Predicting the US Unemployment Rate Using Bayesian Model Averaging

Thesis

presented to the Faculty of Economics and Social Sciences
at the University of Fribourg (Switzerland)
in fulfillment of the requirements for the degree of
Doctor of Economics and Social Sciences by

Jeremy Kolly
from Pont-la-Ville (FR)

Accepted by the Faculty of Economics and Social Sciences on September 22, 2014 at the proposal of Prof. Dr. Philippe J. Deschamps (first supervisor) and Prof. Dr. Martin Wallmeier (second supervisor)

Fribourg (Switzerland) 2014

The Faculty of Economics and Social Sciences at the University of Fribourg (Switzerland) does not intend either to approve or disapprove the opinions expressed in a thesis: they must be considered as the author's own (decision of the Faculty Council of January 23, 1990).

To Charlotte

Preface

It was during my Master, when I followed the courses of Prof. Philippe Deschamps, that I started to develop a strong interest for econometrics. His lecture notes were so carefully constructed that it was a pleasure to study them. Later, I was very glad to have the opportunity to begin a thesis under his supervision. The process was very demanding in terms of time and energy, but was highly intellectually rewarding. I would like to thank Prof. Deschamps for the wise guidance he provided me during the realization of my thesis. Furthermore, I would also like to thank Prof. Martin Wallmeier for having accepted to be my second supervisor and for being always available for answering my questions.

Besides my supervisors, I wish to acknowledge Prof. David Ardia for the many insights he brought me. I am eager to work on the research projects we already started together. I also wish to express my gratitude to Prof. Olivier Furrer who kindly welcomed me in his chair after Prof. Deschamps retirement and who provided me with valuable advices. Many thanks also go to Gilles Kaltenrieder for his pedagogical skills, to Christoph Leuenberger and Fred Lang for their mathematical expertise, to my colleagues Jérôme Brugger, Jean-Baptiste Chabrol, Philippe Jolliet, Stéphane Mudhoosoodun, Nicole Schlegel and Josselin Yerly for bringing me a nice working atmosphere, to Stephanie Fürer and Michael Weber for the brainstorming sessions, to Prof. David Stadelmann and Nicolas Guerry for the coffee times and to Helga Kahr for her secretary work. I apologize to anyone I might have forgotten here.

Finally, I warmly thank my parents, Bertrand and Danièle, my sisters, Margaux and Romane, as well as Marie-Jeanne Lang for their constant support. Last but not least, I will never thank enough my beloved partner, Charlotte Lang, for her encouragement, infinite help and great patience.

Jeremy Kolly
April 2014

Contents

Introduction

Motivation

Accurate predictions of future macroeconomic magnitudes are crucial for many economic actors such as central banks, governments, firms and so forth. They often have to take important decisions with regard to such forecasts. At the same time, precise predictions are difficult to obtain because of the complexity of the mechanisms that govern macroeconomic data. Among the macroeconomic variables, the unemployment rate is of particular importance. It reflects the performance of an economy. In this thesis, we will implement econometric models with the aim of improving short-term predictions of the unemployment rate in the US.

In macroeconometrics, we can distinguish between structural and nonstructural models (Diebold, 1998). Structural models describe the data generating process (DGP) with the help of macroeconomic theory and are useful to understand the economic system and to implement policy scenarios. Dynamic stochastic general equilibrium, or DSGE, models constitute for instance a large class of contemporary structural models (see Del Negro and Schorfheide, 2013). Nonstructural models, on the other hand, depict the dependencies present in the DGP with little inputs from macroeconomic theory. Such models are less adequate for policy analysis but are well known to have a better predictive ability than structural models (Del Negro and Schorfheide, 2013, p. 61). The long history of nonstructural modeling is clearly exposed in Diebold (1998).

To achieve the objective mentioned previously, we will propose a non-structural model of the US unemployment rate. In this thesis, in addition to simple autoregressions (ARs), we will consider a class of nonlinear autoregressive models – the smooth transition autoregressions (STARs) – that are able to smoothly switch between regimes. Within this family of nonlinear models, we will focus in particular on the logistic STAR (LSTAR) model. It has been shown by van Dijk et al. (2002) and Deschamps (2008) that the LSTAR model is able to reproduce the nonlinearities in the US unemployment process. The dependent variable in our models will be a logistic transformation of the monthly US unemployment rate. This enables us to account for the fact that the US unemployment rate is a bounded variable. The problems caused by using the untransformed variable will be described in section 1.1.

Numerous studies (Rothman, 1998; Montgomery et al., 1998; Koop and Potter, 1999; Clements and Smith, 2000; van Dijk et al., 2002; Deschamps, 2008) compare the predictive performance of linear and nonlinear time series models for the US unemployment rate. Although nonlinearity is generally favored in these contributions, linear models still appear to be good competitors. Therefore, we suspect that linear and nonlinear models approximate the process generating the US unemployment rate in a complementary way. In order to verify this presumption, we will investigate in this thesis the predictive performance of averages of linear and nonlinear models. None of the above-mentioned studies consider model averaging except Montgomery et al. (1998) and Koop and Potter (1999). But, the approach of the first contribution is clearly different from ours and the forecasting experiment of the second is very limited.

As explained in Geweke and Whiteman (2006), the Bayesian approach is well suited for prediction. The final output in Bayesian forecasting, the predictive density, is valid in small samples and able to coherently incorporate various sources of uncertainty such as uncertainty about future outcomes, about parameters and about models. Furthermore, the use of Markov chain

Monte Carlo (MCMC) methods for posterior simulation enables robust estimation of the LSTAR model (Deschamps, 2008). In this thesis, we will thus opt for a Bayesian approach to estimate our models and to generate predictions.

In order to combine the predictive densities produced by our linear and nonlinear models, we will use the formal Bayesian model averaging (BMA) method. Nevertheless, this technique presumes a complete model space. Since this perpective is not necessarily reasonable, a heuristic model averaging method that does not make such an assumption will also be implemented. This alternative method, called optimal pooling (OP), was introduced recently by Geweke and Amisano (2011, 2012). Besides evaluating the predictive performance of these two methods, it will also be interesting to analyze the sequences of (time-varying) weights they generate.

Organization

This thesis consists of two parts. Part I is concerned with the Bayesian estimation through MCMC methods of AR and LSTAR models of the transformed US unemployment rate. This part contains chapters 1, 2 and 3. In chapter 1, we start by introducing the US unemployment data. We explain how they are produced, present the transformation we apply to them and describe their particular features. Then, we present and justify our modeling strategy and derive the likelihood functions of the AR and LSTAR models. Finally, the foundations of the Bayesian approach are discussed in comparison with the frequentist approach and some Bayesian issues are introduced. Note that Monte Carlo integration is introduced as a tool enabling to solve one of these issues.

In chapter 2, we present the Gibbs sampler and the Metropolis-Hastings algorithm. Then, we consider posterior simulators for the AR and LSTAR models that are based on these MCMC sampling schemes. In this chapter, two complementary model selection criteria are also discussed: the Bayesian information criterion and the marginal likelihood. For the latter, we provide a detailed exposition of the bridge sampling estimator and its numerical standard error. Lastly, we explain the notion of Bayes factor and display the Jeffreys scale.

In chapter 3, we begin by conducting a specification search on the whole data set for the AR and LSTAR models. This search is completed by an investigation of the prior sensitivity of marginal likelihoods. Next, we report MCMC estimates and diagnostics for the best specification in each model class. From the estimation results of the best LSTAR model, we examine the probability of an unit root in each unemployment regime as well as the ability of the estimated transition function to identify those regimes. We conclude the chapter by comparing our results with those of Deschamps (2008).

Part II is concerned with the predictive performance of averages of AR, LSTAR and random walk (RW) models for the transformed US unemployment rate. This part consists of chapters 4 and 5. In chapter 4, we describe the BMA and OP methods. For each of them, we also provide an illustration with simulated data, a discussion of method's properties and a review of the literature. The log scoring rule to assess predictive densities is also introduced in this chapter because it helps the reader to clearly understand the OP method.

In chapter 5, we conduct a pseudo out-of-sample forecasting competition. We first explain how to simulate predictive densities and mixtures of predictive densities. Then, we choose the exact composition of the model averages with specific procedures based on the evolution of posterior model probabilities over time. After having carefully described the forecasting procedure, we study the evolution over time of the weights generated by the BMA and OP methods. We next compare the predictive performance of our models and model averages by means of the Diebold-Mariano test, the efficiency test of West and McCracken and the log score approach. Note that the statistical tests are implemented with two different point predictions and that the evolution over time of cumulative log predictive Bayes factors is also considered in the log score approach. Finally, we address the issue of model misspecification with the probability integral transformation.

Two appendices can be found at the end of this thesis. Their purpose is to complement the analysis of the main text. In appendix A, we give some theorems and their proofs. In appendix B, we provide some empirical results concerning the RW model.

Part I

Formulating and Estimating Models of US Unemployment

Chapter 1

Data, Models and Methods

> "Nevertheless Bayes' theorem has come back from the cemetery to which it has been consigned, and is today an object of research and application among a growing body of statisticians."
>
> Cornfield (1967, p. 41)

We begin this thesis with a description of the data. Section 1.1 provides a lot of information about the US unemployment rate and also shows how we can transform it in order to facilitate the forthcoming empirical analyses. On the basis of data properties and literature, some econometric models are proposed in section 1.2. Finally, section 1.3 gives a comparative introduction to two competing statistical approaches.

1.1 The US Unemployment Data

Unemployment is a fundamental economic problem that concerns all economic actors. Unemployment figures generally receive a wide media coverage. However, the nature of this phenomenon is still not fully understood by economists today. In this thesis, our goal will be to provide statistical insights about the US unemployment rate in order to improve the understanding of this phenomenon.

Each month, the US Census Bureau collects data on the labor market with the current population survey (CPS). These data are then analyzed and published by the US Bureau of Labor Statistics. The CPS sample is made of approximately 60000 households (about 110000 individuals) and provides a reliable estimate of monthly unemployment as argued by the US Bureau of Labor Statistics (2009, p. 3). In the CPS, individuals are categorized as employed, unemployed or not in the labor force. Those classified as unemployed are jobless persons available for work that are looking for a job. Those waiting to be recalled to a job from which they have been laid off also belong to this category. The sum of the employed and unemployed forms the labor force and the third category is for persons that are not in this sum. From the CPS, the US Bureau of Labor Statistics (2013, ch. 1) computes, among others, the unemployment rate which is defined as the proportion of unemployed in the labor force. This statistic does not consider the persons that are not in the labor force.

The data used in this thesis are seasonally adjusted monthly unemployment rates of the US for civilians of 20 years and over. The series goes from 1:1948 to 3:2011 (759 observations) and is available on the Bureau of Labor Statistics' website: `http://www.bls.gov`. It is measured in percentage points and denoted by u_t. By using a seasonal adjustment, it will be easier to identify patterns in the data. For instance, a regime-switching model for the seasonally adjusted data will not switch between regimes because of seasonal fluctuations but rather when the economic conditions are changing.

Being a proportion, the unemployment rate is bounded between 0 and 1. As noted by Koop and Potter (1999, p. 300), working with such a variable is problematic. The support of predictive densities may not be restricted to the $[0, 1]$ interval.[1] Moreover, Deschamps (2008, p. 436) points out that for

[1] In Bayesian time series analysis, a predictive density is defined as the density of future outcomes of a time series given its past realizations.

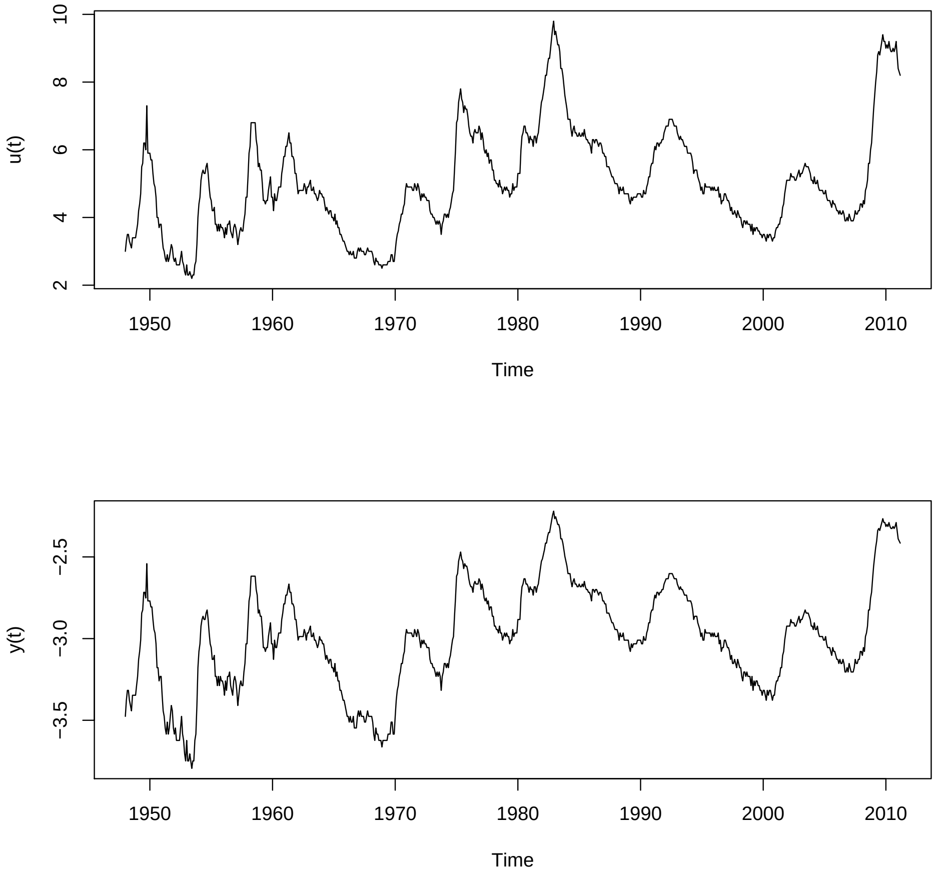


Figure 1.1: Untransformed and transformed US unemployment series

models estimated from u_t the distribution of residuals is strongly leptokurtic (i.e. heavy-tailed). These difficulties can fortunately be overcome by using a contemporary transformation of u_t. A candidate transformation is the logarithmic transformation, $\ln(0.01u_t)$. However, the inverse transformation allows $0.01u_t$ to be greater than 1. As mentioned by Cox and Hinkley (1974, p. 6) "[...] even though this limiting behavior may be far from the region directly covered by the data, it will often be wise to use a family of models consistent with the limiting behavior [...]". Wallis (1987) proposes to use another transformation that is appropriate for the unemployment rate: the

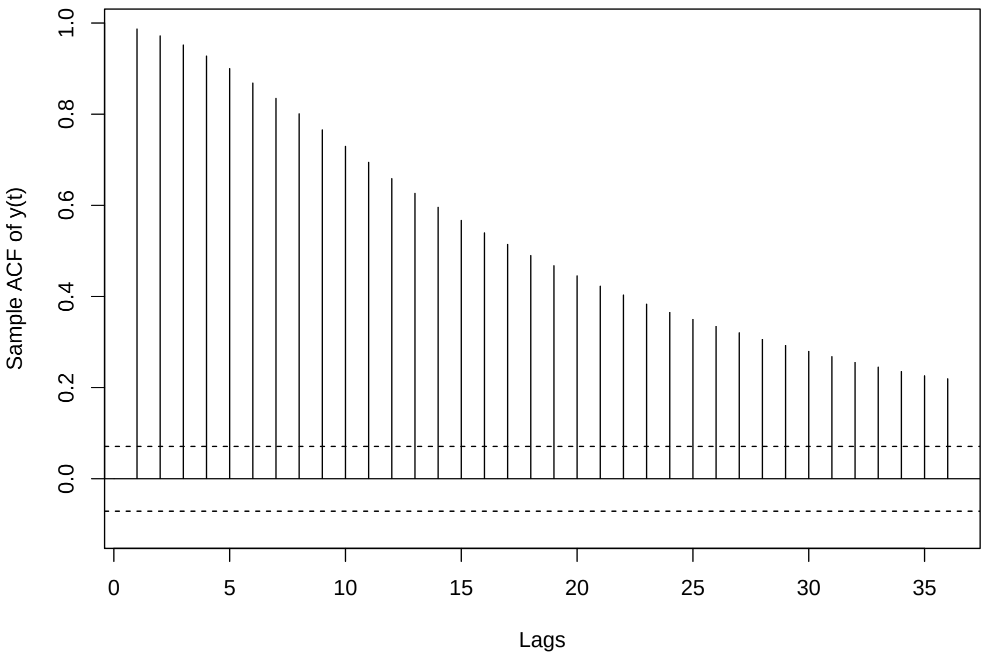


Figure 1.2: Sample ACF of y_t with a 95% confidence interval around zero

logistic transformation. We follow this recommendation and use a logistic transformation of the unemployment rate as our dependent variable:

$$y_t = \ln\left(\frac{0.01u_t}{1 - 0.01u_t}\right). \tag{1.1}$$

The transformed series is unbounded and the inverse transformation will give us predictions between 0 and 1. Figure 1.1 presents the series u_t and y_t.

