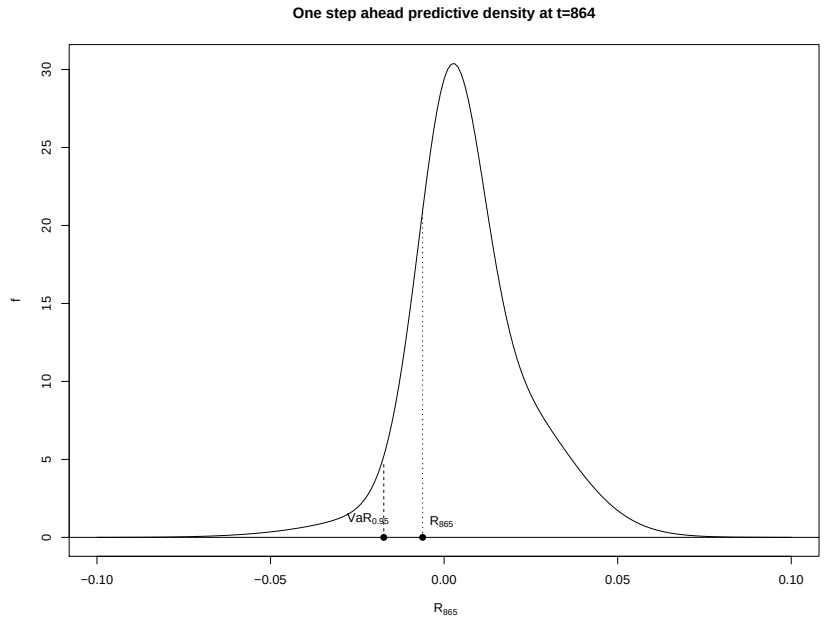

Figure 5.7: Conditional one-step predictive density of R_{865}, with VaR at 95% (dashed line) and observed return (dotted line) highlighted.

We then use the distribution assumptions on $\boldsymbol{y}_{866}$, its expected value and covariance matrix to estimate optimal weights to look once again to increasing our return by bulding an efficient portfolio with same target return $\mu^* = 0.007$ as before:

$$w_{\text{EFF}}^{(2)} = (0.0402, 0.6174, 0.7554, -0.4130)^T$$

Using parameter estimates and (5.15), the conditional distribution of R_{866} is a mixture of $3^2 = 9$ components:

$$\begin{aligned} F(R_{866} \mid \mathcal{F}_{864}) =& 0.0173\ \Phi\left(\frac{R_{866} - 0.0170}{0.0285}\right) + 0.0741\ \Phi\left(\frac{R_{866} - 0.0190}{0.0268}\right) \\ &+ 0.0402\ \Phi\left(\frac{R_{866} - 0.0252}{0.0276}\right) + 0.0741\ \Phi\left(\frac{R_{866} - 0.0086}{0.0101}\right) \\ &+ 0.3166\ \Phi\left(\frac{R_{866} - 0.0069}{0.0096}\right) + 0.1720\ \Phi\left(\frac{R_{866} - 0.0051}{0.0096}\right) \\ &+ 0.0402\ \Phi\left(\frac{R_{866} - 0.0098}{0.0206}\right) + 0.1720\ \Phi\left(\frac{R_{866} - 0.0040}{0.0187}\right) \\ &+ 0.0935\ \Phi\left(\frac{R_{866} + 0.0077}{0.0196}\right) \end{aligned}$$

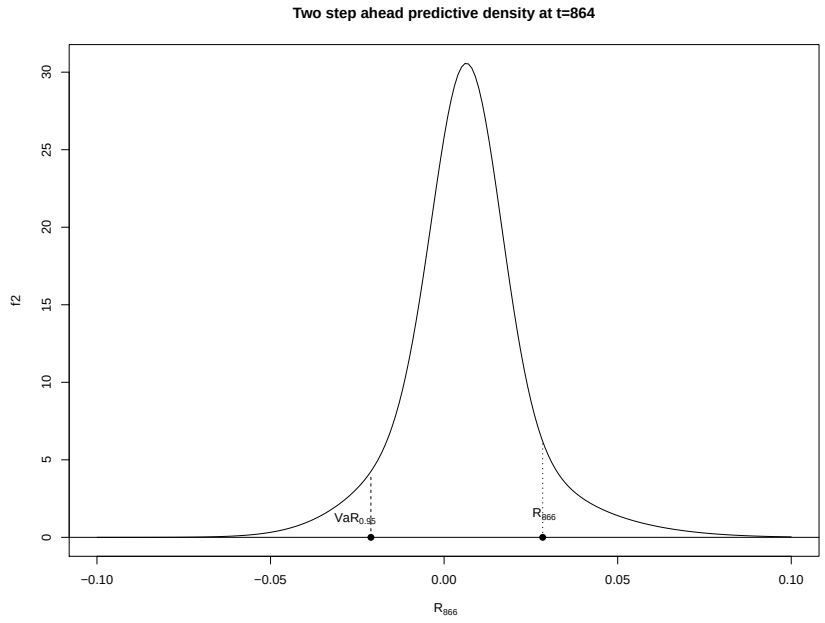

Figure 5.8: Conditional two-step predictive density of R_{866}, with VaR at 95% (dashed line) and observed return (dotted line) highlighted.

The conditional distribution of R_{866} can be seen in Figure 5.8. We notice, as expected, an increase in the standard deviation of the distribution, $\sigma^{(2)}_{\text{EFF}} = 0.0177$ with respect to $\sigma_{\text{EFF}} = 0.0177$. Overall, the two shapes in Figure 5.7 and 5.8 look similar, however the observed return R_{866} is not in a high density region of its predictive distribution, as it actually exceeds the expectations. VaR is now estimated at -0.021, with expected shortfall equal to -0.0315.

5.5 Comparison of VAR, MVAR and DCC

Dynamic Conditional Correlation models (DCC, Engle, 2002) are a class of multivariate GARCH models in which conditional correlations between elements of a vector series are time dependent. In particular, given the conditional covariance matrix of the model at time t, H_t, each entry h_{ij} of the matrix is modelled as a univariate GARCH model. DCC models are used in finance to predict behavior of vector time series in

which the assets are correlated and heteroskedastic, thanks to the fact that the conditional covariance matrix of a predictor is always fully specified (see Section 2.7.1 for more details).

In his model, Engle (2002) states that correlation is based (i.e. conditional) on information known the previous period, and that correlation matrices of multi-period forecasts are similarly defined. According to this, each time point produces a different conditional correlation matrix. A similar argument can be presented for MAR and MVAR models, except we work with conditional covariance matrices, rather than correlations. The conditional covariance matrix clearly depends on past observations, and one unique conditional covariance matrix is produced at each time point, except for some limiting cases. As seen in Section 5.2, this also applies to one and multi-step forecasts. For this reason, we consider a comparison between MVAR and DCC to be appropriate.

We compare here the performance of modelling the data in Section 5.4 with an MVAR model and a DCC model. We also add a comparison with a fitted vector autoregressive model of order 3 (VAR(3)). We use a rolling-window setup for comparison of the density forecasts.

First, we consider a window from the first observation on January 2^{nd} 2016 to October 10^{th} 2019, thus including 766 observations (roughly 90% of the available data). We use this to estimate MVAR, DCC and VAR models, and derive one and two steps density forecasts for the three models for October 11^{th} and October 12^{th}, and calculate forecasting accuracy using their respective observed values. We then move the window forward by one day. The window would now contain 766 observations from from the first observation on January 3^{rd} 2016 to October 11^{th} 2019. The procedure is repeated until the window contains the most up to date observation on January 28^{th} 2020, which allows a one step forecast. Therefore, we obtain 100 one step density forecasts, and 99 two step density forecasts.

Comparison of density forecasts is ususally done using scoring rules. Here we compare the three models in terms of some strictly proper scoring rules: Continuous Ranked Probability Score (CRPS, see for instance Gneiting and Raftery, 2007), logarithmic score (LogS, Good, 1952), and the Dawid-Sebastiani score (Dawid and Sebastiani, 1999, DSS,). For more details on these, see Section 2.5.2.

A DCC-GARCH(1,1) was found to be the best model for the return series assuming multivariate normal innovations. For this, and for the VAR(3) model, the same rolling window procedure as for MVAR is performed. For each forecast of each model, we calculate CRPS, LogS and DSS, and take the average score for comparison. Results can be seen in Tables 5.1 and 5.2: —

	CRPS	LogS	DSS
MVAR$(2;3,2,1)$	0.004895	-3.300112	-8.410847
DCC-GARCH(1,1)	0.005048	-3.296478	-8.430833
VAR(3)	0.005123	-2.930259	-7.698397

Table 5.1: Average scores for one step density forecasts.

	CRPS	LogS	DSS
MVAR$(2;3,2,1)$	0.004805	-3.310742	-8.439584
DCC-GARCH(1,1)	0.004845	-3.326115	-8.490107
VAR(3)	0.005022	-3.024682	-7.887240

Table 5.2: Average scores for two step density forecasts.

From this comparison, it appears that the only significant differences between MVAR and DCC-GARCH, in terms of forecast accuracy, are in the CRPS for the one step predictor (first column of Table 5.1), in which on average MVAR outperforms DCC-GARCH, and for DSS in the two step predictor (last column of Table 5.2), where DCC-GARCH outperforms MVAR instead. However, neither method is objectively better than the other. On the other hand, we notice that the VAR model is far behind in terms of forecasting accuracy, and therefore may not suitable for predicting portfolio returns. We conclude that our method for portfolio optimisation with MVAR

models may be a valid alternative to a widely accepted method such as DCC-GARCH, while it clearly outperforms the standard VAR model.

5.6 Discussion

We presented an innovative way of using mixture vector autoregressive models for portfolio optimisation. The method consists in deriving analytically predictive distributions of future observations, and use the conditional covariance matrix, together with modern portfolio theory, to build an efficient portfolio and obtain a distribution for future returns. We have seen in fact that, assuming multivariate normal distributions for mixture components, the conditional predictive distribution of the portfolio return at a future horizon h itself follows a (univariate) mixture of g^h normal components, depending on observation up to the present.

The methodology was tested both on a simulated and a real dataset. For the latter, we compared performance of MVAR with the widely used dynamic conditional correlation model, which uses multivariate GARCH to estimate conditional correlations, and with the VAR model, using a rolling-window forecasting scheme. In particular, forecasting accuracy was assessed using three strictly proper scoring rules, averaged over the number of forecasts. In terms of minimum variance portfolios, the conclusion was that the MVAR and DCC-GARCH have similar performance on the analysed datasets, suggesting MVAR may be considered a valid alternative to DCC-GARCH. Furthermore, it was seen that MVAR outperformed VAR.

Chapter 6

Constrained mixture autroegressive model for uncorrelated time series

We have so far presented applications of mixture autoregressive models to financial returns, highlighting some of the features of this kind of data, such as heteroskedasticity, heavy tails or multimodality. While MAR models can intrisically account for absence of autocorrelation in the data, the property of financial returns of being uncorrelated or weakly correlated has not yet been taken into account explicitly. A standout example of effectively modelling such characteristics is given by GARCH models, which assumptions are that of 0 mean and absence of autocorrelation. For this reason, we consider in this chapter some simple, yet crucial linear constraints that can be applied to MAR models to account for uncorrelatedness in a time series of interest, with the assumption of 0 mean. We refer to this as an uncorrelated version of MAR models.

The stationary region of the parameters of MAR and MVAR models contains the uncorrelated case as well as the 0 mean case, which allows all of the properties mentioned to be achieved smoothly as part of the estimation process. However, it may be useful to ”force” a MAR model to satisfy these assumptions. In fact, given the nature of the model, which allows for multiple modes, the EM algorithm may fail to estimate

paramters such that this assumption is satisfied, if indeed the likelihood function was multimodal.

Another scenario in which we may wish to impose uncorrelatedness on our model is in time series regression or econometric models for which the assumption of $i.i.d.$ residuals is violated. Suppose in fact that a time series regression model has been fitted, and that residuals for such model are uncorrelated, but violate the homoskedasticity assumption of a linear model. In this case, an additional layer of estimation can be added on the residuals by fitting a MAR model that simultaneously accounts for absence of correlation and heteroskedasticity.

A MAR model can be "forced" to satisfy the assumption of uncorrelatedness and 0 mean by simply imposing certain linear constraints on the autoregressive parameters, which we will derive analytically.

6.1 Constraints for uncorrelated MAR model

From Equation (**??**) we have, for $h > 0$:

$$\begin{aligned} \rho_h &= \sum_{k=1}^{g} \pi_k \sum_{i=1}^{p} \phi_{ki} \rho_{|h-i|} \\ &= \sum_{i=1}^{p} \sum_{k=1}^{g} (\pi_k \phi_{ki}) \rho_{|h-i|} \end{aligned} \tag{6.1}$$

If $\rho_h = 0$ for $h = 1, \ldots, p$, then $\rho_h = 0$ for $h > p$ as well. So, it is sufficient to determine the restrictions for $h = 1, \ldots, p$. If $h \in \{1, \ldots, p\}$ and under the assumption $\rho_h = 0$ for $|h| > 0$, the left hand side of (6.1) is 0, and the only non-zero term on the right hand side is in the sum over i when $i = h$. Taking into account that $\rho_0 = 1$, we obtain the condition for uncorrelatedness:

$$\sum_{k=1}^{g} \pi_k \phi_{kh} = 0 \qquad h = 1, \ldots, p \tag{6.2}$$

This is also a sufficient condition for uncorrelatedness since, if it holds, then (6.1) reduces to $\rho_h = 0$ for $h = 1, \ldots, p$.

Notice that this set of constraints implies that, if a MAR model is of autoregressive order p, at least two of the regimes involved must be autoregressive processes of order p for the model to satisfy the assumption of no correlation.

Finally, a constraint to impose 0 mean on the model may be similarly added. Assuming a stationary MAR model, the expression for the unconditional mean, μ, of the model is

$$\mu = \frac{\sum_{k=1}^{g} \pi_k \phi_{k0}}{1 - \sum_{i=1}^{p} \sum_{i=1}^{p} \pi_k \phi_{ki}} \tag{6.3}$$

If the conditions for uncorrelatedness are satisfied, then the denominator in (6.3) is equal to 1. In general, assuming that the denominator is different from 0, and setting $\mu = 0$, we obtain that the condition for the model to have a 0 mean is analogous to that for the autocorrelations:

$$0 = \sum_{k=1}^{g} \pi_k \phi_{k0} \;\Leftrightarrow\; \phi_{g0} = -\frac{\sum_{k=1}^{g-1} \pi_k \phi_{k0}}{\pi_g} \tag{6.4}$$

Therefore, we have a set of $p+1$ constraints on the model. It is important to stress that the constraints in (6.1) and (6.4) are independent of each other, i.e. we could impose uncorrelatedness without setting any constraints on the shift parameters, and viceversa.

Note that, by imposing 0 mean, the conditional expetation of y_t is also 0 at every time t. In fact, the conditional expectation seen in (**??**) is now:

$$\mathrm{E}\left[y_t \mid \mathcal{F}_{t-1}\right] = \sum_{k=1}^{g} \pi_k \left(\phi_{k0} + \sum_{i=1}^{p} \phi_{ki} y_{t-i} \right) = \sum_{k=1}^{g} \pi_k \phi_{k0} + \sum_{i=1}^{p} \sum_{k=1}^{g} \pi_k \phi_{ki} y_{t-i} = 0$$

Consequently, the conditional variance (**??**) reduces to:

$$\mathrm{Var}\left(y_t \mid \mathcal{F}_{t-1}\right) = \mathrm{E}\left[y_t^2 \mid \mathcal{F}_{t-1}\right] = \sum_{k=1}^{g} \pi_k \sigma_k^2 + \sum_{k=1}^{g} \pi_k \mu_{tk}^2.$$

where $\mu_{tk} = \phi_{k0} + \sum_{i=1}^{p} \phi_{ki} y_{t-i}$.

With these constraints, it appears to be no longer possible to derive explicit formulas for parameter estimates via EM-Algorithm. Numerical optimisation methods to maximise the likelihood functions may be sucsessfully used instead.