We can observe in figure 1.1 that the US unemployment rate is characterized by fast increases during economic contractions and by slower decreases during economic expansions. Rothman (1998, p. 164), among others, pointed out that such asymmetric behavior is a nonlinear phenomenon which may not be accurately represented by a linear time series model. Figure 1.2 shows the sample autocorrelation function (ACF) of y_t.[2] We see that the sample autocorrelations decrease slowly indicating a highly persistent process. We can rely on economic theory to explain the cyclical asymmetries and the persistence of the US unemployment rate. For that purpose, we begin by presenting the determinants of unemployment.

In an ideal competitive labor market, unemployment is not an issue. The clearing mechanisms will keep the market at equilibrium where there is no involuntary unemployment. We have to deviate from the classical model to obtain an explanation about this phenomenon. Romer (2006, ch. 9) presents three sets of theories that allow to depart from the ideal model of the labor market. The first set of theories contains the efficiency-wage models. In these models, it is beneficial for firms to pay higher wages because it makes workers more loyal and productive even in situation of imperfect monitoring. Unemployment is then due to above-equilibrium wages. The second set of theories concerns the contracting models. These models argue that negociated agreements between workers and firms prevent wage adjustment and imply unemployment. The last set of theories focuses on search models. A key element of these models is that workers and jobs are heterogeneous. Given this, a complex process takes place to match workers and jobs. Unemployment is then the result of this costly and time-consuming process.

All these theories can be used to explain the persistence of the US unemployment rate. An explanation for the cyclical asymmetries is provided by the search models. Indeed, Mortensen and Pissarides (1994) developed a model where the asymmetric behaviors of job creation and job destruction lead to cyclical asymmetries in unemployment. These economic theories allow a better understanding of the causes and properties of unemployment. In this thesis, we will rather use a time series approach to study this phenomenon because time series models are well known to perform well in forecasting exercises. Our modeling strategy will be presented in the next section.

[2] Being very similar, the sample ACF of u_t is not presented.

1.2 Modeling US Unemployment

Let a stochastic process be a time-indexed sequence of random variables. The US unemployment data described in section 1.1 can be considered as a realization of an unknown stochastic process. We call it the data generating process (DGP). Since the DGP is extremely complex, we will use simplified stochastic processes to approximate it. These representations of the DGP are time series models.

An important notion is that of stationarity. It occurs when a stochastic process is "[...] in a particular state of statistical equilibrium." (Box and Jenkins, 1976, p. 26). More formally, the process for y_t is called covariance-stationary if and only if (Hamilton, 1994, p. 45):

$$E(y_t) = \mu \text{ for all } t$$

$$\text{Cov}(y_t, y_{t-j}) = \gamma_j \text{ for all } t, j.$$

For the sake of simplicity, time series models are often assumed to be linear. The pth-order autoregression (AR) is a linear model that is consistent with the sample ACF of y_t in figure 1.2.[3] It is given by:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t \tag{1.2}$$

where the ϵ_t are independent and identically distributed (i.i.d.) $N(0, \sigma^2)$. It can be rewritten as:

$$\phi(L) y_t = \phi_0 + \epsilon_t$$

where $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ is a polynomial in L, the lag operator. The AR(p) is stationary when all the roots of $\phi(L) = 0$ are outside the complex unit circle. Assuming that the autoregression coefficients are known, we can

[3] We also consulted the sample partial ACF of y_t and it seems that moving average errors do not need to be added to the AR(p).

find the mean of a stationary AR(p) as follows:

$$\begin{aligned} E(y_t) &= \phi_0 + \phi_1 E(y_{t-1}) + \ldots + \phi_p E(y_{t-p}) \\ \mu &= \phi_0 + \phi_1 \mu + \ldots + \phi_p \mu \\ \mu &= \phi_0 (1 - \phi_1 - \ldots - \phi_p)^{-1}. \end{aligned}$$

Further results for the AR process are given in Box and Jenkins (1976) or Hamilton (1994).

The likelihood function is a fundamental concept in statistical inference which has a key role in both the Bayesian and frequentist approaches.[4] Let $y = (y_1, \ldots, y_T)'$ be a sample of data modeled by a time series model and θ a vector containing the unknown parameters of the model. The likelihood function $p(y|\theta)$ is the joint density of y given θ evaluated at the observed sample and treated as a function of θ. The importance of this notion comes from the likelihood principle. It states that the likelihood function contains all the data-based evidence about θ (Box and Jenkins, 1976, pp. 208-209).

We now construct the likelihood function of the AR(p). The vector θ corresponds to $(\phi', \sigma^2)'$ where the vector ϕ contains the autoregression coefficients (intercept included). These unknown parameters are random since we follow a Bayesian approach. While conditioning on $y_c = (y_{1-p}, \ldots, y_0)'$,[5] we decompose the likelihood as follows:

$$\begin{aligned} p(y|y_c, \phi, \sigma^2) &= \prod_{t=1}^{T} p(y_t | y_{1-p}, \ldots, y_{t-1}, \phi, \sigma^2) \\ &= \prod_{t=1}^{T} p(y_t | y_{t-p}, \ldots, y_{t-1}, \phi, \sigma^2). \end{aligned}$$

The distribution of $y_t | y_{t-p}, \ldots, y_{t-1}, \phi, \sigma^2$ is normal with mean and variance

[4] These approaches will be presented in section 1.3.

[5] For simplicity, we condition on some initial observations to form a conditional likelihood function. An overview of the different ways to treat initial observations in time series models can be found in Bauwens et al. (1999, pp. 134-135).

given by $\phi_0 + \sum_{j=1}^{p} \phi_j y_{t-j}$ and σ^2, respectively. Letting the conditioning on y_c as implicit, we find that the likelihood of the AR(p) is:

$$p(y|\phi, \sigma^2) = \frac{1}{(2\pi)^{\frac{T}{2}} \sigma^T} \exp \left[-\frac{1}{2\sigma^2} \sum_{t=1}^{T} \left(y_t - \phi_0 - \sum_{j=1}^{p} \phi_j y_{t-j} \right)^2 \right]. \qquad (1.3)$$

This linear model is straightforward to understand and to implement. However, the cyclical asymmetries of the US unemployment suggest nonlinearity (see section 1.1). Thus, in this thesis, the AR model will be considered as a benchmark.

Time series models can deviate from the linearity hypothesis in many ways. The nonlinear models relevant for studying the US unemployment rate are those that can switch between regimes. Some nonlinear regime-switching models will be introduced as special cases of the following general model:

$$y_t = \phi_{10} + \sum_{j=1}^{p} \phi_{1j} y_{t-j} + G(s_t; \vartheta) \left(\phi_{20} + \sum_{j=1}^{r} \phi_{2j} y_{t-j} \right) + \epsilon_t \qquad (1.4)$$

where standard assumptions are made for the errors. In (1.4), the function $G(\bullet)$ manages the transition between two regimes and is called the transition function. This function depends on the observable variable s_t and is parameterized by the elements of the vector ϑ. According to the forms of $G(\bullet)$ and s_t, we are faced with different models. If the transition function is equal to $I_{\{s_t > c\}}$, an indicator function having the value 1 when the event in brackets occurs and 0 otherwise, we obtain the threshold autoregression (TAR) devised by Tong (1978). In the TAR model, s_t is called the threshold variable and c the threshold parameter. By setting $s_t = y_{t-\delta}$ where δ is a delay parameter, we get the self-exciting TAR (SETAR) model. The drawbacks of the TAR model are conceptual and practical. As indicated by Potter (1999, p. 514), the abrupt transition between the two regimes is not realistic and the step transition function makes the estimation more difficult.[6] We can overcome these problems by using a continuous transition function bounded between 0 and 1 that switches smoothly between the two regimes. In this case, the model is a smooth transition autoregression (STAR). Different STAR models are considered in the literature. Teräsvirta (1994) proposes the logistic STAR (LSTAR) which is based on the logistic transition function:

$$G(s_t;\gamma,c)=\frac{1}{1+\exp[-\gamma^2(s_t-c)]}.$$

This function contains the parameters γ and c. As explained in van Dijk et al. (2002, p. 3), γ governs the smoothness of the transition between the two regimes and c is a threshold between them because $G(c;\gamma,c)=0.5$. Moreover, for large values of γ, the logistic transition function is close to $I_{\{s_t>c\}}$, the transition function of the TAR model. The logistic transition function is plotted in panel A of figure 1.3 for various values of γ. Teräsvirta (1994) also proposes the exponential STAR (ESTAR) which relies on the exponential transition function:

$$G(s_t;\gamma,c)=1-\exp[-\gamma^2(s_t-c)^2].$$

Here, γ has the same interpretation as in the logistic transition function, while c is the location of the inner regime because $G(c;\gamma,c)=0$. However, the ESTAR model does not nest a TAR model when γ is large. The exponential transition function is plotted in panel B of figure 1.3 for various values of γ.

The ESTAR is commonly used to model real exchange rates (see among others Taylor et al., 2001 and Sarantis, 1999) whereas the LSTAR is mostly used for economic variables closely related to the business cycle as unemployment (see for example van Dijk et al., 2002 and Deschamps, 2008). Past studies thus point out that the LSTAR model is more appropriate for our data than the ESTAR model.

[6] Bayesian estimations of SETAR models can be found in Geweke and Terui (1993) and Chen and Lee (1995). For the frequentist alternative, see Hansen (1997).

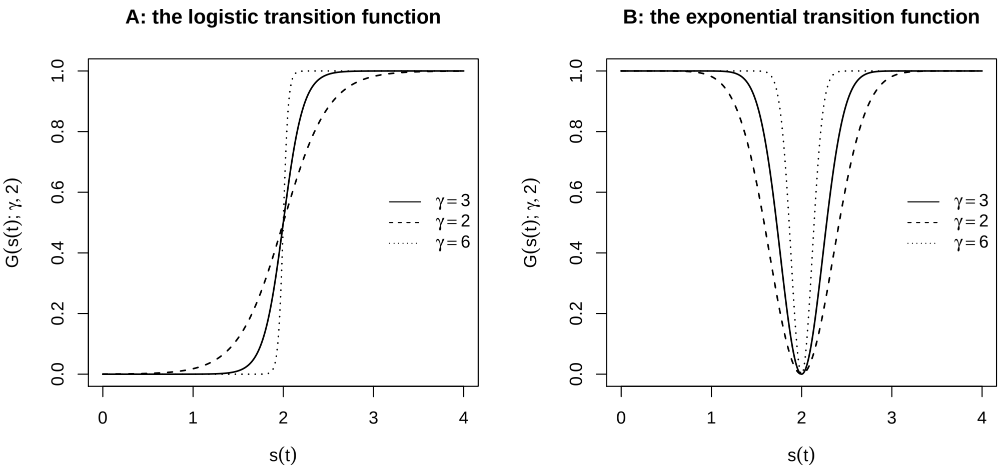


Figure 1.3: The logistic and exponential transition functions

The LSTAR model is also a valuable alternative to nonlinear regime-switching models where s_t is unobservable. Indeed, Deschamps (2008) finds good results in favor of the LSTAR when he compares it to the Markov switching autoregression (MSAR) in a forecasting experiment on US unemployment data.[7]

Among the nonlinear regime-switching models, the LSTAR appears to be a good choice for modeling the transformed US unemployment rate. The LSTAR model we consider in this thesis is written as follows:

$$y_t = \phi_{10} + \sum_{j=1}^{p} \phi_{1j} y_{t-j} + G(s_t; \gamma, c) \left(\phi_{20} + \sum_{j=1}^{p} \phi_{2j} y_{t-j} \right) + \epsilon_t \tag{1.5}$$

$$G(s_t; \gamma, c) = \frac{1}{1 + \exp[-\gamma^2 (s_t - c)]} \tag{1.6}$$

where the ϵ_t are i.i.d. $N(0, \sigma^2)$ and where the autoregressive order is assumed to be the same in both regimes in order to simplify the analysis. The two regimes of this model, occuring when $G(s_t; \gamma, c)$ is close to 0 or to 1, can

[7] The MSAR model results from (1.4) when the transition function is equal to s_t and is governed by a hidden Markov chain with states 0 and 1 (van Dijk et al., 2002, p. 27).

naturally be linked with economic contraction and expansion. The specific form of the transition variable s_t has to be determined. Deschamps (2008) used the Bayes factors to choose between definitions of s_t. We retain his final choice and specify our transition variable as follows: $s_t = u_{t-1} - u_{t-13}$.

The likelihood of the LSTAR is constructed in the same manner as that of the AR. It is given by:

$$p(y|\phi, \sigma^2, \gamma, c) = \frac{1}{(2\pi)^{\frac{T}{2}} \sigma^T} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=1}^{T} \left[y_t - \phi_{10} \right.\right.$$
$$\left.\left. - \sum_{j=1}^{p} \phi_{1j} y_{t-j} - G(s_t; \gamma, c) \left(\phi_{20} + \sum_{j=1}^{p} \phi_{2j} y_{t-j} \right) \right]^2 \right\}$$

where this time the vector $\phi = (\phi_{10}, \phi_{11}, \ldots, \phi_{1p}, \phi_{20}, \phi_{21}, \ldots, \phi_{2p})'$ contains the autoregression coefficients (intercepts included) of the LSTAR.

1.3 The Bayesian and Frequentist Paradigms

Two complementary approaches coexist in econometrics – the Bayesian and frequentist paradigms. A fundamental difference between these two approaches lies in their interpretations of probability.[8] In frequentist theories, a probability is viewed as the relative frequency of an event when an experiment is repeated an infinite number of times (e.g. a coin tossing experiment). On the other hand, Bayesian theories consider that a probability expresses the degree of belief an individual has in the realization of an event. For example, an economist could assess that it is unlikely that her country will enter into recession next year. More details on the objective and subjective interpretations of probability can be found in Leamer (1978, ch. 2).

These views on probability have an incidence on how inference is con-

[8] Note that in both approaches the mathematical definition of a probability function is the same. It is a set function, written Pr(•), which satisfies some axioms (see Leamer, 1978, p. 23).

ducted in each approach. In frequentist theories, the unknown parameter vector θ is considered as fixed and an estimator $\hat{\theta}$ is built on the basis of the sample y in order to learn about θ. The estimator $\hat{\theta}$ is a random vector whose distribution is obtained through repeated sampling and called the sampling distribution. The quality of $\hat{\theta}$ is investigated by looking at the properties of its sampling distribution. In the frequentist approach, the point of view is ex ante; we stand before the drawing of the sample. As we will see in what follows, the point of view in the Bayesian approach is ex post; inference is performed conditional on the sample.

As explained by Lindley (1975, pp. 106-107), the underlying logic in Bayesian theories is that unknown quantities are regarded as random variables. In this logic, probabilities are interpreted as being subjective degrees of belief. As the elements of θ are unknown, they are thus random variables in the Bayesian approach. Before the sample is observed, we give to θ a prior density, $p(\theta)$. Depending on the needs of the analysis, the prior density can be more or less informative, i.e. more or less peaked. After the sample is observed, we can build a posterior density $p(\theta|y)$ for θ. This density merges prior and data information.

The economist previously mentioned thinks a priori that her country is unlikely to enter recession next year. But, she just reads a new report indicating that the main economic indicators are weakening. Given this information, she thinks a posteriori that a recession has one chance out of two to occur next year. This simple example illustrates the Bayesian updating process, that is the revision of a prior belief on the basis of data in order to form a posterior belief. The Bayes' rule, well presented in Cornfield (1967), is the formalization of this process. It is given by:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \tag{1.7}$$

where the denominator must be different from zero and where $p(y|\theta)$ is the likelihood function introduced in section 1.2. Since the denominator in (1.7) does not depend on θ, we can write:

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

Therefore, it suffices to multiply $p(y|\theta)$ by $p(\theta)$ and to remove the factors not depending on θ to find the kernel of $p(\theta|y)$.[9] Although being mathematically trivial, the Bayes' rule had a long and tumultuous history since its first occurrences in the 18th century. A detailed, but nontechnical, presentation of this story is provided by McGrayne (2011).

In the practice of Bayesian econometrics, we often encounter the three following problems with regard to equation (1.7). First, in many cases we can not sample directly from $p(\theta|y)$ because of its nonstandard form. Simulation methods that solve this problem will be described in sections 2.1 and 2.2. Second, the denominator in (1.7), the marginal likelihood, can generally not be obtained analytically.[10] A simulation-based method that is efficient for estimating this number will be given in section 2.3. Third, analytical results for posterior features such as the mean or variance of a particular parameter in θ are frequently unavailable. However, they can be easily estimated by (Markov chain) Monte Carlo integration (see e.g. Koop, 2003 or Carlin and Louis, 2009, ch. 3) as will be explained below.