It is important to notice that there are no constraints required on the component scale parameters $\sigma_1, \ldots, \sigma_g$. This implies that the capacity of MAR models to account for heteroskedasticity is conserved under a constrained setup.

6.1.1 Constrained MAR vs. GARCH model

We have explained several times how MAR models represent a suitable alternative to GARCH models. In particular, the constrained version of MAR model introduced in this chapter could be the direct "competitor" of GARCH.

Recall the equations that define a $GARCH(p,q)$ model, as defined in (2.51):

$$\begin{aligned} \varepsilon_t &= \sigma_t \eta_t \qquad \eta_t \sim WN(0,1) \\ \sigma_t^2 &= \mathrm{Var}\left(\varepsilon_t \mid \mathcal{F}_{t-1}\right) = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_i \sigma_{t-i}^2 \end{aligned} \tag{6.5}$$

where $\omega, \alpha_i, \beta_j \geq 0$ for all i, j. The first equation controls the conditional distribution of the process, which is introduced explicitly and does not change over time, while the second equation describes the conditional variance structure. η_t are i.i.d. with mean 0 and unit variance (i.e. strong white noise), so it may be assumed to follow any distribution satisfying such conditions (e.g. Gaussian, Student-t, skewed Student-t), a characteristic that gives flexibility to the model. Clearly, the conditional mean

$\mathrm{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0$.

We have discussed how imposing constraints to a MAR model ensures uncorrelatedness, as well as 0 mean. As a result, the process is weak white noise, analogous to ε_t for the GARCH model in (6.5). In addition, the choice of a mixture of distributions provides the flexibility required to handle multimodality, skewness and heavy tails. In addition, and in contrast with GARCH, conditional distributions are not explicitly defined, but rather they arise naturally from parameter estimation, and they depend on the recent past of the process. Furthermore, the conditional variance in (6.1) depends on the past of the process, thus accounting for a dependence structure in the data. This is another similarity to GARCH model. It is now clear how MAR and GARCH may be alternative of each other, in that they are similarly structured.

6.2 Testing constrained vs. unconstrained model

In this section, we discuss a test for constrained MAR model against an unconstrained one. Formally, this test determines whether the autoregressive parameters of one of the mixture components can be expressed as a linear combination of the parameters from the other components. Implicitly, this may be seen as a test for serial uncorrelatedness in a time series (i.e. weak white noise), in that the linear combination of parameters discussed implies absence of autocorrelation in the data.

We can therefore build a likelihood ratio test between the constrained and unconstrained models. The literature for likelihood ratio tests in the context of mixture models (Hope, 1968; Aitkin et al., 1981; McLachlan and Peel, 2000, and references therein) suggests that the test statistic, in general, does not follow a standard χ^2_p distribution, where p is the difference in number of parameters between the full and constrained model. While our case is slightly different, as we are not strictly testing for number of mixture components, the same rule applies.

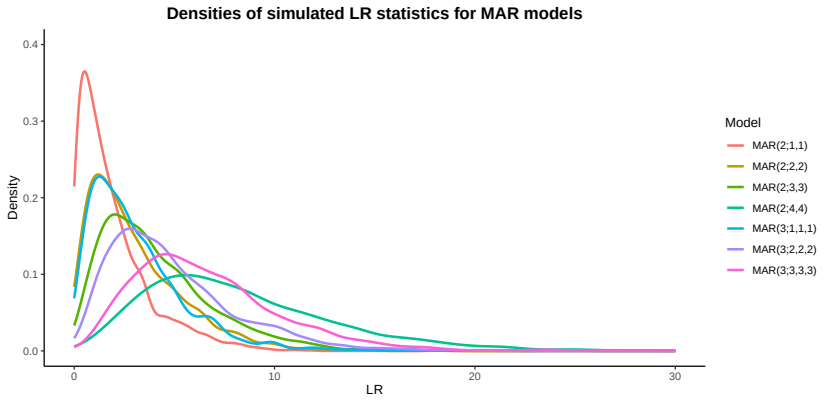

Figure 6.1: Monte Carlo estimated probability density functions of the inspected MAR models.

Every example in the related literature proposes simulation approaches to approximate the distribution of the LR statistic. Two methods are mainly used, namely Monte Carlo and bootstrapping. Due to the underlying dependence of time series data, bootstrapping is not a suitable approach, and therefore we choose to estimate the distribution of the test statistics via Monte Carlo simulations.

We hence proceed with a simulation study. Table 6.1 summarises critical quantiles for all MAR models used (for comparison, or demonstration, although not all displayed), throughout the project. For each model listed, various combinations of parameters were considered, and for each combination 2000 time series of length $n = 500$ were simulated. It was noticed that the distribution of the likelihood ratio statistic is not affected by the choice of model parameters. The shape of the estimated probability density functions for some MAR models is shown in Figure 6.1. Furthermore, it appears that, while the χ^2 and the distribution of the LR statistics are nearly indistinguishable for $g = 2$ mixture components except for extremely large quantiles, the discrepancy increases as g and p increase. Aitkin et al. (1981) suggests that this may be due to the 2 component model being more likely to have one unique maximum, whereas multiple local maxima may appear more often as the number of components increases.

	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
MAR$(2;1,1)$	4.552815	5.889053	8.375820	11.794349
(χ^2_2)	(4.605170)	(5.991465)	(9.210340)	(13.815511)
MAR$(2;2,2)$	6.290129	7.808398	10.72092	14.18025
(χ^2_3)	(6.251389)	(7.814728)	(11.344867)	(16.266236)
MAR$(2;3,3)$	7.906311	9.502881	12.55007	18.57420
(χ^2_4)	(7.779440)	(9.487729)	(13.276704)	(18.466827)
MAR$(2;4,4)$	14.47753	17.23880	22.00793	33.04748
(χ^2_5)	(9.236357)	(11.070498)	(15.086272)	(20.515006)
MAR$(3;1,1,1)$	6.417377	7.555101	12.13223	20.00665
(χ^2_2)	(4.605170)	(5.991465)	(9.210340)	(13.815511)
MAR$(3;2,2,2)$	9.309040	10.909260	15.04179	18.66130
(χ^2_3)	(6.251389)	(7.814728)	(11.344867)	(16.266236)
MAR$(3;3,3,3)$	11.544098	13.418782	17.56541	23.99487
(χ^2_4)	(7.779440)	(9.487729)	(13.276704)	(18.466827)

Table 6.1: Critical quantiles for distribution of likelihood ratio test statistic of some MAR models, each based on 2000 simulated time series of $n = 500$ data points

As usual for likelihood ratio tests, the test is one-sided, meaning that the null hypothesis is rejected if the likelihood ratio between the complete and the constrained model is large. Formally, we test for:

$$H_0 : \mathcal{M}_C \equiv \mathcal{M}_F \quad \text{vs } H_1 : \mathcal{M}_C \neq \mathcal{M}_F \tag{6.6}$$

where $\mathcal{M}_C$ and $\mathcal{M}_F$ denote constrained and unconstrained model respectively. The null hypothesis is rejected if the test statistic T is such that $T > q_{1-\alpha;(g,p)}$. At a closer look, testing constrained versus unconstrained model implicitly corresponds to testing the assumption of weak white noise (i.e. uncorrelatedness) for the time series of interest. This provides an additional support tool to Ljung-Box test, which is predominantly used to test the assumption of strong white noise. Strictly speaking, rejecting the null hypothesis of a Ljung-Box test means that the data are not i.i.d., however it does not necessarily imply correlatedness. Hence, where Ljung-Box test, routinely used to test for independence, rejects the null hypothesis, an additional test for constrained MAR could provide additional information on whether the data can be deemed uncorrelated.

We illustrate this in more detail through an example. We consider data From Freddie Mac, an American mortgage loan company. The dataset, which comprises 570 weekly returns from May 2006 to April 2017, is available in the R package **sarima** (Boshnakov and Halliday, 2020), where it was used as example for introducing a test of uncorrelatedness of a time series using a GARCH model, after the null hypothesis of strong white noise had been rejected. This is a particularly interesting example since, after the real estate plummeted with the financial crisis in 2008, the price of this stock fell from \$60 to as little as \$0.5, and decreased by a further 50% in 2010. As a consequence, the vast majority of shares have been owned by the US governement since 2008. In the following years, the company has been target of speculation in the financial market, due to belief that the stocks would be sold back to private properties. This caused the stock to be highly unpredictable, with volatility clusters alternating over time.

The series of first order differences of log-returns and the corresponding autocorrelation function are shown in Figure 6.2. The autocorrelation plot shows several values exceeding the confidence bands for the null hypothesis of the Ljung-Box test, which is in fact rejected at several lags (we attempted lags 5, 10, 15, 20), with p-values < 0.001 in all cases.

We then fitted a constrained MAR$(2;2,2)$ model (found to be the best fit to the data), and use it to test uncorrelatedness against its corresponding unconstrained model. In contrast with the previous result, the test statistic is 1.6148, which is not significant at 10% level (see the second row of Table 6.1 for reference). Therefore, there is evidence in support of the data being weak white noise, but not strong white noise.

The test can hence be used routinely, similarly to the Ljung-Box test, to assess the presence of an underlying dependence structure in a time series.

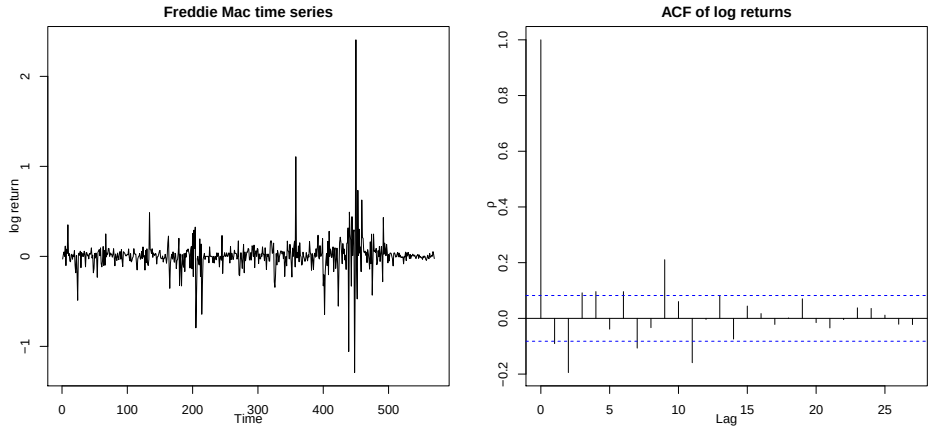

Figure 6.2: Time series of Freddie Mac's log returns and its autocorrelation function. Notice significant correlation at several lags, leading to reject Ljung-Box test.

6.3 Simulation study

Alongside the estimation of quantiles of the distribution of the test statistics, we performe a study aimed to assess what could be the potential benefits in fitting a constrained MAR model in the correct condition. In order to do this, we set up simulations of data from several MAR processes, with $g = 2,3,4$ components and with maximum autoregressive order $p = 4$ (e.g. MAR$(2;3,3)$ or MAR$(3;1,1,1)$). Model parameters are selected in a way such that $\rho_{|h|} = 0$ for $|h| > 0$ and $\mu = 0$ as defined in (6.3). A total of $m = 1000$ datasets of size $n = 500$ were simulated from each of the processes considered in the experiment, for 12 different models within the requirements set described above.

The aim of the experiment is to quantify the improvement in accuracy of the parameter estimates of a constrained MAR model, when the data satisfies the assumptions. In order to do this, for each simulated dataset we estimate the corresponding "true" model, both constrained and unconstrained versions, building in this way Monte Carlo samples of the paramter estimates.

Since the true values of the parameters are known, we then can compare these samples in term of mean squared error (MSE). For a generic parameter θ, recall the

formula for calculation of the mean squared error:

$$MSE\left(\hat{\theta}\right) = \mathrm{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = \frac{\sum_{m=1}^{M}\left(\hat{\theta}_m-\theta\right)^2}{M} \tag{6.7}$$

where $\hat{\theta}_i$ is the parameter estimate from the i^{th} dataset and θ is the true value. The MSE is hence a measure of discrepancy between the parameter estimates and the true value of the parameter, and it quantifies the efficiency of the estimator in question.

The simulation study shows significant improvements in terms of accuracy of the parameter estimates. Overall, mean squared errors reduce by $\sim 10\%$ across all models with $g = 2$ and $\sim 50\%$ across models with $g = 3$ regimes, giving an indication that the more complex the true model is, the more convenient it becomes to use constrained parameter estimation. This is reasonable, since the number of model parameters increases steeply as g increases.

Delving deeper into the simulation results, we notice that the most reductions in MSE are found in the mixing weights. In particular, the average reduction in MSE id around 50% for π_1 in two-component models, and even lower for π_1 and π_2 in three-component models, where the reduction is over 85%. For the remaining parameters, the improvement is still significant, although not on the same level as that for the mixing weights. For two-component models, the improvement swings between $5-10\%$, whereas the reduction in MSE is much more relevant in three-component models, with a decrease between 25% and 50% for most parameters. For what regards the four-component models, result are more constant throughout the model parameters, with reductions in MSE oscillating between $15-25\%$.

Overall, we may conclude that there is evidence that using constrained estimation under correct assumptions brings improvements to parameter estimation in MAR models, suggesting that it is convenient to consider the restricted model for estimation when the data satisfies the assumptions.

6.4 Time Series regression with heteroskedastic errors

A possible application of constrained MAR models, besides that to financial returns, is found in econometrics with time series regression models.

Econometric models consist in the use of statistical methods for a quantitative analysis of economic phenomena. Much like financial data, econometric datasets sometimes show simple linear relationships over time between a variable and lagged values of that variable itself, with the addition of so called "exogenous" variables, other economic phenomena that may influence the response of interest.

Consider a simple linear regression with econometric data:

$$y_t = \beta_0 + \beta_1 y_{t-1} + x_t + \omega_t \tag{6.8}$$

which is essentially an $AR(1)$ model with the addition of an exogenous variable x_t. Unlike linear regression, ω_t in econometric models is not necessarily assumed to follow a Normal distribution. It is often the case in fact that $\omega_t, t = 1, \ldots, n$ are not i.i.d. or even just normally distributed. More often, they are in fact heteroskedastic or heavy-tailed. In such cases a constrained MAR model may be suitable for modelling $\{\omega_t\}$.

We demonstrate this through an example, aimed to show how an uncorrelated MAR model can be used to improve the fit of a previously fit model, whose residuals look far from i.i.d. Normally distributed.

We consider two time series of S&P500 stock indices and the monthly change of the Federal Reserve Board's index of industrial production. Both series are calculated as 100 times the natural logarithm of the index change between a month and the previous month. Before transformation, the two original series consist of monthly observations between January 1946 and June 1993, for a total of 559 observations. The dataset, available in the R package **wooldridge** (Shea, 2018), is analysed by Hamilton

and Lin (1996) who suggest the presence of an intrinsic relationship between stock indices in S&P500 (*pcsp*) and industrial production (*pcip*). The authors also hint to a particular characteristic of this relationship, in that the S&P500 index may foresee the future behavior of industrial production. This means in practice that *pcip* should be influenced by past values of *pcsp*.

Following the authors' analysis and recommendations, the transformed industrial production index *pcip* is treated as the response variable. We then fitted a linear model of the form:

$$pcip_t = \beta_0 + \beta_1 pcsp_{t-2} + \phi_1 pcip_{t-1} + \phi_2 pcip_{t-3} + \phi_{12} pcip_{t-12} + \phi_{24} pcip_{t-24} + \omega_t$$

Here, a seasonal ARIMA model, $SARIMA(3,0,0)(2,0,0)_{12}$ is fitted to $pcip_t$, plus an additional exogenous variable in $pcsp_{t-2}$. The ARIMA term accounts for short term and seasonal autocorrelation in the process, while the exogenous variable is somewhat a confirmation that stock behaviors may in fact foresee future industrial production trends by two months (two time points in practice). The fitted model is hence:

$$\begin{aligned} pcip_t = 3.269 + 0.032 pcsp_{t-2} + 0.329 pcip_{t-1} + 0.099 pcip_{t-3} \\ - 0.175 pcip_{t-12} - 0.125 pcip_{t-24} + \omega_t \end{aligned} \quad (6.9)$$

While the fitted model in itself is interesting, our main focus here is in the residuals $\{\omega_t\}$. In their work Hamilton and Lin (1996) show that the residuals, although uncorrelated, violate the assumptions of a linear model. From diagnostics such as Shapiro-Wilk normality test or Breusch-Pagan test for heteroskedasticity, which both reject their respective null hypothesis with p-values < 0.001, it turns out that the residuals are in fact not i.i.d. Gaussian, nor they have a common variance, but rather are conditionally heteroskedastic. The residuals are shown in the left panel of Figure 6.3

As a solution, Hamilton and Lin (1996) propose to introduce a second step in the estimation, by modelling the residuals, ε_t, using a Markov-switching GARCH model.