In empirical applications, the posterior features of interest can generally be obtained from:

$$E[f(\theta)|y] = \int f(\theta)p(\theta|y)d\theta \tag{1.8}$$

where $f(\theta)$ is a given function of θ. Consider a possibly correlated sample $\theta^{(1)}, \ldots, \theta^{(D)}$ from $p(\theta|y)$. An estimator of (1.8) is given by the sample mean:

$$\bar{f}_D = \frac{1}{D}\sum_{d=1}^{D} f(\theta^{(d)}). \tag{1.9}$$

[9] The notion of kernel of a density is presented in definition 1 of appendix A.

[10] In the Bayesian approach, the marginal likelihood plays a key role in model comparison (see section 2.3) and model averaging (see section 4.1).

Under weak conditions, $\bar{f}_D$ converges to $E[f(\theta)|y]$ as D tends to infinity. Therefore, by drawing sufficiently from $p(\theta|y)$ and by averaging the draws as in (1.9), we can obtain an accurate estimate of $E[f(\theta)|y]$. Note that if the draws come from a Markov chain Monte Carlo algorithm, we will usually discard some initial replications before computing (1.9) in order to remove the effect of starting values.

Chapter 2

Posterior Simulators and Model Selection Criteria

> "[...] Themis, the Greek goddess of justice is usually represented as carrying a pair of scales, these being for weights of evidence on the two sides of an argument."
>
> Good (1985, p. 249)

This chapter focuses on two theoretical issues. The first is the presentation of posterior simulators for the AR and LSTAR models in sections 2.1 and 2.2. These simulation methods will initially be described in general terms before being applied to the models. The second issue is the introduction of two complementary model selection criteria in section 2.3. While allowing model comparison, one of these criteria, the marginal likelihood, also plays a key role in Bayesian model averaging as will be seen in section 4.1.

2.1 The Gibbs Sampler

Suppose we have a model with parameters contained in the vector θ. In order to do posterior inference, we often have to draw from the posterior density $p(\theta|y)$. However, depending on the form of $p(\theta|y)$, it might not be possible to perform direct draws. In this case, another strategy is required, such as Gibbs sampling.

Belonging to the family of Markov chain Monte Carlo (MCMC) algorithms, the Gibbs sampler is a posterior simulator fully described in many textbooks, e.g. Koop (2003, ch. 4) or Carlin and Louis (2009, ch. 3). We begin by partitioning θ in blocks, $\theta = (\theta_1, \ldots, \theta_b)'$. The Gibbs sampler will then iterate on the full conditional posteriors as follows:

1. Set arbitrary starting values for $\theta_2^{(0)}, \ldots, \theta_b^{(0)}$.

2. For $d = 1, \ldots, D$, repeat these steps:

$$\begin{aligned}
&\text{Step 1: Draw } \theta_1^{(d)} \text{ from } p(\theta_1 | y, \theta_2^{(d-1)}, \theta_3^{(d-1)}, \ldots, \theta_b^{(d-1)}).\\
&\text{Step 2: Draw } \theta_2^{(d)} \text{ from } p(\theta_2 | y, \theta_1^{(d)}, \theta_3^{(d-1)}, \ldots, \theta_b^{(d-1)}).\\
&\qquad\qquad \vdots\\
&\text{Step } b\text{: Draw } \theta_b^{(d)} \text{ from } p(\theta_b | y, \theta_1^{(d)}, \theta_2^{(d)}, \ldots, \theta_{b-1}^{(d)}).
\end{aligned}$$

Under weak conditions, this algorithm converges to a sequence of draws from the posterior, $\theta^{(d)}$ for $d = d_0 + 1, \ldots, D$. Of course, the first d_0 replications are discarded to cancel the effect of starting values.

In matrix notation, the AR model of equation (1.2) becomes $y = X\phi + \epsilon$, where the tth row of the $T \times (p+1)$ matrix X is $(1, y_{t-1}, \ldots, y_{t-p})$. For this model, the Gibbs sampler iterates on the full conditional posteriors of $\phi = (\phi_0, \phi_1, \ldots, \phi_p)'$ and σ^2. By assuming an independent normal-inverted gamma prior:

$$\phi \sim N(\phi_a, V_a) \tag{2.1}$$

$$\sigma^2 \sim IG(a, b) \tag{2.2}$$

we can derive full conditional posteriors which are also multivariate normal and inverted gamma (for the proof, see theorem 1 in appendix A):

$$p(\phi|y,\sigma^2) \propto \exp\left[-\frac{1}{2}(\phi-\phi_\star)'V_\star^{-1}(\phi-\phi_\star)\right] \tag{2.3}$$

$$p(\sigma^2|y,\phi) \propto \frac{1}{(\sigma^2)^{a_\star+1}}\exp\left(-\frac{b_\star}{\sigma^2}\right) \tag{2.4}$$

where

$$\phi_\star = V_\star\left(\frac{X'y}{\sigma^2}+V_a^{-1}\phi_a\right) \tag{2.5}$$

$$V_\star = \left(\frac{X'X}{\sigma^2}+V_a^{-1}\right)^{-1} \tag{2.6}$$

$$a_\star = a+\frac{T}{2} \tag{2.7}$$

$$b_\star = b+\frac{(y-X\phi)'(y-X\phi)}{2}. \tag{2.8}$$

In this thesis, the prior hyperparameters of the AR model take the following values: $\phi_a = (0,\ldots,0)'$, $V_a = I_{p+1}$ and $a = b = 10^{-6}$. We chose a relatively noninformative prior in order for data information to be predominant.

2.2 A Metropolis-within-Gibbs Algorithm

Depending on the model and the prior, it may happen that some full conditional posteriors do not have a suitable form for drawing. In such case, the Gibbs sampler can be supplemented by Metropolis-Hastings algorithms in order to simulate these awkward conditionals. The Metropolis-Hastings algorithm also belongs to the MCMC family and is fully described in Koop (2003, ch. 5) or Carlin and Louis (2009, ch. 3). For the general exercise which is to simulate a posterior, it works as follows:

1. Set an arbitrary starting value $\theta^{(0)}$.

2. For $d = 1, \dots, D$, repeat these steps:

   Step 1: Draw a candidate θ^c from a candidate generating density $q(\theta^c|\theta^{(d-1)})$.

   Step 2: Compute the acceptance probability:

$$\alpha(\theta^c, \theta^{(d-1)}) = \min\left[1, \frac{q(\theta^{(d-1)}|\theta^c)}{q(\theta^c|\theta^{(d-1)})}\frac{p(\theta^c|y)}{p(\theta^{(d-1)}|y)}\right].$$

   Step 3: Set $\theta^{(d)} = \theta^c$ with probability $\alpha(\theta^c, \theta^{(d-1)})$ and $\theta^{(d)} = \theta^{(d-1)}$ with probability $1 - \alpha(\theta^c, \theta^{(d-1)})$.

Similarly to the Gibbs sampler, this algorithm converges under weak conditions to a sequence of draws from the posterior, $\theta^{(d)}$ for $d = d_0 + 1, \dots, D$. Here again, the first d_0 draws are discarded.

For the LSTAR model of equations (1.5)-(1.6), Deschamps (2008, pp. 437-440) developed a posterior simulator that draws sequentially from the full conditional posteriors of $\phi = (\phi_{10}, \phi_{11}, \dots, \phi_{1p}, \phi_{20}, \phi_{21}, \dots, \phi_{2p})'$, σ^2 and $\vartheta = (\gamma, c)'$, using the most recently drawn conditioning values. When ϑ is known, the LSTAR can be reduced to the linear model $y = X\phi + \epsilon$. X is now a $T \times (2p + 2)$ matrix whose row t is written:

$$(1, y_{t-1}, \dots, y_{t-p}, G_t, G_t y_{t-1}, \dots, G_t y_{t-p})$$

where $G_t \equiv G(s_t; \gamma, c)$. By assuming the independent priors of (2.1)-(2.2), we thus obtain the same results as in (2.3)-(2.8) for $p(\phi|y, \sigma^2, \vartheta)$ and $p(\sigma^2|y, \phi, \vartheta)$.

Two independent normal priors $N(\gamma_a, \sigma_\gamma^2)$ and $N(c_a, \sigma_c^2)$ are postulated for γ and c. Nevertheless, this causes ϑ to have a nonstandard full conditional posterior whose kernel is given by:

$$k^\star(\vartheta) = \exp\left\{-\frac{(\gamma-\gamma_a)^2}{2\sigma_\gamma^2} - \frac{(c-c_a)^2}{2\sigma_c^2} - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left[y_t - \phi_{10} \right.\right.$$
$$\left.\left. - \sum_{j=1}^{p}\phi_{1j}y_{t-j} - G(s_t;\gamma,c)\left(\phi_{20} + \sum_{j=1}^{p}\phi_{2j}y_{t-j}\right)\right]^2\right\}.$$

Deschamps (2008) proposes to simulate ϑ with a Metropolis-Hastings independence chain.[1] He suggests using a multivariate Student candidate generating density with parameters that come from the following linearization. The first-order Taylor expansion of (1.5)-(1.6) around $(\gamma^\star, c^\star)$ is found and then the terms not depending on γ and c are placed on the left-hand side:

$$y_t^\star = \gamma x_{1t}^\star + c x_{2t}^\star + \upsilon_t \tag{2.9}$$

where the υ_t are i.i.d. $N(0,\sigma^2)$ for $t = 1,\ldots,T$ and where:

$$y_t^\star = y_t - \phi_{10} - \sum_{j=1}^{p}\phi_{1j}y_{t-j} - \left(\phi_{20} + \sum_{j=1}^{p}\phi_{2j}y_{t-j}\right)$$
$$\times\left(G(s_t;\gamma^\star,c^\star) - \left.\frac{\partial G_t}{\partial\gamma}\right|_{\gamma^\star,c^\star}\gamma^\star - \left.\frac{\partial G_t}{\partial c}\right|_{\gamma^\star,c^\star}c^\star\right)$$

$$x_{1t}^\star = \left(\phi_{20} + \sum_{j=1}^{p}\phi_{2j}y_{t-j}\right)\left.\frac{\partial G_t}{\partial\gamma}\right|_{\gamma^\star,c^\star}$$
$$x_{2t}^\star = \left(\phi_{20} + \sum_{j=1}^{p}\phi_{2j}y_{t-j}\right)\left.\frac{\partial G_t}{\partial c}\right|_{\gamma^\star,c^\star}.$$

[1] In the previous general presentation, this would have meant that $q(\theta^c|\theta^{(d-1)}) = q(\theta^c)$.

Given the likelihood of (2.9) and the prior of ϑ, the following Bayesian update equations can be derived:

$$\vartheta^\star = S\left[\frac{X_\star' y_\star}{\sigma^2} + \begin{pmatrix}\sigma_\gamma^2 & 0\\ 0 & \sigma_c^2\end{pmatrix}^{-1}\begin{pmatrix}\gamma_a\\ c_a\end{pmatrix}\right] \tag{2.10}$$

$$S = \left[\frac{X_\star' X_\star}{\sigma^2} + \begin{pmatrix}\sigma_\gamma^2 & 0\\ 0 & \sigma_c^2\end{pmatrix}^{-1}\right]^{-1} \tag{2.11}$$

where the $T \times 2$ matrix $X_\star$ has row t equal to $(x_{1t}^\star, x_{2t}^\star)$ and $y_\star = (y_1^\star, \ldots, y_T^\star)'$. We can find an approximate solution for $\vartheta^\star = (\gamma^\star, c^\star)'$ by iterating on (2.10) and (2.11) using prior expectations as a starting point. The kernel of the candidate generating density is multivariate Student with ν degrees of freedom:

$$k(\vartheta^c) = \left[1 + \frac{(\vartheta^c - \vartheta^\star)' S^{-1} (\vartheta^c - \vartheta^\star)}{\nu}\right]^{-\frac{\nu+2}{2}}. \tag{2.12}$$

A candidate ϑ^c drawn from (2.12) is accepted with probability:

$$\alpha(\vartheta^c, \vartheta^{(d-1)}) = \min\left[1, \frac{k(\vartheta^{(d-1)})}{k(\vartheta^c)}\frac{k^\star(\vartheta^c)}{k^\star(\vartheta^{(d-1)})}\right].$$

If it is not accepted, we retain the most recently drawn vector $\vartheta^{(d-1)}$. Finally, we can choose experimentally the number ν of degrees of freedom in (2.12) in order to have a good acceptance rate. Choosing $\nu = 3$ works well in our computations.

Here, the prior hyperparameters on ϕ and σ^2 take the same values as in the AR.[2] This choice allows us to easily compare the two models. The prior hyperparameters in the transition function are $\gamma_a = 3$, $c_a = 0$ and $\sigma_\gamma^2 = \sigma_c^2 = 0.1$ a choice similar to that in Deschamps (2008). This allows the simulator to converge and is noninformative enough as will be seen later.

[2] Of course, the length of ϕ_a and the order of V_a are now equal to $2p + 2$.

2.3 Two Complementary Criteria

In chapter 3, we will determine the values of the autoregressive order p that yield the best model specifications. To perform this task, two complementary model selection criteria will be used. In this section, we will present them for a given model with likelihood $p(y|\theta)$, prior $p(\theta)$ and posterior $p(\theta|y)$ where the parameter vector θ is of length q.

The Bayesian information criterion (BIC), introduced by Schwarz (1978), can be written:

$$BIC(\hat{\theta}) = 2 \ln p(y|\hat{\theta}) - q \ln T \tag{2.13}$$

where $\hat{\theta}$ is the posterior mean of θ. The first term in (2.13) represents the goodness-of-fit and the second penalizes model complexity. Therefore, the decision rule is to choose the model with the highest BIC. This criterion has the advantage of neglecting the prior.[3] However, the size of differences between BIC values remains difficult to interpret.

The second criterion is the marginal likelihood:

$$p(y) = \int p(y|\theta)p(\theta)d\theta. \tag{2.14}$$

Contrary to the BIC, (2.14) is sensitive to the prior. But, as will be seen later, it allows us to calculate the Bayes factors which can clearly be interpreted. Moreover, the marginal likelihood is also useful for computing the weights used in the Bayesian model averaging (see section 4.1). In many cases, the integral in (2.14) cannot be solved analytically and a simulation method is required to estimate $p(y)$. A numerically efficient method is the bridge sampling of Meng and Wong (1996) well described in Frühwirth-Schnatter (2004). It comes from the following identity:

$$1 = \frac{\int p(\theta|y)\beta(\theta)g(\theta)d\theta}{\int \beta(\theta)g(\theta)p(\theta|y)d\theta}$$

[3] In large samples the prior is dominated by the likelihood so that $\hat{\theta}$ (and thus the BIC) is nearly unaffected by the prior.

which, using Bayes' rule, can be written as:

$$p(y) = \frac{\int [p(y|\theta)p(\theta)\beta(\theta)]g(\theta)d\theta}{\int [\beta(\theta)g(\theta)]p(\theta|y)d\theta} = \frac{E_g[p(y|\theta)p(\theta)\beta(\theta)]}{E_{\theta|y}[\beta(\theta)g(\theta)]} \tag{2.15}$$

where $E_h[f(\theta)]$ is the expectation of $f(\theta)$ with respect to the density $h(\theta)$, $\beta(\theta)$ is a bridge function and $g(\theta)$ is an importance density that provides a simple approximation of $p(\theta|y)$. For the AR and LSTAR models, we follow Deschamps (2008, p. 440) and define $g(\theta)$ as the prior with moments given by the empirical moments of the posterior. Concerning the choice of $\beta(\theta)$, it can lead to different estimators of (2.15). For instance, with $\beta(\theta) = 1/g(\theta)$, we obtain the importance sampling estimator:

$$\hat{p}_{IS}(y) = \frac{1}{N}\sum_{n=1}^{N} \frac{p(y|\theta^{(n)})p(\theta^{(n)})}{g(\theta^{(n)})} \tag{2.16}$$

using i.i.d. draws $\theta^{(n)}$, $n = 1, \ldots, N$ from the importance density. Nevertheless, an optimal choice for the bridge function is available. It was proposed by Meng and Wong (1996) and takes the following form:

$$\beta(\theta) = \frac{1}{Ng(\theta) + Mp(y|\theta)p(\theta)/p(y)}. \tag{2.17}$$

The corresponding estimator is called the bridge sampling estimator $\hat{p}_{BS}(y)$. As (2.17) contains $p(y)$, it has to be computed iteratively as follows:

$$\hat{p}_{BS,t}(y) = \frac{\displaystyle\frac{1}{N}\sum_{n=1}^{N} \frac{p(y|\theta^{(n)})p(\theta^{(n)})}{Ng(\theta^{(n)}) + Mp(y|\theta^{(n)})p(\theta^{(n)})/\hat{p}_{BS,t-1}(y)}}{\displaystyle\frac{1}{M}\sum_{m=1}^{M} \frac{g(\theta^{(m)})}{Ng(\theta^{(m)}) + Mp(y|\theta^{(m)})p(\theta^{(m)})/\hat{p}_{BS,t-1}(y)}} \tag{2.18}$$

where $\theta^{(m)}$, $m = 1, \ldots, M$ are MCMC draws from the posterior and $\theta^{(n)}$, $n = 1, \ldots, N$ are i.i.d. draws from the importance density. We may use (2.16) to obtain a starting value. Convergence of (2.18) is very fast in practice.

Table 2.1: The Jeffreys scale to interpret BFs

$-\log_{10}(BF_{kl})$	Evidence against M_k
> 0.5	substantial
> 1	strong
> 2	decisive

The numerical efficiency of the bridge sampling estimator can be assessed with the numerical standard error (NSE). Frühwirth-Schnatter (2004) presents a method to compute it based on the relative mean squared error:

$$RE^2[\hat{p}(y)] = \frac{E[\hat{p}(y) - p(y)]^2}{p(y)^2}$$

where the uncertainty in $\hat{p}(y)$ comes from the random sequences $\theta^{(1)}, \ldots, \theta^{(N)}$ and $\theta^{(1)}, \ldots, \theta^{(M)}$. From this, she derives an approximate relative mean squared error for the bridge sampling estimator:

$$\hat{RE}^2[\hat{p}_{BS}(y)] = \frac{1}{N}\frac{V_g\left[p(\theta|y)/h_1(\theta)\right]}{E_g^2\left[p(\theta|y)/h_1(\theta)\right]} + \frac{\rho_{h_2}(0)}{M}\frac{V_{\theta|y}\left[h_2(\theta)\right]}{E_{\theta|y}^2\left[h_2(\theta)\right]} \tag{2.19}$$

where $h_1(\theta) = \frac{N}{N+M}g(\theta) + \frac{M}{N+M}p(\theta|y)$, $h_2(\theta) = g(\theta)/h_1(\theta)$, $\rho_{h_2}(0)$ is the normalized spectral density of the process $h_2(\theta^{(m)})$ at the frequency 0 and $V_h[f(\theta)]$ is the variance of $f(\theta)$ with respect to the density $h(\theta)$. Knowing that:

$$E[\ln \hat{p}(y) - \ln p(y)]^2 \approx RE^2[\hat{p}(y)]$$

we only need to take the square root of (2.19) to get the NSE of $\ln \hat{p}_{BS}(y)$.

To compare two models, we can use their posterior odds, i.e. the ratio of their posterior model probabilities:

$$\frac{p(M_k|y)}{p(M_l|y)} = \frac{p(y|M_k)}{p(y|M_l)}\frac{p(M_k)}{p(M_l)}. \tag{2.20}$$

In (2.20), the ratio of marginal likelihoods, $BF_{kl} = p(y|M_k)/p(y|M_l)$, is called the Bayes factor (BF). When $p(M_k) = p(M_l)$ is assumed, the posterior odds are equal to this ratio. A review on BFs can be found in Kass and Raftery (1995) and may be supplemented by Kass (1993). For a pair of models, the BF assesses the evidence provided by the data in favor or against one of the models. To interpret the strength of the evidence, we can use the Jeffreys scale (Jeffreys, 1961, app. B) presented in table 2.1.

Chapter 3

An Application to US Unemployment

> "An algorithm must be seen to be believed, and the best way to learn what an algorithm is all about is to try it."
>
> Knuth (1968, p. 4)

In this chapter, we will put the posterior simulators and model selection criteria developed in chapter 2 into practice in order to estimate our models and to conduct specification searches. More precisely, we will begin in section 3.1 by looking at the entire data set to determine the best specifications for the AR and LSTAR models. It will be interesting to see if the data favor linearity or nonlinearity. Then, in section 3.2 we will present the MCMC estimates of the AR and LSTAR specifications that were found to be the best in section 3.1. Some MCMC diagnostics will also be provided to attest that our algorithms work well.

3.1 Model Selection on the Whole Data Set

We will now conduct a specification search for the AR and LSTAR models. In each of them, the dependent variable is defined by equation (1.1) and goes from 2 : 1949 to 3 : 2011 (746 observations). We begin by estimating the models for $p = 1, \ldots, 8$ with the posterior simulators described in sections 2.1 and 2.2. Then, for each model specification, we compute the BIC and estimate the log marginal likelihood by bridge sampling. Tables 3.1 and 3.2 present the results. The NSE of $\ln \hat{p}_{BS}(y)$ is also reported.[1] Regarding the AR(p) model, the best specification corresponds to $p = 6$ where the two criteria are the highest. On the Jeffreys scale exhibited in table 2.1, the BFs comparing the AR specifications to the AR(6) provide substantial evidence against the AR(4)[2] and at least a strong evidence against the other specifications. Given that the evidence against the AR(4) is only substantial, this specification can also be retained. About the LSTAR(p) model, the preferred specification is given by $p = 4$. The BF comparing the LSTAR(3) to the LSTAR(4) gives only substantial evidence against the LSTAR(3). Therefore, we will also retain this specification. When looking for the best specification in both models, the BIC leads us to select the AR(6) and the log marginal likelihood brings us to the LSTAR(4). So the complementary criteria provide contradictory results when we try to choose between linearity and nonlinearity on the whole data set. Note that we also tried a random walk (RW) model. However, as evidenced in appendix B, it does not perform well compared to the AR and LSTAR models. The BIC and the log marginal likelihood of the RW model are only equal to 2528.7161 and 1251.8091 respectively.

Some will perhaps say that our log marginal likelihoods are subjective because they depend on the prior. We argue that this is not true for the following reasons. First, for each of our models the log marginal likelihoods

[1] In order to validate the method, we took for each model specification the standard deviation of repeated estimations of the log marginal likelihood. The results obtained were always close to the NSE values.

[2] Indeed, $-(1279.1486 - 1281.1934)/\ln(10) = 0.8880$ is between 0.5 and 1.

Table 3.1: The two criteria for the AR(p)

p	BIC	$\ln \hat{p}_{BS}(y)$	NSE
1	2518.3297	1241.0324	0.0005
2	2523.5730	1243.4992	0.0005
3	2575.4753	1269.2870	0.0006
4	2595.2363	1279.1486	0.0007
5	2591.2703	1277.1870	0.0008
6	**2599.3862**	**1281.1934**	0.0008
7	2593.4511	1278.2312	0.0008
8	2590.9926	1276.9725	0.0009

yield almost the same ranking as the BIC which neglects the prior. Second, the doubts about prior sensitivity in the LSTAR are removed by a sensitivity analysis. Indeed, we multiplied by 5 the prior variances of γ and c and computed again the log marginal likelihoods. The results, presented in table 3.3, show that the ranking is approximately the same as for the LSTAR with $\sigma_\gamma^2 = \sigma_c^2 = 0.1$ and that the slightly lower values do not change the contradictory results found when comparing both models. With these two arguments, we can conclude that our log marginal likelihoods reflect data information and are not biased by subjective information. In the next section, we will focus on the estimation of those AR and LSTAR models that were

Table 3.2: The two criteria for the LSTAR(p)

p	BIC	$\ln \hat{p}_{BS}(y)$	NSE
1	2566.5165	1266.2816	0.0047
2	2575.7273	1271.9310	0.0071
3	2595.4520	1282.7061	0.0072
4	**2597.5203**	**1284.8690**	0.0077
5	2584.6487	1279.3009	0.0085
6	2577.6175	1276.7763	0.0092
7	2572.7844	1274.6847	0.0098
8	2576.5560	1276.4542	0.0068

Table 3.3: Log marginal likelihoods of the LSTAR(p) when $\sigma_\gamma^2 = \sigma_c^2 = 0.5$

p	$\ln \hat{p}_{BS}(y)$	NSE
1	1265.8159	0.0056
2	1271.2455	0.0086
3	1281.9794	0.0084
4	**1284.0464**	0.0096
5	1278.5441	0.0104
6	1276.0559	0.0110
7	1274.3493	0.0110
8	1276.2199	0.0079

found to be the best in the present section.

3.2 MCMC Estimation and Diagnostics

MCMC estimates of the AR(6) and LSTAR(4) will now be reported. They will be accompanied by MCMC diagnostics to assess numerical accuracy and convergence. The estimations are carried out exactly as in section 3.1. For each model, 12500 posterior replications are generated and 2500 are immediately discarded to remove the effect of starting values. Posterior results are displayed in tables 3.4 and 3.6 where θ_α is the estimated posterior quantile at probability α and $\hat{\theta}$ is the estimated posterior mean.[3] As proposed in Geweke (1992), the NSE of the mean is $\sqrt{\rho_\theta(0)/10000}$ where $\rho_\theta(0)$ is the spectral density at zero of the sequence formed by the parameter replications. It thus takes the autocorrelation of the chain into account. Despite slightly higher values for γ and c, the NSE values remain low in both tables. Therefore, the estimated posterior means exhibit good numerical accuracy. From Geweke (1992), the relative numerical efficiency (RNE) is the squared ratio of a naive NSE ignoring autocorrelation of the chain to the NSE using spectral density which was introduced earlier. When the RNE is near one,

[3] In this section, we use θ to represent a parameter of the models of interest.

Table 3.4: Posterior results for the AR(6)

θ	$\theta_{0.025}$	$\theta_{0.5}$	$\theta_{0.975}$	$\hat{\theta}$	NSE	RNE	CD
ϕ_0	-0.0854	-0.0571	-0.0279	-0.0569	0.0001	1.0075	-1.7486
ϕ_1	0.9370	1.0102	1.0827	1.0101	0.0003	1.1893	-0.6522
ϕ_2	0.0886	0.1923	0.2942	0.1923	0.0006	0.8319	0.7516
ϕ_3	-0.1713	-0.0677	0.0338	-0.0680	0.0005	1.2190	-1.0409
ϕ_4	-0.1970	-0.0952	0.0065	-0.0950	0.0005	1.0676	0.8909
ϕ_5	-0.0177	0.0810	0.1831	0.0815	0.0006	0.7683	-1.1157
ϕ_6	-0.2120	-0.1396	-0.0696	-0.1401	0.0004	0.9606	0.9282
$\sigma^2 \times 1000$	1.5283	1.6899	1.8709	1.6933	0.0008	1.0939	0.2167

Table 3.5: Estimated posterior correlation matrix of the AR(6)

	ϕ_0	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	σ^2
ϕ_0	1.00	0.12	-0.01	-0.04	-0.04	-0.01	0.15	-0.02
ϕ_1		1.00	-0.72	-0.16	0.05	0.08	0.07	-0.02
ϕ_2			1.00	-0.39	-0.14	-0.01	0.07	0.00
ϕ_3				1.00	-0.39	-0.14	0.03	0.02
ϕ_4					1.00	-0.41	-0.13	0.00
ϕ_5						1.00	-0.71	-0.01
ϕ_6							1.00	-0.01
σ^2								1.00

the autocorrelation is negligible as it is indeed the case in table 3.4. The RNE values of table 3.6 show a little more autocorrelation for some parameters. Nevertheless, it remains acceptable. We arrive at the same conclusions by observing the ACFs of posterior replications of the AR(6) (figure 3.1) and of the LSTAR(4) (figure 3.2). The convergence of MCMC algorithms must also be assessed. To this end, Geweke (1992) proposed a convergence diagnostic (CD) that compares the means computed with the first and the last part of the sequence of replications. It is given by:

$$CD = \frac{\hat{\theta}_A - \hat{\theta}_B}{\sqrt{d_A^{-1}\rho_\theta^A(0) + d_B^{-1}\rho_\theta^B(0)}}$$

and follows asymptotically a $N(0,1)$. The replications of the first 10% and the last 50% of the chain are contained in sets A and B respectively. The cardinals of A and B are respectively denoted by d_A and d_B. The following decision rule can be used for this diagnostic. When the CD value is less than about 1.96 in absolute value for each parameter in the model, we can conclude that the MCMC algorithm converges. For the AR(6), the CD values indicate convergence. For the LSTAR(4), the CD value of γ is slightly lower than -1.96 whereas the other CD values are all between -1.96 and 1.96. Thus, convergence seems also to be reached.

Tables 3.5 and 3.7 report the estimated posterior correlation matrices of

Table 3.6: Posterior results for the LSTAR(4)

θ	$\theta_{0.025}$	$\theta_{0.5}$	$\theta_{0.975}$	$\hat{\theta}$	NSE	RNE	CD
ϕ_{10}	-0.0884	-0.0498	-0.0097	-0.0497	0.0002	0.9003	-0.0182
ϕ_{11}	0.7313	0.8442	0.9494	0.8427	0.0009	0.3863	-1.1856
ϕ_{12}	0.0715	0.2033	0.3383	0.2034	0.0007	0.8317	1.3591
ϕ_{13}	-0.1881	-0.0557	0.0763	-0.0554	0.0006	1.0882	-1.1360
ϕ_{14}	-0.1095	-0.0050	0.1041	-0.0043	0.0006	0.7391	0.9072
ϕ_{20}	-0.1309	-0.0551	0.0189	-0.0552	0.0005	0.6930	0.3168
ϕ_{21}	0.0454	0.2106	0.3881	0.2123	0.0013	0.4465	0.9007
ϕ_{22}	-0.2407	-0.0080	0.2209	-0.0079	0.0012	0.9179	-0.6985
ϕ_{23}	-0.2176	0.0021	0.2241	0.0022	0.0011	1.0547	0.1203
ϕ_{24}	-0.3935	-0.2306	-0.0721	-0.2312	0.0008	1.0003	-0.4114
$\sigma^2 \times 1000$	1.4728	1.6295	1.8044	1.6318	0.0008	1.2068	0.9037
γ	2.4032	2.9395	3.5547	2.9541	0.0047	0.3979	-2.3386
c	-0.0615	0.1741	0.3936	0.1713	0.0030	0.1607	-1.0435

The acceptance rate of the Metropolis-Hastings independence chain used to simulate γ and c is 0.7465.

Table 3.7: Estimated posterior correlation matrix of the LSTAR(4)

	ϕ_{10}	ϕ_{11}	ϕ_{12}	ϕ_{13}	ϕ_{14}	ϕ_{20}	ϕ_{21}	ϕ_{22}	ϕ_{23}	ϕ_{24}	σ^2	γ	c
ϕ_{10}	1.00	0.06	0.03	-0.01	0.04	-0.58	-0.04	-0.01	0.02	-0.05	0.00	-0.07	-0.10
ϕ_{11}		1.00	-0.65	-0.10	-0.08	-0.13	-0.71	0.45	0.04	0.04	-0.01	0.09	0.36
ϕ_{12}			1.00	-0.39	-0.10	0.03	0.47	-0.66	0.27	0.06	0.00	-0.02	-0.14
ϕ_{13}				1.00	-0.64	-0.01	0.06	0.26	-0.66	0.48	0.01	0.03	0.05
ϕ_{14}					1.00	0.03	0.07	0.03	0.43	-0.71	0.00	-0.10	-0.26
ϕ_{20}						1.00	0.29	-0.09	-0.03	0.02	0.00	0.07	-0.31
ϕ_{21}							1.00	-0.73	-0.08	0.11	0.02	-0.07	-0.39
ϕ_{22}								1.00	-0.41	-0.09	-0.01	0.04	0.20
ϕ_{23}									1.00	-0.72	-0.01	-0.02	-0.09
ϕ_{24}										1.00	0.01	0.06	0.20
σ^2											1.00	0.02	-0.02
γ												1.00	0.08
c													1.00

the AR(6) and LSTAR(4) models, respectively. The correlations displayed in these matrices are rather low, indicating that the models do not seem to be over-parameterized. These correlations also suggest that the blocking schemes of the MCMC algorithms of both models are appropriate.

In figures 3.3 and 3.4, we plot posterior densities of interest using the kernel method. For the AR(6), the posterior of the sum of autoregression coefficients indicates that the probability of an unit root is zero but that the process is very persistent. For the LSTAR(4), by comparing the posteriors of the sums of autoregression coefficients in both regimes, we observe that one of the regimes is more persistent than the other. We will see below that this regime corresponds to slow decreases in unemployment.

Figure 3.5 presents the estimated transition function $\hat{G}_t \equiv G(s_t; \hat{\gamma}, \hat{c})$ in several ways. In the top panel, there is a plot of all the pairs $(s_t, \hat{G}_t)$ that can be obtained with the whole data set. It reveals the logistic form of the transition function as well as the smoothness and localisation of the transition between the two regimes. In the middle panel, $\hat{G}_t$ is plotted along with the (transformed) US unemployment series. First, we can observe that $\hat{G}_t$ reacts a little too late to turning points in y_t. Second, the two unemployment regimes are clearly identified by the estimated transition function. Indeed, $\hat{G}_t = 1$ corresponds most of the time to a sharp increase in unemployment and $\hat{G}_t = 0$ to a slower decrease.[4] The ability of the LSTAR to identify unemployment regimes increases our expectations about its predictive power. In the bottom panel, $\hat{G}_t$ is accompanied by, in shaded areas, the US recession periods dated by the National Bureau of Economic Research. We observe that $\hat{G}_t$ goes to 1 at the beginning of recessions and remains at 1 for some time after the end of recessions.