In this example, we exploit the same idea of a two-step estimation, but we use instead our uncorrelated mixture autoregressive model with Gaussian components for the residuals, and take the chance to showcase the diagnostic tools for goodness of fit of autoregressive models introduced in Section 2.4.

The time series of residuals can be seen in Figure 6.3. The figure shows that most

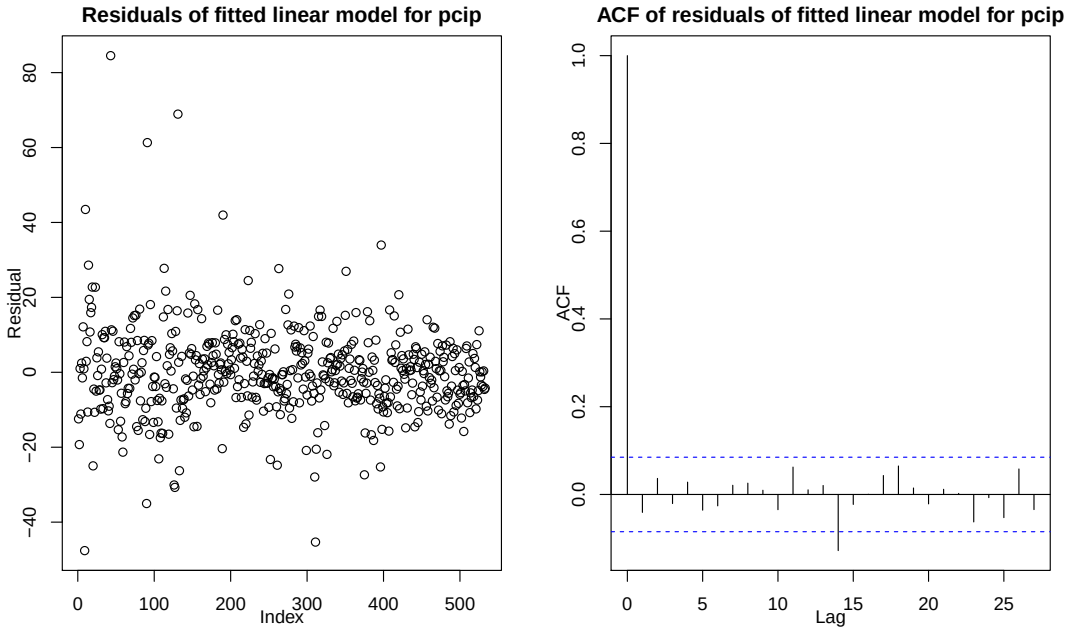

Figure 6.3: Residuals from fitted model in (6.9) and autocorrelation function.

residuals are in the interval $[-20, 20]$, however with several larger values throughout. This behavior suggests a mixture of two components, one to model the data around the 0 mean, the other to account for the sporadic large residuals. After fitting several models, a constrained MAR$(2;2,2)$ was selected as best fit to the data, in terms of diagnostics and BIC. Model diagnostics are displayed in Figures 6.4 and 6.5. Furthermore, the likelihood ratio test statistic for constrained vs. unconstrained model is 3.316. This is not significant at 10% level according to the critical quantiles in Table 6.1, meaning that there is evidence in favor of the constrained model. In other words,

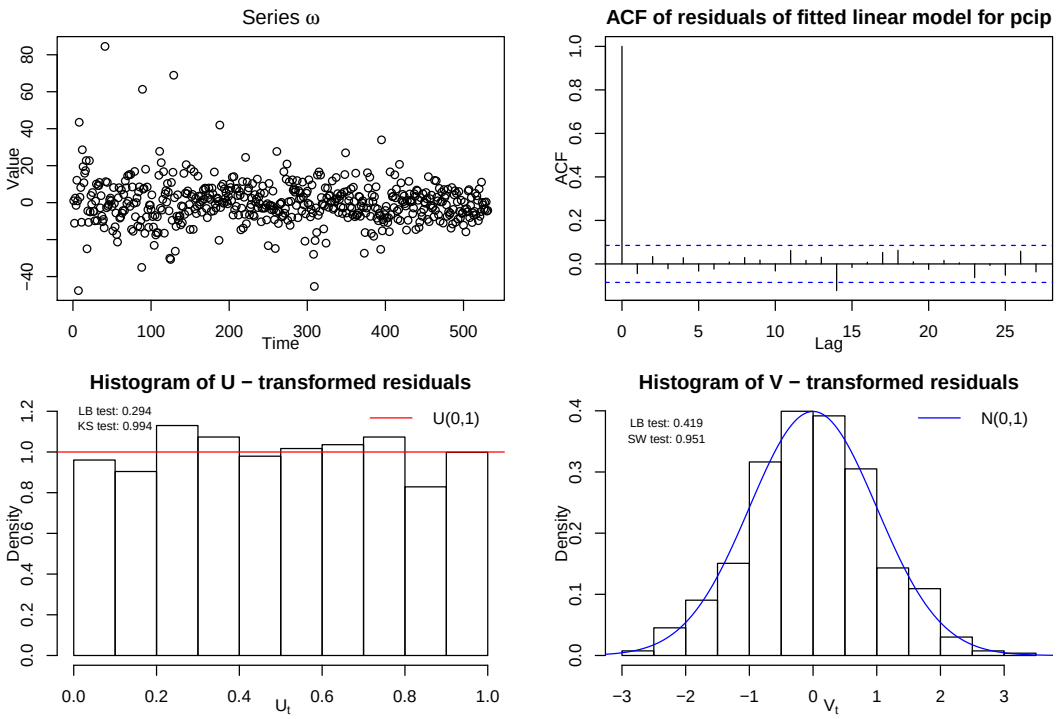

Figure 6.4: Diagnostic plots of fitted MAR$(2;2,2)$ for the series $\{\omega_t\}$.

our fitted uncorrelated MAR model fits the data sufficiently well, so that a MAR model with correlation is not required.

The conditional CDF of the fitted mixture for the residuals can be written in terms of parameter estimates as:

$$F(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0.911\mathbf{\Phi}\left(\frac{\omega_t + 0.354 - 0.014\omega_{t-1} - 0.085\omega_{t-2}}{8.367}\right) + 0.089\mathbf{\Phi}\left(\frac{\omega_t - 3.609 + 0.139\omega_{t-1} + 0.874\omega_{t-2}}{29.953}\right)$$

where autoregressive and shift parameters of the second component are functions of parameters of the first component, as descibed in previous sections.

We start from analysing the set of "traditional" residuals described in section 2.4.

The top two plots in Figure 6.4 show the original series and the autocorrelation function of the residuals. As expected, the absence of correlation, which is a feature

of the original series $\{\omega_t\}$, is maintained in the MAR residuals. The bottom two plots show instead histograms of the two sets of transformed residuals U and V, as defined in (2.36). P-values of relevant tests, specifically Ljung-Box and Kolmogorov-Smirnov for U_t, Ljung-Box and Shapiro-Wilk for V_t, are also printed on the top left of the histograms. We see that the respective null hypothesss for U and V are not rejected, meaning that there is evidence that the model provides a good fit for the data.