Our application is close to that performed by Deschamps (2008, sec. 4 and 6). However, he uses a shorter sample than us: He does not consider the data prior to 1960 since they can imply the existence of a third regime and

[4] The identification remains however less obvious at the beginning of the series where the two asymmetric regimes of the US unemployment are not well defined.

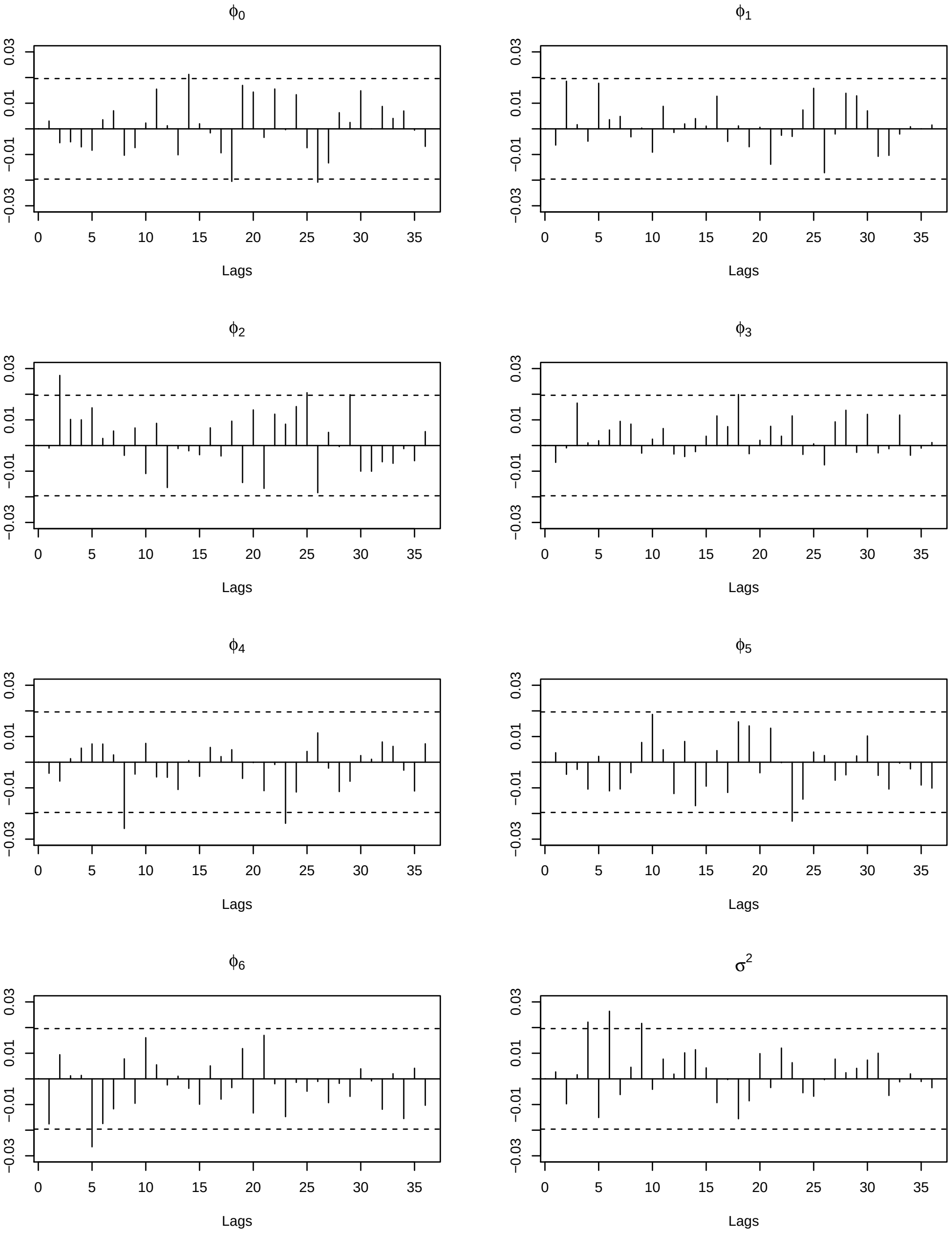


Figure 3.1: ACFs of posterior replications with 95% confidence intervals around zero (AR(6) model)

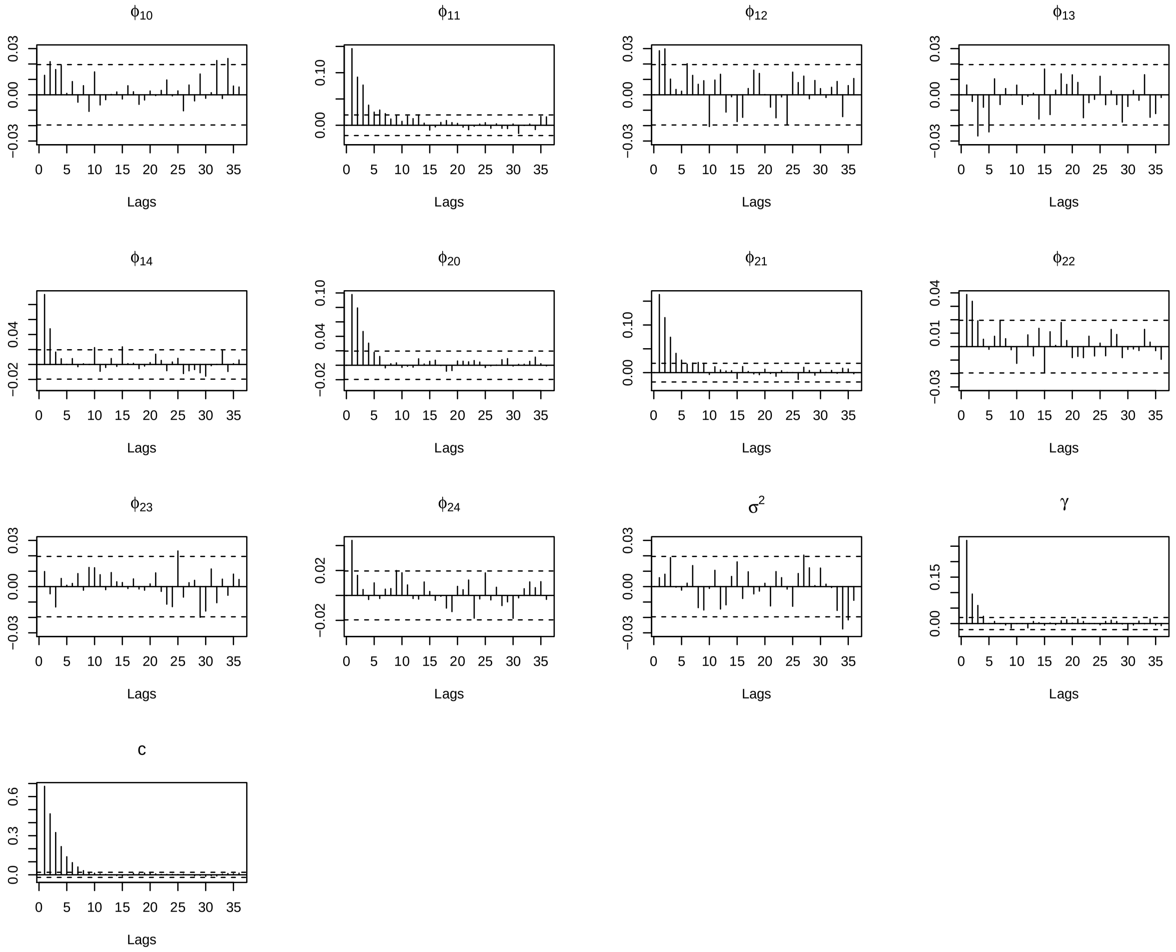


Figure 3.2: ACFs of posterior replications with 95% confidence intervals around zero (LSTAR(4) model)

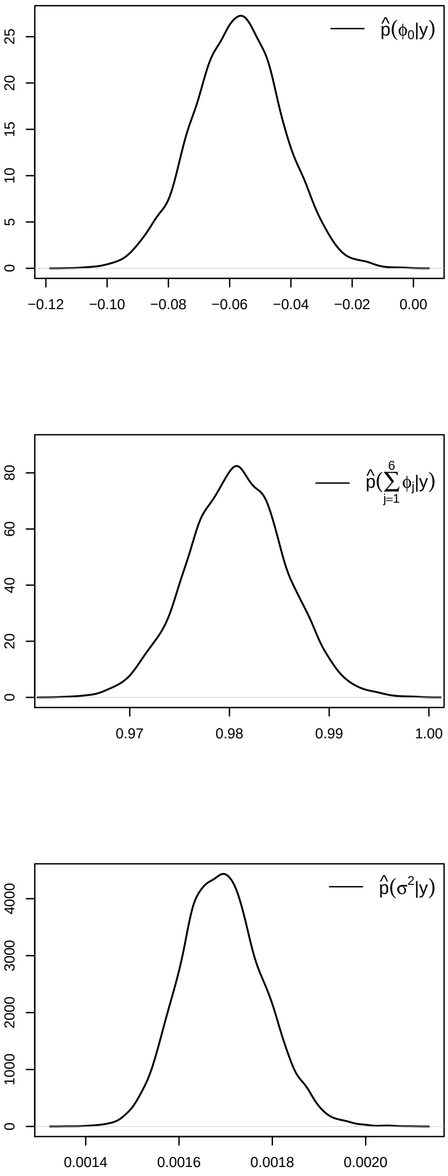


Figure 3.3: Kernel density estimates for the AR(6)

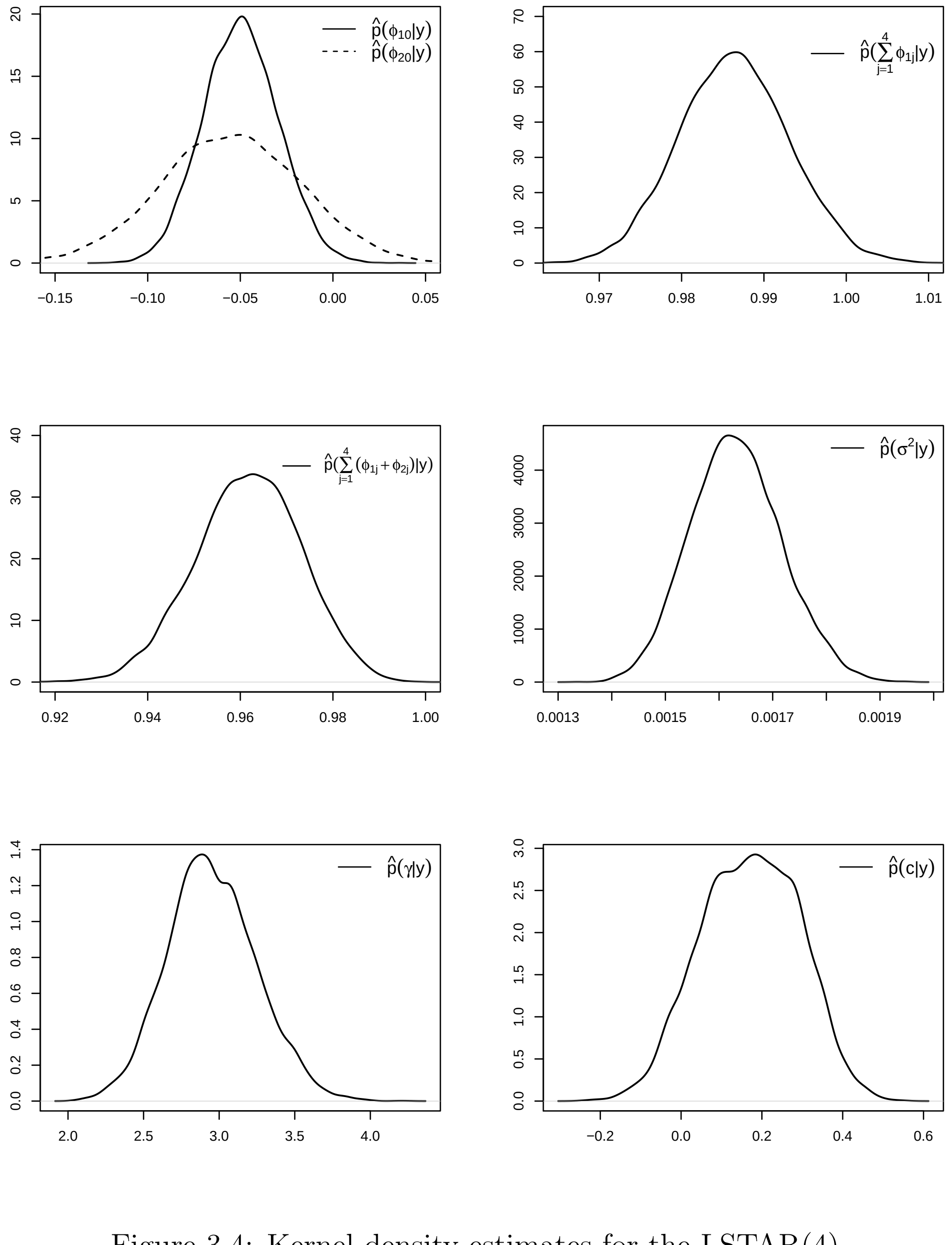


Figure 3.4: Kernel density estimates for the LSTAR(4)

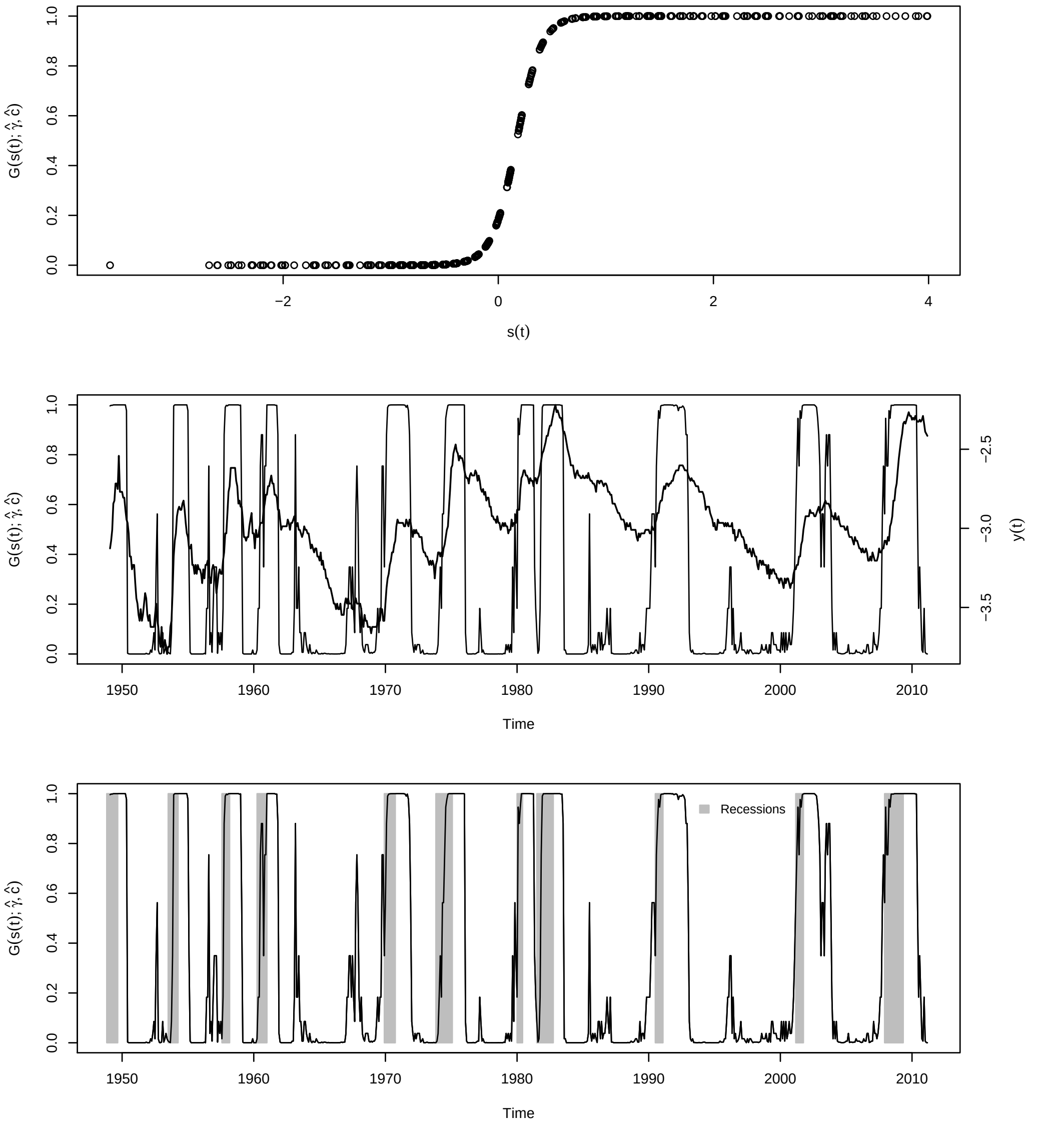


Figure 3.5: The estimated transition function of the LSTAR(4)

his sample ends in 2004. Furthermore, although we chose the same prior distributions as his for the AR and LSTAR models, the parameters of our prior distributions are slightly different.[5] Another difference is that Deschamps (2008) focuses mainly on the comparison of the regime-switching behaviors of the LSTAR and MSAR models. In our study, the MSAR model is not considered and we rather investigate whether the US unemployment DGP is nonlinear or not. Regarding results, the AR and LSTAR selected by Deschamps (2008) are more parsimonious than those of the present application. It is also interesting to point out that he finds very strong evidence against linearity with the help of BFs. In our application, the evidence provided by BFs against linearity is more moderate. We even obtain evidence against nonlinearity with the BIC. Note that Deschamps (2008) does not use this criterion, although it is complementary to BFs as explained in section 2.3.

In conclusion, as we faced model uncertainty on the whole data set, we recommend future research on similar samples to investigate model averaging methods. Moreover, it would be worth to develop a posterior simulator for the LSTAR model that slightly improves the mixing for the parameters of the transition function.

[5] Indeed, we selected a variance of 1 rather than 0.01 for the intercept in the AR model, different variances and covariances for the elements of ϕ in the LSTAR model and a lower value for σ_c^2.

Part II

Formal and Heuristic Model Averaging Methods

Chapter 4

Model Averaging Methods

> "[...] the Chimaera, who was not a human being, but a goddess, for she had the head of a lion and the tail of a serpent, while her body was that of a goat, and she breathed forth flames of fire; but Bellerophon slew her, for he was guided by signs from heaven."
>
> *The Iliad*
> Homer, translated by S. Butler

As stated by Cornfield (1967, p. 34): "A set of observations may be logically consistent with several different hypotheses, even though some of the hypotheses are inherently less plausible than others and even if the observations are more reasonably accounted for by some hypotheses than by others." In such a case, we are confronted with model uncertainty. This phenomenon occurred for instance in section 3.1 where we were not able to discriminate between linear and nonlinear models for the whole data set. Situations in which there is model uncertainty should be treated with care. Indeed, by not considering the ambiguity about models, i.e. by ignoring the somewhat less plausible models, we could be led to underestimate the uncertainty related to quantities of interest (Leamer, 1978, ch. 4; Raftery et al., 1997, among others). A solution to this problem is provided by model averaging methods.

The formal method to account for model uncertainty is Bayesian model averaging (BMA). This technique will be presented in section 4.1. Although this may be unrealistic, BMA assumes a complete model space. Therefore, we will also consider a heuristic model averaging method, named optimal pooling (OP), that does not suppose that the model space is complete. In section 4.2, we introduce the log scoring rule, a key criterion for the OP method. In section 4.3, we describe this method.

4.1 Bayesian Model Averaging

We will start with a bit of history. An early occurrence of BMA can be found in Roberts (1965) where a third party synthesizes expert opinions with the help of BMA. Afterward, Leamer (1978, ch. 4) presented the method as the way to consider formally model uncertainty. Nevertheless, interest for BMA came only later because the method requires a computational power not available at that time. It is only in the 1990s that BMA became practical. The seminal paper of Hoeting et al. (1999) provides a review of the computational advances made during this decade. Subsequently, BMA began to be used in many different disciplines such as meteorology (e.g. Sloughter et al., 2007), astrophysics (Parkinson and Liddle, 2013), macroeconomics (Fernández et al., 2001, among others) and so on.

As can be seen in Koop (2003, ch. 11) or Carlin and Louis (2009, ch. 2), BMA operates as follows. Let M be a discrete random variable taking its values in $\mathcal{M} = \{M_1, \ldots, M_K\}$, the model space under study.[1] Consider that Δ is a quantity of interest having a common meaning in all models and denote the vector of data $(y_1, \ldots, y_T)'$ by $y_{1:T}$. Using the laws of probability,

[1] In section 1.3, we saw that Bayesian theories treat the parameter vector θ as random because it is unknown. The same principle applies to the DGP: Since it is unknown, a probability distribution is assigned over the set of possible models.

we can formulate the posterior density of Δ as follows:

$$p(\Delta|y_{1:T}) = \sum_{k=1}^{K} p(\Delta|y_{1:T}, M_k)p(M_k|y_{1:T}). \tag{4.1}$$

The general BMA equation given in (4.1) is a finite mixture whose components are the posteriors of Δ under the different models in $\mathcal{M}$ and whose weights are the posterior model probabilities (PMPs). In the more specific context of one-step ahead prediction, we replace Δ by the quantity to be forecasted y_{T+1} in (4.1) so as to obtain the predictive density of BMA:

$$p(y_{T+1}|y_{1:T}) = \sum_{k=1}^{K} p(y_{T+1}|y_{1:T}, M_k)p(M_k|y_{1:T}) \tag{4.2}$$

where $p(y_{T+1}|y_{1:T}, M_k)$ is the predictive density of model M_k. With the help of Bayes' rule, the kth PMP can be written as:

$$p(M_k|y_{1:T}) = \frac{p(y_{1:T}|M_k)p(M_k)}{\sum_{l=1}^{K} p(y_{1:T}|M_l)p(M_l)}$$

where $p(y_{1:T}|M_k)$ is the marginal likelihood of model M_k and $p(M_k)$ its prior probability. By setting equal prior weights to the models ($p(M_k) = 1/K$ for all k), the formula simplifies to:

$$p(M_k|y_{1:T}) = \frac{p(y_{1:T}|M_k)}{\sum_{l=1}^{K} p(y_{1:T}|M_l)}. \tag{4.3}$$

This means that we only need to know the marginal likelihoods of the models in $\mathcal{M}$ to compute their PMPs. As discussed in section 2.3, the marginal likelihoods can be efficiently estimated by the bridge sampling method.

We now consider an illustration of the mechanisms of BMA. A given economic time series is driven by the following DGP: $y_t \overset{\text{i.i.d.}}{\sim} N(0.29, 3.86)$. An economist postulates the models M_1: $y_t \overset{\text{i.i.d.}}{\sim} N(0.2, 3.4)$ and M_2: $y_t \overset{\text{i.i.d.}}{\sim} N(0.4, 5)$ for this time series. The top panel of figure 4.1 presents the predic-

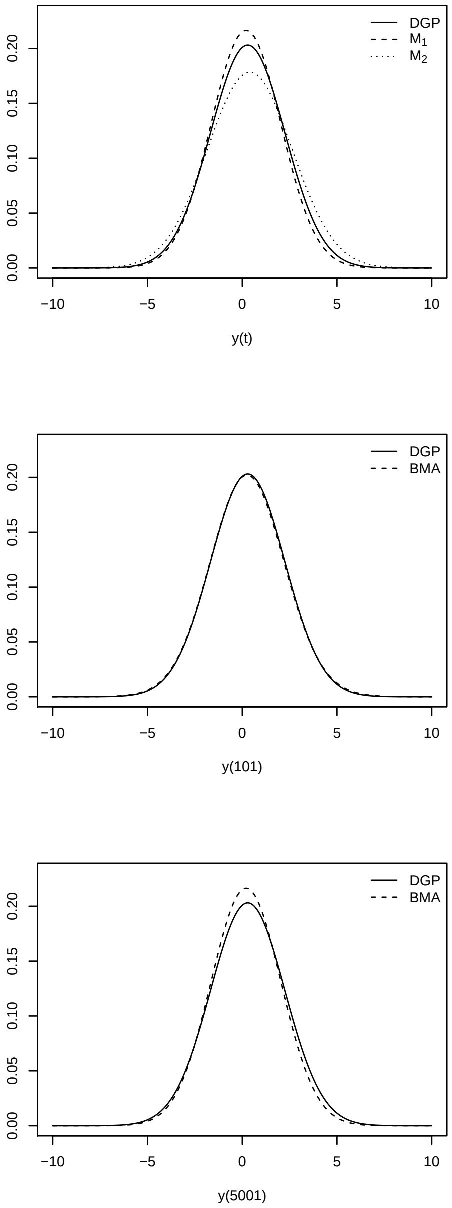


Figure 4.1: Predictive densities of the BMA illustration

tive densities of the DGP, M_1 and M_2. The economist holds a sample of 100 realizations of the series and wants to predict the next outcome with the help of BMA. First, she computes $p(M_1|y_{1:100})$ and $p(M_2|y_{1:100})$ assuming equal prior model probabilities and obtains 0.63 and 0.37, respectively. Then, she derives the one-step ahead predictive density of BMA drawn in the middle panel of figure 4.1. We see that BMA provides a better approximation of the DGP density than M_1 and M_2. Wishing to observe the large sample behavior of BMA, the economist collects 4900 additional realizations of the series. When she computes $p(M_1|y_{1:5000})$ and $p(M_2|y_{1:5000})$, she gets 1 and 0, respectively. This means that the one-step ahead predictive density of BMA is the same as the density of M_1 as can be seen in the bottom panel of figure 4.1. This phenomenon occurs despite the fact that M_1 is not the DGP.[2]

The BMA method has several interesting properties. The Kullback-Leibler information criterion (KLIC) measuring the distance from (4.1) to $p(\Delta|y_{1:T}, M_k)$ is given by:[3]

$$KL[p(\Delta|y_{1:T}), p(\Delta|y_{1:T}, M_k)] = E[\ln p(\Delta|y_{1:T}) - \ln p(\Delta|y_{1:T}, M_k)]$$

where the expectation is with respect to (4.1). Since the KLIC is nonnegative, we have:

$$E[\ln p(\Delta|y_{1:T})] \geq E[\ln p(\Delta|y_{1:T}, M_k)].$$

Raftery et al. (1997, p. 180) interpret this result by saying that on average BMA yields better predictive performance with regard to a log scoring rule than any model in $\mathcal{M}$. Furthermore, BMA is also optimal for forecasting with regard to an expected squared error loss when the set of models is exhaustive (Min and Zellner, 1993).[4] Another interesting property of BMA

[2] We inform the reader who wants to reproduce this example that the PMPs computed with 100 realizations can change considerably from one simulation of the series to another, while the large sample PMPs will always be 1 and 0.

[3] A general definition and some features of the KLIC can be found in Geweke (2005, p. 92).

[4] Note that the out-of-sample predictive performance of BMA has also been empirically demonstrated, for instance by Fernández et al. (2001) in cross-country growth regression

is that prior information on the relative correctness of the models in $\mathcal{M}$ can easily be incorporated in the analysis through the prior distribution of M.

As previously illustrated and as noticed by Diebold (1991), BMA attributes in large samples unit weight to a single model in $\mathcal{M}$ and zero weights to the others. This means that BMA considers that the DGP belongs to $\mathcal{M}$ (Geweke and Amisano, 2011, 2012) or in the terminology of Bernardo and Smith (1994) that it is $\mathcal{M}$-closed. When indeed the DGP belongs to $\mathcal{M}$, BMA will consistently give it unit weight in large samples. However, such situation is rare in economics and occurs mainly in simulation studies where the DGP is chosen by the researcher. Most of the time in the real world, the DGP is not among the models considered by the researcher. In this context, BMA will give in large samples unit weight to the model with the smallest KLIC distance from the DGP (Geweke and Amisano, 2011). Except if this model is a very good approximation of the DGP, this amounts to saying that BMA treats a false model as being true.

Let us now examine the implications of the $\mathcal{M}$-closed assumption of BMA on its forecasting performance. If the DGP belongs to $\mathcal{M}$ or if a model in $\mathcal{M}$ is close to the DGP, the forecasting performance of BMA will be excellent given a large sample. If the DGP is neither in $\mathcal{M}$ nor closely approximated by a model in $\mathcal{M}$, the forecasting performance of BMA will be probably better with a moderate sample possibly allowing some false models to be combined than with a large sample where all weight is attributed to a false model. Indeed, it is well known since Bates and Granger (1969) that model averaging can improve predictive accuracy because the individual models, although misspecified, can independently capture different aspects of the underlying DGP.

In summary, the predictive performance of BMA is difficult to predict in advance as it depends on a wide variety of criteria (sample size, models under study, nature of the DGP and so on) and on their interactions. An empirical

or Hoeting et al. (1999) in medical experiments.

investigation is often necessary to determine the value of BMA in a particular context. In chapter 5, we will thus implement a BMA of some carefully selected specifications of the AR and LSTAR models to see whether it forecasts more accurately the US unemployment rate than individual models or other model averaging methods.

4.2 The Log Scoring Rule for Bayesian Predictions

As explained by Gneiting and Raftery (2007), a scoring rule is a numerical score computed from the predictive density provided by a given model for a future outcome and the realization of this outcome.[5] It assesses the goodness of the predictive density and allows to compare it with those provided by another models for the same outcome. Assume that the researcher reports a predictive density $p^{\star}(y_t|y_{1:t-1}, M_k)$ such as to maximize expected score where expectation is taken with respect to her subjective predictive density $p(y_t|y_{1:t-1}, M_k)$. Although this is desirable from a scientific viewpoint, the researcher is not forced to report her personal prediction. A scoring rule encouraging the researcher to be honest is one for which expected score is maximized when the researcher reports her personal prediction. Such scoring rule is said to be proper. Furthermore, when the score obtained relies on $p^{\star}(y_t|y_{1:t-1}, M_k)$ solely through its value at the realization of y_t, the scoring rule is termed local (Bernardo, 1979).

The log scoring rule, introduced by Good (1952), is the log predictive density of a future outcome evaluated at this outcome. It is thus a log predictive likelihood.[6] This rule gives high score to a predictive density when the realized outcome is located in an area of high predictive density and low

[5] Note that in the Bayesian approach, a scoring rule can also be viewed as an utility function (see Bernardo and Smith, 1994).

[6] For more details on the notions of predictive density and likelihood, see Geweke (2005, pp. 66-67).

score when it is located in an area of low predictive density (Diks et al., 2011, p. 217). Note also that the log scoring rule has the appealing feature of being the only proper local scoring rule as demonstrated by Bernardo (1979) for the continuous case. Geweke and Amisano (2010) suggest to cumulate the log scores obtained by model M_k over a given sample $y_{t_1:t_2} = (y_{t_1}, \dots, y_{t_2})'$:

$$LS(y_{t_1:t_2}, M_k) = \sum_{t=t_1}^{t_2} \ln p(y_t|y_{1:t-1}, M_k) \tag{4.4}$$

where $y_{1:t_1-1}$ is a training sample allowing predictive likelihoods to be insensitive to the initial prior density $p(\theta_k|M_k)$ for the parameter vector θ_k of model M_k. We can compare the predictive performance over $y_{t_1:t_2}$ of two competing models as follows:

$$LS(y_{t_1:t_2}, M_k) - LS(y_{t_1:t_2}, M_l) = \sum_{t=t_1}^{t_2} \ln \left[\frac{p(y_t|y_{1:t-1}, M_k)}{p(y_t|y_{1:t-1}, M_l)} \right] \tag{4.5}$$

where $p(y_t|y_{1:t-1}, M_k)/p(y_t|y_{1:t-1}, M_l)$ is the t-time predictive Bayes factor. In (4.5), the evidence in favor of M_k against M_l is cumulated along the given time period.

As pointed out by Geweke and Amisano (2011, 2012), the criterion in (4.4) is equal to the log marginal likelihood $\ln p(y_{t_1:t_2}|y_{1:t_1-1}, M_k)$.[7] It follows that the log score difference in (4.5) is equal to the log Bayes factor:

$$\ln \left[\frac{p(y_{t_1:t_2}|y_{1:t_1-1}, M_k)}{p(y_{t_1:t_2}|y_{1:t_1-1}, M_l)} \right]$$

whose interpretation is given in section 2.3. These last results highlight the relationship between model adequacy and out-of-sample predictive performance in the Bayesian approach (Geweke, 2005, p. 67, see also Kass and Raftery, 1995, p. 777).

To compute $LS(y_{t_1:t_2}, M_k)$, we have to evaluate some predictive likeli-

[7] Here, $p(\theta_k|y_{1:t_1-1}, M_k)$ is considered as the prior.

hoods. In some cases, the analytical form of $p(y_t|y_{1:t-1}, M_k)$ is available as for instance for the RW model described in theorem 2 of appendix A. But in general, the predictive likelihoods must be evaluated numerically. Fortunately, this is an easy task: The t-time predictive likelihood of model M_k can be rewritten as $E[f(\theta_k)|y_{1:t-1}, M_k]$ where $f(\theta_k) = p(y_t|y_{1:t-1}, \theta_k, M_k)$ and estimated by (Markov chain) Monte Carlo integration as explained in section 1.3. The specific form of $f(\theta_k)$ for the AR and LSTAR models will be provided in section 5.1. Note that computing $LS(y_{t_1:t_2}, M_k)$ in this way may be time-consuming since the posterior simulator must be run $(t_2 - t_1 + 1)$ times. An alternative would be to write:

$$LS(y_{t_1:t_2}, M_k) = \ln p(y_{1:t_2}|M_k) - \ln p(y_{1:t_1-1}|M_k)$$

and to estimate each term using the bridge sampling method presented in section 2.3.