Secondly, we derive the set of residuals $\tilde{\varepsilon}_t$, as defined in (2.37). Figure 6.5 shows a histogram of such residuals, as well as the autocorrelation function. It appears from the two plots that the series $\tilde{\varepsilon}_t$ is essentially uncorrelated, and it fits well the standard Normal distribution (blue line), which is what we would expect under correct model specification. Both claims are confirmed once again by Ljung-Box and Shapiro-Wilk tests, which p-values are printed in the histogram plot, both not rejecting the null hypothesis.

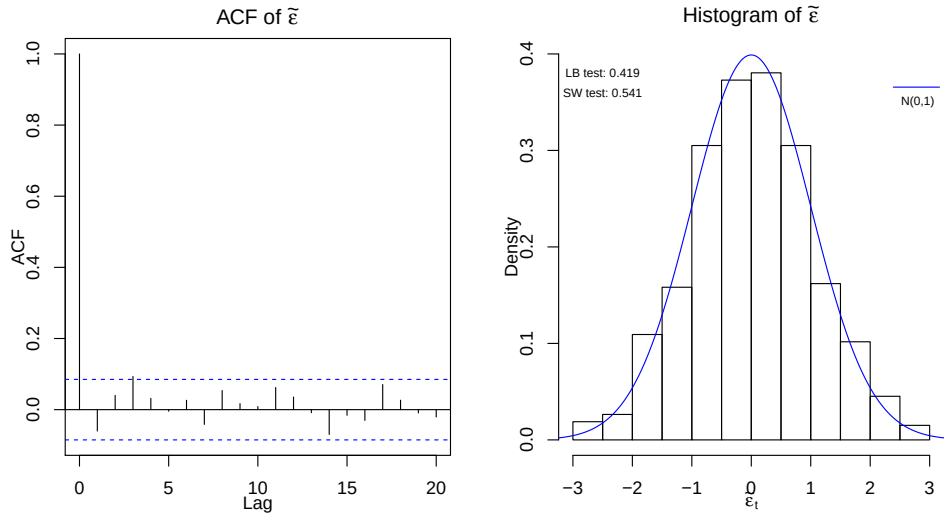

Figure 6.5: Histogram and autocorrelation of residuals $\tilde{\varepsilon}_t$.

It appears in conclusion that the constrained MAR$(2;2,2)$ fits the series of residuals ω_t adequately. In terms of prediction, we know from model assumptions that $\mathrm{E}[\omega_t] = 0$ for all t, meaning that the point predictor $\hat{pcip}_t$ remains invariate. On the other hand,

the variance of such predictor now incorporates the conditional heteroskedasticity and the dependent structure of the data, brought in by the MAR part of the model. For instance, the conditional one step ahead predictor of $pcip_t$ given past information has expected value and variance:

$$\begin{aligned}
\mathrm{E}[pcip_t \mid \mathcal{F}_{t-1}] =& 3.269 \\
&+0.329\,\mathrm{E}[pcip_{t-1} \mid \mathcal{F}_{t-1}] + 0.099\,\mathrm{E}[pcip_{t-3} \mid \mathcal{F}_{t-1}] - 0.175\,\mathrm{E}[pcip_{t-12} \mid \mathcal{F}_{t-1}] \\
&-0.125\,\mathrm{E}[pcip_{t-24} \mid \mathcal{F}_{t-1}] + 0.032\,\mathrm{E}[pcsp_{t-2} \mid \mathcal{F}_{t-1}] + \mathrm{E}[\omega_t \mid \mathcal{F}_{t-1}] \\
=& 3.269 + 0.329 pcip_{t-1} + 0.099 pcip_{t-3} - 0.175 pcip_{t-12} \\
&-0.125 pcip_{t-24} + 0.032 pcsp_{t-2} + 0 \\
\mathrm{Var}(pcip_t \mid \mathcal{F}_{t-1}) =& \mathrm{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \\
&0.902\left[8.307^2 + (-0.321 + 0.034\varepsilon_{t-1} + 0.088\varepsilon_{t-2})^2\right] \\
&+0.088\left[26.546^2 + (2.952 - 0.314\varepsilon_{t-1} - 0.812\varepsilon_{t-2})^2\right]
\end{aligned}$$

Therefore, the variance of the conditional prediction is now time-dependent, through ω_t, on previous values of the production index itself, as well as several past values of the S&P500 stock index.

6.5 Discussion

We introduced an uncorrelated version of MAR models. Formally, the assumption of uncorrelatedness is imposed by a set of linear constraints on the autoregressive parameters of the mixture components, as well as on their shift parameters to impose 0 mean.

We then proposed a likelihood ratio test for testing a constrained model against its corresponding unconstrained model. Implicitly, this test provides information on whether a time series of interest may be considered a weak white noise process, and may be used as an additional tool in support of the traditional Ljung-Box test, when its

null hypothesis is rejected and therefore strong white noise cannot be assumed.

We also showed the advantages of fitting an uncorrelated MAR model when the data appears to satisfy the assumption of uncorrelatedness or weak white noise. In fact, a significant decrease in mean squared error of the parameter estimates was registered.

Finally, we proposed an application of uncorrelated MAR models to error terms from an econometric model. The reason for this application arises from the fact that, in econometrics, the assumption of i.i.d. error terms may be violated, as there is an underlying dependent structure. In such cases, uncorrelated MAR provides an additional layer of modelling, directly applied to the error terms, which aims to better explain their dependent structure.

Chapter 7

Discussion and future work

This thesis proposed and presented advances in theory and applications within the class of mixture autoregressive models.

Chapters 3 and 4 focused on Bayesian analysis of MAR models. The former proposes a way of modelling MAR models with Gaussian components. The existing literature on the topic had in fact some shortcomings in that authors did not consider that the region of first and second order stationarity of the MAR model, what we refer to as *stability* region, does not coincide with that of AR model, and therefore a different approach is required to simulate samples from the posterior distribution of the model parameters. Our proposed method overcame this issue, by use of a Metropolis-Hastings move for the autoregressive parameters. Furthermore, at each iteration of the MCMC, stability of the candidate model is assessed, so that a given set of parameters is automatically rejected if it violates this assumption.

Chapter 4 extended the Bayesian analysis of MAR models to the class of Student-t autoregression. The new assumption on the mixture components was for the innovations to follow Student-t distributions with different degrees of freedom. By using the integral representation of the t distribution, we saw how it is possible to write each component not only in terms of its degrees of freedom, but also mean and variance.

This aspect allows to control all the main features of the distribution.

For the mean and variance of the Student-t distribution to be finite, it is necessary to assume that the degrees of freedom are larger than 2. On the other hand, we discussed how, for degrees of freedom larger than 30, a Gaussian approximation may be preferred due to better numerical stability in MCMC simulations. This also means that a mixture of Gaussian components is a limit case of Student-t mixture. All things considered, we constrained the degrees of freedom to assume values in the interval $[2, 30]$. To take this restriction into account, we introduced a truncated Gamma prior distribution, which also incorporates prior information into the estimation. It was discussed in fact how relevant it is to make use of any prior information for efficient estimation of the posterior distirbution of the degrees of freedom.

Another common issue with Bayesian analysis of mixture models is that of label switching. Label switching arises from the fact that the likelihood function of a mixture model has a number of symmetric modes equal to $g!$, where g is the number of mixture components. In absence of prior information, this may result in mixture components switching permutation, perhaps several times, during MCMC simulations. If not detected, label switching may lead to meaningless, or even wrong inference. Notice however that label switching involves interpretation and identifiability of the model. Prediction of future events using an entire MCMC sample is not affected by label switching, since the density function of MAR models is invariant to permutation of the labels.

While identifiability constraints are commonly used to prevent the occurrence label switching, we opted for a relabelling algorithm a posteriori. From the literature (we refer in particular to Hossain, 2012) it appeared in fact that imposing identifiability constraints on MAR models, such as $\pi_1 > \pi_2 > \cdots > \pi_g$, may sometimes affect convergence of the Markov Chain, so that the stationary distribution cannot be achieved at all (or at least in a feasible amount of simulations). Our relabelling algorithm instead

does not interact with the chain until after the simulation has been completed. The key is in fact that, if label switching has occurred at all in the chain, we are able to detect its presence, at least graphically, and we can consequently apply a relabelling algorithm to adjust labels according to one permutation of our choice. Note that it is not important which permutation is chosen for relabelling, as long as that same one is maintained throughout the relabelling process.