4.3 Optimal Pooling

In part I, we presented and estimated linear and nonlinear models of a logistic transformation of the US unemployment rate. We saw that these models are able to capture some features of the US unemployment rate. Nevertheless, they remain intrinsically false since the underlying DGP is likely to be extremely complex. In this context, the $\mathcal{M}$-closed framework of BMA described in section 4.1 is not realistic and BMA can only be relevant for prediction when we condition on a moderate sample. Therefore, it can be interesting to consider a model averaging framework where the unknown DGP is not assumed to belong to the set of models $\mathcal{M} = \{M_1, \dots, M_K\}$ under investigation. Using the terminology of Bernardo and Smith (1994), such framework is said to be $\mathcal{M}$-open.

A predictive density $p(y_{T+1}|y_{1:T}, M_k)$ is supposed to be provided by each model in $\mathcal{M}$. These predictive densities can be combined by means of a finite mixture density:[8]

$$p_{w_T}(y_{T+1}|y_{1:T}) = \sum_{k=1}^{K} w_{T,k} p(y_{T+1}|y_{1:T}, M_k) \tag{4.6}$$

where the weight vector $w_T = (w_{T,1}, \ldots, w_{T,K})'$ only depends on data up to time T and fulfills the conditions $\sum_{k=1}^{K} w_{T,k} = 1$ and $w_{T,1}, \ldots, w_{T,K} \geq 0$ so that (4.6) is a valid density.[9] Equation (4.6) is a general device for model averaging. Setting $w_{T,k} = 1/K$ for all k, gives us the equally-weighted model averaging (EWMA) method. By defining $w_{T,k} = p(M_k|y_{1:T})$ for all k, we obtain the formal, but $\mathcal{M}$-closed, BMA method presented in section 4.1. We will now introduce the weight definition that leads to the heuristic $\mathcal{M}$-open OP method.

The principles of the OP method were first presented by Hall and Mitchell (2007). But the full theoretical analysis was derived by Geweke and Amisano (2011, 2012). In this method, we find the optimal weight vector $w_T^\star$ by solving the following problem:

$$\begin{gathered}\max_{w_T} \sum_{t=t_0+1}^{T} \ln \left[\sum_{k=1}^{K} w_{T,k} p(y_t|y_{1:t-1}, M_k) \right] \\ \text{subject to } \sum_{k=1}^{K} w_{T,k} = 1 \text{ and } w_{T,1}, \ldots, w_{T,K} \geq 0\end{gathered} \tag{4.7}$$

where the objective function cumulates the log scores of the mixture over $y_{t_0+1:T}$ given the training sample $y_{1:t_0}$. In (4.7), the optimal weights are thus selected such as to maximize the past predictive performance of the mixture.[10] The first step to solve this problem is to evaluate all predictive likelihoods as explained in section 4.2. Then, $w_T^\star$ can be obtained by using

[8] Although it is not considered here, it is also possible to aggregate predictive densities through logarithmic combination as for example in Kascha and Ravazzolo (2010).

[9] Indeed, $\sum_{k=1}^{K} w_{T,k} = 1 \Leftrightarrow \int_{-\infty}^{\infty} p_{w_T}(y_{T+1}|y_{1:T})dy_{T+1} = 1$ and weight nonnegativity ensures that (4.6) is nonnegative for all y_{T+1}.

[10] As explained in Hall and Mitchell (2007), the optimal weights can also be viewed as those minimizing the KLIC distance from the DGP to (4.6).

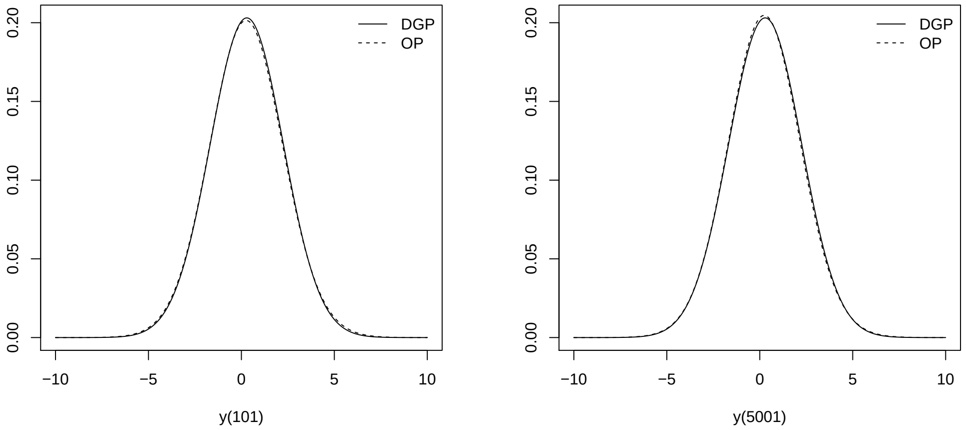


Figure 4.2: Predictive densities of the OP illustration

an appropriate numerical optimization method.

Let us return to the illustration of section 4.1. The economist now wants to use an OP of M_1 and M_2 to predict y_{101} given $y_{1:100}$ and y_{5001} given $y_{1:5000}$. The optimal weight vectors $w^\star_{100}$ and $w^\star_{5000}$ that she obtains are equal to $(0.61, 0.39)'$ and $(0.70, 0.30)'$ respectively. We remark that, unlike BMA, OP does not give a weight of one to M_1 in the large sample case. The one-step ahead predictive densities of OP are presented in figure 4.2. We observe that OP yields in both panels a better approximation of the DGP density than M_1 and M_2 in the top panel of figure 4.1.[11]

The asymptotic behavior of $w^\star_T$ was carefully studied by Geweke and Amisano (2011). In the rare cases where the DGP belongs to $\mathcal{M}$, the OP weight of the DGP will consistently converge to one as T increases. In the more common situation where the DGP is not in $\mathcal{M}$, several misspecified models will receive a positive OP weight in large samples. This phenomenon was previously illustrated. Therefore, the OP method can be useful for prediction with large as well as small samples when all models are false, unlike

[11] The reader wishing to reproduce this example should note that $w^\star_{100}$ can vary substantially from one simulation of the series to another, while $w^\star_{5000}$ remains around $(0.70, 0.30)'$.

BMA.

Some interesting applications of the OP method are already available in the literature. Hall and Mitchell (2007) use this method to combine density forecasts for UK inflation. Geweke and Amisano (2011) or more recently Durham and Geweke (2014) implement the OP method with many models for Standard & Poor's 500 log returns. Furthermore, Geweke and Amisano (2012) also apply the method to multivariate macroeconometric models of US data. While Chua et al. (2013) combine short-term interest rate models with the OP method, their focus is rather on BMA. Nevertheless, none of these studies have used this combination tool to analyze and predict the US unemployment rate as will be done in chapter 5.

Chapter 5

Evaluating Forecasts of US Unemployment

"[...] The past, the future, dwelling there, like space, inseparable together."

Kosmos
Poem of Walt Whitman

In this chapter, a forecasting competition will be held between linear and nonlinear models for the transformed monthly US unemployment rate and averages of these models obtained with the methods of chapter 4. This competition will be based on pseudo out-of-sample forecasting (Stock and Watson, 2009, pp. 103-104). This means that we will act as if we were forecasting in real-time. For each month t of the forecasting period, predictions will be made using data up to $t-1$ to estimate the individual models and to compute the model average weights. However, unlike in a true real-time situation, we will be able at each month t to directly evaluate our forecasts with the realization of the series.

Pseudo out-of-sample forecasting competitions are helpful for improving the modeling of the process under investigation (Rothman, 1998). Such exercises have already been performed with linear and nonlinear models for the US unemployment rate in Rothman (1998), Montgomery et al. (1998), Koop and Potter (1999), Clements and Smith (2000), van Dijk et al. (2002) and Deschamps (2008). Among these contributions, those of Montgomery et al. (1998) and Koop and Potter (1999) are the only ones that consider model averaging. However, their approaches differ considerably from ours. Montgomery et al. (1998) combine point forecasts from a TAR model and from a consensus of experts whereas we will combine predictive densities of RW, AR and LSTAR models. Moreover, the weights of Montgomery et al. (1998) are not recomputed in real-time. Koop and Potter (1999) implement a BMA of AR and TAR models, but only on a very short forecasting period.

The plan of this chapter is as follows. Section 5.1 will provide some theory about simulation in Bayesian forecasting. In section 5.2, we will explain how the individual models entering the model averages are selected and will describe all aspects of the forecasting experiment. In section 5.3, we will compare the real-time weights of BMA and OP over the forecasting period. In section 5.4, the forecasting performance of our models will be assessed with the help of statistical tests and log scores. Finally, in section 5.5, this assessment will be complemented by analyses based on the probability integral transformation.

5.1 Simulating Predictive Densities

For the RW model, the analytical form of the one-step ahead predictive density is provided in a conjugate framework by theorem 2 of appendix A. However, in more complex situations the predictive density can generally not be obtained analytically and must be simulated by the researcher.

The techniques used to simulate $p(y_{T+1}|y_{1:T}, M_k)$ from the AR or LSTAR model and to simulate (4.6) will now be developed. We begin by explaining the method used for individual models and then the one used for the mixture of predictive densities. In general terms, the one-step ahead predictive density of an AR or LSTAR model is written as follows:[1]

$$p(y_{T+1}|y_{1:T}, M_k) = \int p(y_{T+1}|y_{1:T}, \theta_k, M_k) p(\theta_k|y_{1:T}, M_k) d\theta_k \tag{5.1}$$

where $p(y_{T+1}|y_{1:T}, \theta_k, M_k)$ is normal for both models. The mean of $p(y_{T+1}| y_{1:T}, \theta_k, M_k)$ is given by $\phi_0 + \sum_{j=1}^{p} \phi_j y_{T+1-j}$ for the AR(p) and by:

$$\phi_{10} + \sum_{j=1}^{p} \phi_{1j} y_{T+1-j} + G(s_{T+1}; \gamma, c) \left(\phi_{20} + \sum_{j=1}^{p} \phi_{2j} y_{T+1-j} \right)$$

for the LSTAR(p). The variance is σ^2 for both models. Koop (2003, p. 73) presents the strategy that allows us to draw from (5.1). For each posterior replication $\theta_k^{(d)}$, we take a draw $y_{T+1}^{(d)}$ from $p(y_{T+1}|y_{1:T}, \theta_k^{(d)}, M_k)$. Then, the draws $y_{T+1}^{(d)}$, $d = d_0 + 1, \ldots, D$ form a sample from the predictive density in (5.1).[2]

The finite mixture of one-step ahead predictive densities given in (4.6) can be simulated with the following algorithm:

1. Compute the weight vector w_T whose elements sum to one and are nonnegative using a given method.

2. If it is not a byproduct of weight computation, generate $\theta_k^{(d)}$, $d = d_0 + 1, \ldots, D$ from $p(\theta_k|y_{1:T}, M_k)$ for the K models under consideration.

3. Construct the cumulative weights $P_k = \sum_{l=1}^{k} w_{T,l}$ for $k = 1, \ldots, K$.

4. For $d = d_0 + 1, \ldots, D$, repeat these steps:

   Step 1: Draw u from the uniform distribution $U(0, 1)$.

[1] These models were presented in section 1.2.

[2] The first d_0 posterior draws constitute the burn-in sample.

Step 2: Select model M_k where:

$$k = \begin{cases} 1 & \text{if } u \leq P_1 \\ 2 & \text{if } P_1 < u \leq P_2 \\ \vdots & \quad \vdots \\ K & \text{if } P_{K-1} < u. \end{cases}$$

Step 3: Obtain a draw $y_{T+1}^{(d)}$ from the predictive density of the selected model by drawing from $p(y_{T+1}|y_{1:T}, \theta_k^{(d)}, M_k)$.

The resulting draws $y_{T+1}^{(d)}$, $d = d_0 + 1, \ldots, D$ are a sample from (4.6). This algorithm can be used for the predictive densities of BMA, OP and EWMA, as well as for those of other model averaging methods that also enter the general framework described in (4.6). Of course, this algorithm can be computationally demanding when repeated over the forecasting period. To implement it in the forecasting experiment of this chapter, we had to parallelize the computations on a cluster of computers.

5.2 Setting Up the Experiment

Individual models in competition are AR, LSTAR and RW models for the logistic transformation of the monthly US unemployment rate. We use the prior specifications and posterior simulators presented in chapter 2 for the AR and LSTAR models. Furthermore, we use the results of theorem 2 in appendix A and prior choices made in appendix B for the RW model. For the sake of clarity, our prior choices are summarized in table 5.1. Note that a sensitivity analysis was performed for γ and c using the whole data set in section 3.1.

Specific procedures enable us to determine the composition of the different model averages in competition. For BMA, we discard the RW model and compute the PMPs of the AR(p) and LSTAR(p), $p = 1, \ldots, 8$, over time using

Table 5.1: Prior choices

	σ^2	ϕ	γ	c
RW	$IG(10^{-6}, 10^{-6})$			
AR	$IG(10^{-6}, 10^{-6})$	$N(0, I)$		
LSTAR	$IG(10^{-6}, 10^{-6})$	$N(0, I)$	$N(3, 0.1)$	$N(0, 0.1)$

expanding estimation windows.[3] The first window contains the observations from the beginning of the data set until 12 : 1979. Then, 12 observations are added to the next window, 12 more to the third and so on. The results are presented in figure 5.1. Note that the PMPs are calculated with equation (4.3) where the marginal likelihoods are obtained with the bridge sampling method developed in section 2.3 and that we only compute the PMPs once a year for computational convenience. Figure 5.1 can be interpreted in two ways. First, at each point in time the models with the highest PMPs are the ones that should be used to forecast future observations. Second, this figure gives us at each point in time the weighting scheme used by equation (4.2) to generate a predictive density at one-month horizon. As the time period on the x-axis corresponds to the forecasting period where one-month ahead predictions will be performed, we can select the models that emerge in figure 5.1 for BMA and drop out those whose PMPs are near zero. Consequently, we only retain the AR(4), AR(6), LSTAR(4) and LSTAR(3) because the first two dominate until the late 1990s and then the last two dominate at the end of the forecasting period. We let $\mathcal{M}_1$ be the set containing these models.

Being conceptually different from BMA (see chapter 4), the OP method requires another model selection procedure. As pointed out by Geweke and Amisano (2011, 2012), even weaker models can be useful in the averages generated by this method. We thus retain models of each kind (AR, LSTAR and RW) for these averages. Inside each model class, we select the best specifi-

[3] In section 5.3, we will provide evidence that the RW model can initially be discarded when we select the BMA composition.

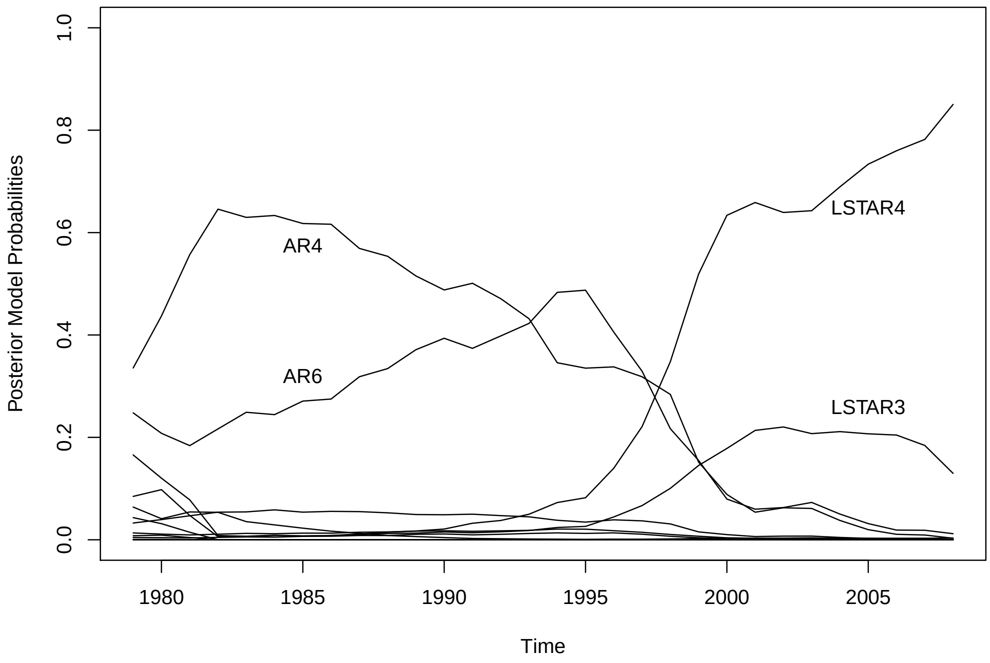
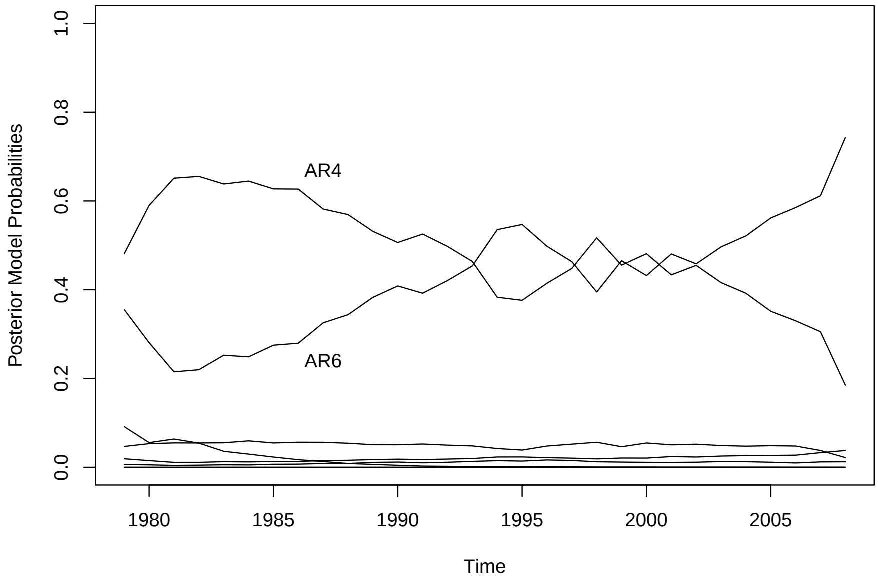


Figure 5.1: Evolution of PMPs over time for the AR(p) and LSTAR(p), $p = 1, \ldots, 8$

cations with the evolution of Bayes factors over what will be the forecasting period. In figure 5.2, we use the same marginal likelihoods than those used in the model selection for BMA in order to compute the PMPs for the AR(p), $p = 1 \ldots, 8$, in the top panel and for the LSTAR(p), $p = 1 \ldots, 8$, in the bottom panel. We observe that no other models than the AR(4) and AR(6) emerge in the top panel and that no other models than the LSTAR(4) and LSTAR(3) emerge in the bottom panel. Consequently, we retain the models in $\mathcal{M}_2 = \mathcal{M}_1 \cup \{\text{RW}\}$ for the OP method.