Chapter 5 introduced a novel application of MAR models aimedd to optimisation of portfolios of assets and the associated risk. The methodology combined theory, estimation and prediction with mixture vector autoregressive models (MVAR), the multivariate version of MAR, and modern portfolio theory. MVAR models in fact provide expected value and covariance matrix of a predictor at an arbitrary time horizon, which can be then used to estimate optimal weights to construct a portfolio of assets with given expected return and the risk associated with it. Furthermore, thanks to the properties of MVAR models and of the multivariate Normal distribution, the distribution of the return on a portfolio is fully specified at any time horizon, and therefore we could resort to predictive densities to estimate risk measures such as value at risk and expected shortfall, as well as comparing performance and accuracy of different models in predicting portfolio returns.

Finally, Chapter 6 proposed an uncorrelated version of MAR models. The assumption of uncorrelatedness is satisfied by applying linear constraints on the autoregressive parameters of the model. The main argument behind these constraints is that financial data, and in particular asset returns, are often uncorrelated or weakly correlated, and they are distributed around 0. On the other hand, financial and econometric data are in general heteroskedastic, a feature that standard linear models cannot handle correctly. In addition, by testing the constrained model against its unconstrained counterpart, we also obtained an additional tool for assessing whether a time series could be considered as weak white noise, where the assumption of strong white noise (e.g. by Ljung-Box

test) had previously been rejected.

The forementioned constraints involved the autoregressive parameters of the MAR model, including the shift, which the autocorrelation function of the MAR model depends on. In particular, it turned out that, to ensure that the estimated autocorellation is equal to 0 at all lags, as well as the mean is equal to 0, the autoregressive parameters of one of the mixture components can be written as a linear combination of the autoregressive parameters of the remaining $g-1$ regimes and their mixing weights. At the same time, the heteroskedastic nature of the model is maintained, since the scale parameters of all mixture components are estimated independently.

While the possibility to apply the uncorrelated model to financial returns appeared straightforward, we presented an alternative, more subtle use for it in modelling residuals of a time series regression model in econometrics. We considered a linear relationship (i.e. a linear regression model) between industrial production index and the S&P500 index in the United States. This relationship was highly significant, suggesting that the s&P500 index can foresee future behavior in industrial production. However, the residuals of this model violated the necessary assumptions for a linear model, as there appeared to be an underlying dependence structure. We hence fitted our uncorrelated MAR model to the residuals, and later showed how this additional layer affects prediction, and in particular the uncertainty (variance) around the prediction of future values of the industrial production index.

Future research could extend mixture autoregressive models even further to other distribution assumptions than the ones presented in this thesis. One example of this could be a mixture of Poisson distributions for high frequency count data. In fact, the growing accessibility of nearly instantaneous obserations (such as number of transactions on a particular stock in a minute, numbers of accesses to a website or to a building, among others) may result in clusters that could be modelled effectively by a mixture.

Other possible additions, regarding the Bayesian analysis in particular, could be the implementation of faster Metropolis-Hastings algorithms for simulation of the autoregressive parameters. For instance, one could use the Metropolis Adjusted Langevin Algorithm, or MALA, which makes use of gradient functions to improve the acceptance rate and therefore the efficiency. Fonseca et al. (2008) proposed a way to reduce the level of subjectivity in setting up the prior distribution on the degrees of freedom in a linear regression model with Student-t error terms using Jeffreys priors. The methodology may as well be applied to MAR models, however with the requirement of having to derive objective Jeffreys priors for the specific model.

There are various ways in which the methodology presented in Chapter 5, regarding portfolio optimisation with MVAR models, could be extended. One possibility is to employ the GMVAR model (Kalliovirta et al., 2016) in place of the MVAR model. GMVAR has the useful property that the mixing weights depend on past values of the process. On the other hand, the region for the autoregressive parameters of GMVAR is restricted to a subset of that of MVAR. Also, MVAR and GMVAR have different dynamics and stationary distributions. So the two classes of models complement each other.

Another possible extension of our method is to incorporate factor models. Factor models provide a way of modelling a large number of possibly correlated assets at a time. In addition, distribution assumptions other than normal can be made on the innovation terms. For example, considering a distribution with heavy tails might need a smaller number of components to fit the data. However, estimation could become much more complicated, and numerical algorithms would be required.

The properties of constrained estimation of MAR models are something that requires some more focus and research, given the crucial advantages discussed in chapter 6. For our analysis, we resorted to a numerical optimisation method to obtain parameter estimates, as it appeared that explicit expressions for these are no longer available

due to the nature of the constraints. However, this has not been explored in depth yet, and deserves further study. In addition, while the stability condition remains the same, it may be worth checking whether, with the constraints, it can be rewritten in terms of the $g-1$ unconstrained mixture components with an explicit form. At this stage, we cannot exclude that further constraints could arise from this expression. If that was the case, this might also bring more clarity about the distribution of the likelihood ratio test, which so far has only been derived via simulations.

Bibliography

Aitkin, M., Anderson, D. and Hinde, J.: 1981, Statistical modelling of data on teaching styles, Journal of the Royal Statistical Society. Series A (General) **144**(4), 419–461.
URL: *http://www.jstor.org/stable/2981826*

Akinyemi, M. I.: 2013, Mixture autoregressive models: asymptotic properties and application to financial risk, PhD thesis, Probability and Statistics Group, School of Mathematics, University of Manchester.

Alexander, C.: 2000, A primer on the orthogonal GARCH model, manuscript ISMA Centre, University of Reading, UK **2**.

Barndorff-Nielsen, O. and Schou, G.: 1973, On the parametrization of autoregressive models by partial autocorrelations, Journal of Multivariate Analysis **3**(4), 408 – 419.
URL: *http://www.sciencedirect.com/science/article/pii/0047259X73900304*

Bollerslev, T.: 1986, Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics **31**(3), 307 – 327.
URL: *http://www.sciencedirect.com/science/article/pii/0304407686900631*

Bollerslev, T., Engle, R. F. and Wooldridge, J. M.: 1988, A capital asset pricing model with time-varying covariances, Journal of Political Economy **96**(1), 116–131.
URL: *http://www.jstor.org/stable/1830713*

Boshnakov, G. N.: 2009, Analytic expressions for predictive distributions in mixture autoregressive models., Stat. Probab. Lett. **79**(15), 1704–1709.

Boshnakov, G. N.: 2011, On first and second order stationarity of random coefficient models, Linear Algebra Appl. **434**(2), 415–423.

Boshnakov, G. N. and Halliday, J.: 2020, sarima: Simulation and Prediction with Seasonal ARIMA Models. R package version 0.8.4.
URL: *https://CRAN.R-project.org/package=sarima*

Boshnakov, G. N. and Ravagli, D.: 2020, mixAR: Mixture Autoregressive Models. R package version 0.22.4.
URL: *https://CRAN.R-project.org/package=mixAR*

Box, G. E. P. and Jenkins, G. M.: 1976, Time series analysis : forecasting and control / George E.P. Box and Gwilym M. Jenkins, rev. ed. edn, Holden-Day San Francisco.

Celeux, G.: 2000, Bayesian Inference of Mixture: The Label Switching Problem., Payne R., Green P. (eds) COMPSTAT. Physica, Heidelberg.

Chib, S.: 1995, Marginal likelihood from the Gibbs output., J. A. Stat. Ass. **90**(432), 1313–1321.

Chib, S. and Jeliazkov, I.: 2001, Marginal likelihood from the Metropolis-Hastings output., J. A. Stat. Ass. **96**(453), 270–281.

Dawid, A. P. and Sebastiani, P.: 1999, Coherent dispersion criteria for optimal experimental design, The Annals of Statistics **27**(1), 65–81.
URL: *http://www.jstor.org/stable/120118*

Dempster, A. P., Laird, N. M. and Rubin, D. B.: 1977, Maximum likelihood from incomplete data via the em algorithm, Journal of the royal statistical society. Series B (methodological) pp. 1–38.

Diaconis, P. and Ylvisaker, D.: 1979, Conjugate priors for exponential families, Ann. Statist. **7**(2), 269–281.
URL: *https://doi.org/10.1214/aos/1176344611*

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

Elton, C. and Nicholson, M.: 1942, The ten-year cycle in numbers of the lynx in canada, Journal of Animal Ecology **11**(2), 215–244.
URL: *http://www.jstor.org/stable/1358*

Engle, R.: 2002, Dynamic conditional correlation, Journal of Business & Economic Statistics **20**(3), 339–350.
URL: *https://doi.org/10.1198/073500102288618487*

Engle, R. F.: 1982, Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation, Econometrica **50**(4), 987–1007.
URL: *http://www.jstor.org/stable/1912773*

Engle, R. F. and Kroner, K. F.: 1995, Multivariate simultaneous generalized arch, Econometric Theory **11**(1), 122–150.
URL: *http://www.jstor.org/stable/3532933*

Ferguson, T. S.: 1973, A bayesian analysis of some nonparametric problems, The Annals of Statistics **1**(2), 209–230.
URL: *http://www.jstor.org/stable/2958008*

Fong, P. W., Li, W. K., Yau, C. W. and Wong, C. S.: 2007, On a mixture vector autoregressive model, The Canadian Journal of Statistics / La Revue Canadienne de

Statistique **35**(1), 135–150.
URL: *http://www.jstor.org/stable/20445243*

Fonseca, T. C. O., Ferreira, M. A. R. and Migon, H. S.: 2008, Objective bayesian analysis for the student-t regression model, Biometrika **95**(2), 325–333.
URL: *http://www.jstor.org/stable/20441467*

Geman, S. and Geman, D.: 1984, Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-6**(6), 721–741.

Geweke, J.: 1993, Bayesian treatment of the independent student-t linear model, Journal of Applied Econometrics **8**, S19–S40.
URL: *http://www.jstor.org/stable/2285073*

Geweke, J.: 1994, Priors for macroeconomic time series and their application, Econometric Theory **10**(3-4), 609–632.

Gneiting, T. and Raftery, A. E.: 2007, Strictly proper scoring rules, prediction, and estimation, Journal of the American Statistical Association **102**(477), 359–378.
URL: *https://doi.org/10.1198/016214506000001437*

Goldfeld, S. M. and Quandt, R. E.: 1973, A markov model for switching regressions, Journal of Econometrics **Volume 1, Issue 1**, 3–15.
URL: *https://doi.org/10.1016/0304-4076(73)90002-X*

Good, I. J.: 1952, Rational decisions, Journal of the Royal Statistical Society. Series B (Methodological) **14**(1), 107–114.
URL: *http://www.jstor.org/stable/2984087*

Green, P. J.: 1995, Reversible jump markov chain monte carlo computation and bayesian model determination, Biometrika **82**(4), 711–732.

Hamilton, J. D.: 1989, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica **57**(2), 357–384.
URL: *http://www.jstor.org/stable/1912559*

Hamilton, J. D. and Lin, G.: 1996, Stock market volatility and the business cycle, Journal of Applied Econometrics **11**(5), 573–593.
URL: *http://www.jstor.org/stable/2285217*

Hope, A. C. A.: 1968, A simplified monte carlo significance test procedure, Journal of the Royal Statistical Society. Series B (Methodological) **30**(3), 582–598.
URL: *http://www.jstor.org/stable/2984263*

Hossain, A. S.: 2012, Complete Bayesian analysis of some mixture time series models, PhD thesis, Probability and Statistics Group, School of Mathematics, University of Manchester.

Jones, M. C.: 1987, Randomly choosing parameters from the stationarity and invertibility region of autoregressive-moving average models, Journal of the Royal Statistical Society. Series C (Applied Statistics) **36**(2), 134–138.
URL: *http://www.jstor.org/stable/2347544*

Kalliovirta, L., Meitz, M. and Saikkonen, P.: 2016, Gaussian mixture vector autoregression, Journal of Econometrics **192**(2), 485 – 498. Innovations in Multiple Time Series Analysis.
URL: *http://www.sciencedirect.com/science/article/pii/S030440761630015X*

Lanne, M. and Saikkonen, P.: 2003, Modeling the U.S. Short-Term Interest Rate by Mixture Autoregressive Processes, Journal of Financial Econometrics **1**(1), 96–125.
URL: *https://doi.org/10.1093/jjfinec/nbg004*

Lau, J. W. and So, M. K.: 2008, Bayesian mixture of autoregressive models, Computational Statistics and Data Analysis **53**(1), 38 – 60.
URL: *http://www.sciencedirect.com/science/article/pii/S0167947308002983*

Lawless, J. F. and Fredette, M.: 2005, Frequentist prediction intervals and predictive distributions, Biometrika **92**(3), 529–542.
URL: *http://www.jstor.org/stable/20441212*

Le, N. D., Martin, R. and Raftery, A. E.: 1996, Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models., J. Am. Stat. Assoc. **91**(436), 1504–1515.

Lo, A. Y.: 2005, Weighted chinese restaurant processes, COSMOS **01**(01), 107–111.
URL: *https://doi.org/10.1142/S0219607705000073*

Lütkepohl, H.: 2007, New introduction to multiple time series analysis, Springer, Berlin [u.a.].
URL: *http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHASRT=YOPIKT=1016TRM=ppn+366296310*

Markowitz, H.: 1952, Portfolio selection, The Journal of Finance **7**(1), 77–91.
URL: *http://www.jstor.org/stable/2975974*

McLachlan, G. J. and Peel, D.: 2000, Finite mixture models, Wiley Series in Probability and Statistics, New York.

Nelson, D. B.: 1991, Conditional heteroskedasticity in asset returns: A new approach, Econometrica: Journal of the Econometric Society pp. 347–370.

Ravagli, D. and Boshnakov, G. N.: 2020a, Bayesian analysis of mixture autoregressive models covering the complete parameter space.
URL: *https://arxiv.org/abs/2006.11041*

Ravagli, D. and Boshnakov, G. N.: 2020b, Portfolio optimization with mixture vector autoregressive models.
URL: *https://arxiv.org/abs/2005.13396*

Richardson, S. and Green, P. J.: 1997, On Bayesian Analysis of Mixtures with an Unknown Number of Components., J. R. Stat. Soc., Ser. B, Stat. Methodol. **59**(4), 731–792.

Saikkonen, P.: 2007, Stability of mixtures of vector autoregressions with autoregressive conditional heteroskedasticity, Statistica Sinica **17**(1), 221–239.

Sampietro, S.: 2006, Bayesian analysis of mixture of autoregressive components with an application to financial market volatility, Applied Stochastic Models in Business and Industry **22**(3), 242.

Santos, A. A. and Moura, G. V.: 2014, Dynamic factor multivariate garch model, Computational Statistics & Data Analysis **76**, 606 – 617. CFEnetwork: The Annals of Computational and Financial Econometrics.
URL: *http://www.sciencedirect.com/science/article/pii/S0167947312003398*

Shea, J. M.: 2018, wooldridge: 111 Data Sets from "Introductory Econometrics: A Modern Approach, 6e" by Jeffrey M. Wooldridge. R package version 1.3.1.
URL: *https://CRAN.R-project.org/package=wooldridge*

Smith, J. Q.: 1985, Diagnostic checks of non-standard time series models, Journal of Forecasting **4**(3), 283–291.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980040305*

Tong, H.: 1990, Non-linear time series: a dynamical system approach, Oxford University Press.

Van der Weide, R.: 2002, Go-garch: a multivariate generalized orthogonal garch model, Journal of Applied Econometrics **17**(5), 549–564.

Wong, C., Chan, W. and Kam, P.: 2009, A Student t-mixture autoregressive model with applications to heavy-tailed financial data, Biometrika **96**(3), 751–760.

Wong, C. S.: 1998, Statistical inference for some nonlinear time series models, PhD thesis, University of Hong Kong, Hong Kong.

Wong, C. S. and Li, W. K.: 2000, On a mixture autoregressive model., J. R. Stat. Soc., Ser. B, Stat. Methodol. **62**(1), 95–115.

Wood, S., Rosen, O. and Kohn, R.: 2011, Bayesian mixtures of autoregressive models, Journal of Computational and Graphical Statistics **20**(1), 174–195.
URL: *https://doi.org/10.1198/jcgs.2010.09174*