To sum up, we average models in $\mathcal{M}_1$ with the BMA method and those in $\mathcal{M}_2$ with the OP method. Furthermore, we also combine models in $\mathcal{M}_1$ with the EWMA method. Sophisticated model averaging methods such as BMA and OP should at least provide better forecasting performance than EWMA.



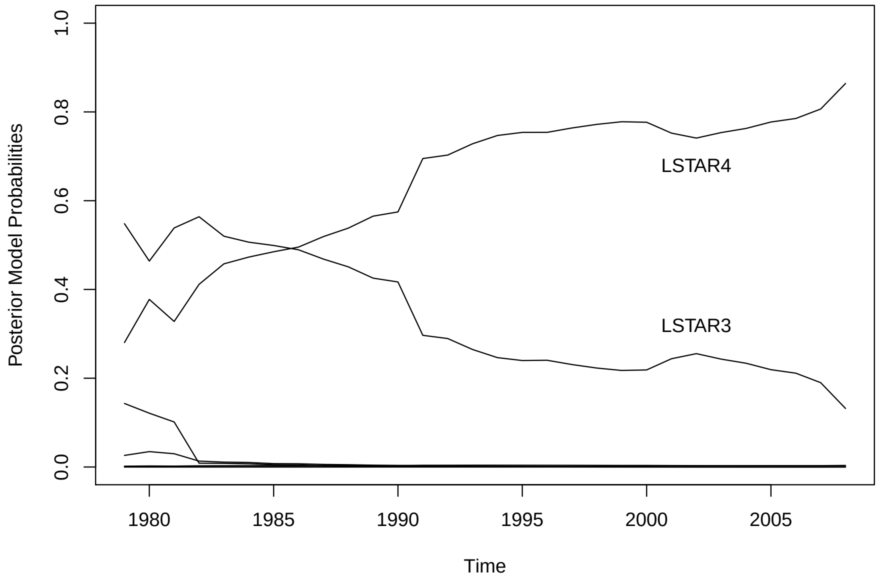


Figure 5.2: Evolution of PMPs over time for the AR(p), $p = 1, \ldots, 8$, in the top panel and for the LSTAR(p), $p = 1, \ldots, 8$, in the bottom panel

In other words, EWMA is a benchmark for other model averaging methods. In the forecasting experiment of this chapter, besides model averages, we will also present results for individual models in $\mathcal{M}_2$.

Our forecasting procedure is now detailed. We use 360 expanding windows to estimate the individual models and to compute the weights assigned to them by the model averaging methods. The first window starts at the beginning of the data set and goes until 12 : 1979. Then, we always add one month to the preceding window to obtain the next window.[4] The BMA weights, i.e. the PMPs, are computed as previously in this section except that they are now monthly updated and the OP weights are obtained from each window by solving problem (4.7) where t_0 corresponds to 9 : 1965. The dynamic behavior of these weights will be analyzed in section 5.3. For each window, the out-of-sample predictive densities of our models and model averages are simulated at one-month horizon with the techniques described in section 5.1. The first densities are thus simulated in 1 : 1980 and the last in 12 : 2009. The choice of one-month horizon is motivated by the remark of Deschamps (2008, p. 456). He indicates that the use of greater horizons provides marginal additional value when we forecast the monthly US unemployment rate. Finally, note that this forecasting procedure is very computationally demanding.

For the US unemployment rate, the literature does not provide recommendations concerning the choice between expanding or rolling estimation windows.[5] As we chose expanding windows, the log scores cumulated over the forecasting period computed in section 5.4 will be equal to log marginal likelihoods given that the smallest estimation window is considered as a training sample. However, expanding windows do not protect us against the adverse effects of structural changes, unlike rolling ones. Our model averages should nevertheless present robustness with regard to such changes.

[4] All windows start in 2 : 1949.

[5] In the rolling scheme, distant observations are discarded as new ones become available so that the window size is always the same (West and McCracken, 1998, p. 819).

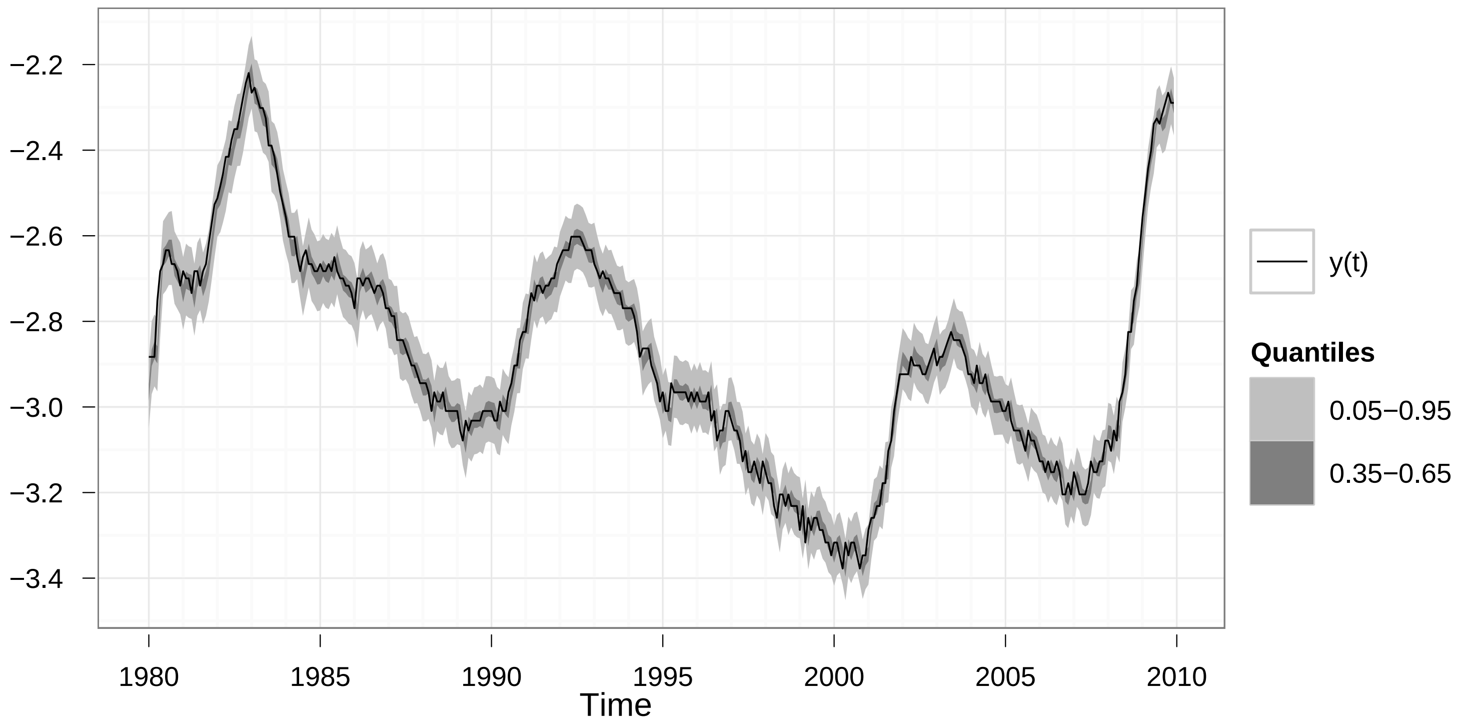


Figure 5.3: Interquantile ranges of the predictive densities of BMA compared to y_t

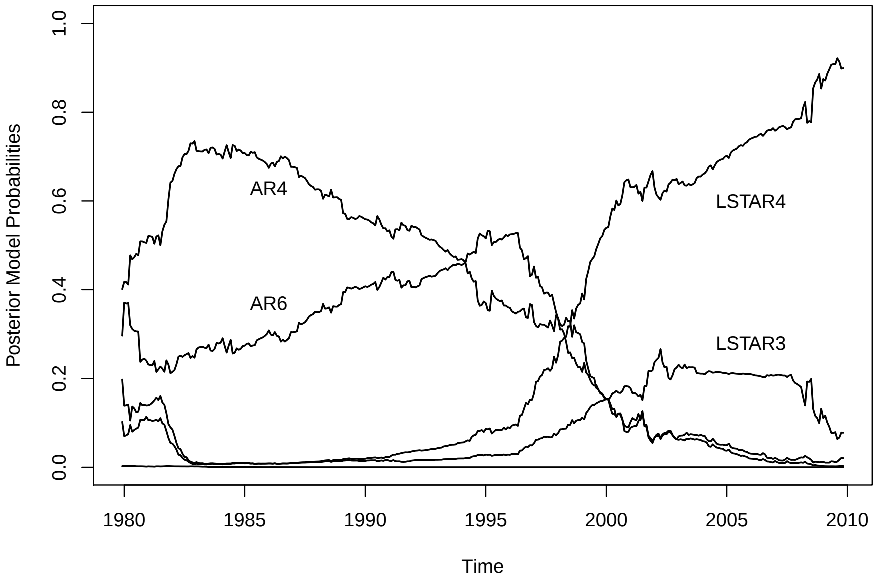
Figure 5.3 shows interquantile ranges of the one-month ahead out-of-sample predictive densities of BMA. We can see that the transformed unemployment rate y_t is most of the time in the 0.35-0.65 range. Moreover, when y_t increases (or decreases) linearly the predictive densities are more concentrated around the true values. We do not present the same plots for individual models and other model averaging methods because they are similar and do not provide additional information.

5.3 Real-Time Weights of Model Averaging Methods

It is instructive to compare the weights generated by the BMA and OP methods. In figure 5.4, these weights are plotted over the forecasting period. All details concerning their computation were provided in section 5.2. Before studying them, we must highlight that the weights of both methods are computed for the same set of models ($\mathcal{M}_2$) so that we can compare them and that they are updated in real-time, i.e. each time a new observation is available.

The models emerging in the top panel are those in $\mathcal{M}_1$. This means that the BMA method puts negligible weight on the RW model. Furthermore, this method does not select a single model since no PMP converges to one. In the bottom panel, the models that emerge are the AR(6) and LSTAR(4). The OP weights of the AR(4) and RW models are always equal to zero. Except in the early 1980s, the OP weights of the LSTAR(3) model are also always equal to zero. It is noteworthy that the OP method is more selective than the BMA method. Only a single AR model and a single LSTAR model receive positive OP weights throughout the forecasting period. We also note that the OP weights of the LSTAR(4) model start to grow more early than its PMPs.

Although the two panels of figure 5.4 show some differences, a common



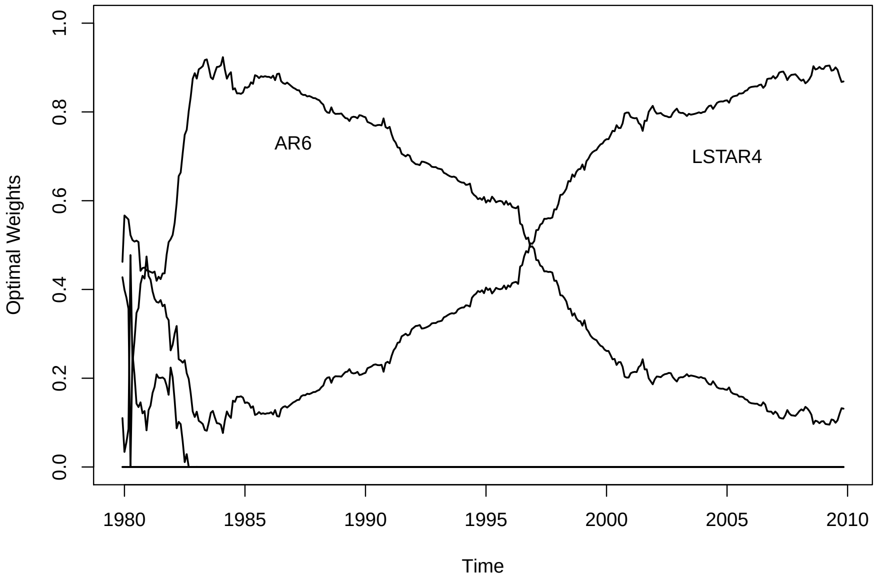


Figure 5.4: Real-time weights allocated to models in $\mathcal{M}_2$ by the BMA method (top panel) and by the OP method (bottom panel)

pattern can clearly be identified. We observe that the weights of the linear models that emerged in these panels are larger than those of the nonlinear ones until the late 1990s, while it is the opposite situation in the remaining part of the forecasting period. In other words, linearity is favored over nonlinearity when the estimation windows end before the late 1990s and nonlinearity is favored over linearity when they end after the late 1990s. This common pattern is surprising because the weights of both methods have different interpretations and properties (see chapter 4). Finally, this pattern could suggest that nonlinearities in the US unemployment process only occur temporarily.

5.4 Forecast Comparison

We will now compare the predictive performance of the models in $\mathcal{M}_3 = \mathcal{M}_2 \cup \{\text{BMA}, \text{OP}, \text{EWMA}\}$ with the Diebold-Mariano test, the efficiency test of West and McCracken and the log score approach. In this section, the index $t = 1, \ldots, 360$ represents the forecasting period which, as mentioned earlier, goes from 1 : 1980 to 12 : 2009. Over this period, we can construct point predictions $\tilde{y}_{t,k}$ with the samples drawn from the predictive densities of each model. To choose optimal point predictions, we define a loss function $L(y_t - \tilde{y}_{t,k})$, which evaluates the prediction error $y_t - \tilde{y}_{t,k}$, and minimize expected loss with respect to $\tilde{y}_{t,k}$. When the loss is quadratic $L(y_t - \tilde{y}_{t,k}) = (y_t - \tilde{y}_{t,k})^2$, the optimal point prediction is the predictive mean $\bar{y}_{t,k}$ and when the loss is linear $L(y_t - \tilde{y}_{t,k}) = |y_t - \tilde{y}_{t,k}|$, the optimal point prediction is the predictive median $y_{t,k}^{med}$. Both results are proved in theorem 3 of appendix A. In the statistical tests of this section, we will use these two point predictions and their associated loss functions in order to show the robustness of the outcomes.

Table 5.2 presents the mean squared prediction error (MSPE), defined as $\frac{1}{360}\sum_{t=1}^{360}(y_t - \bar{y}_{t,k})^2$, and the mean absolute prediction error (MAPE), defined as $\frac{1}{360}\sum_{t=1}^{360}|y_t - y_{t,k}^{med}|$, for each model. Both measures of predictive accuracy

Table 5.2: Measures of predictive accuracy

	MSPE × 100	MAPE
BMA	0.081616	0.022216
OP	0.082526	0.022268
EWMA	0.082587	0.022403
LSTAR(4)	0.083266	0.022513
LSTAR(3)	0.084826	0.022725
AR(4)	0.084995	0.022730
AR(6)	0.085021	0.022778
RW	0.095979	0.022747

provide approximately the same ranking. The model averaging methods are the most accurate and the LSTAR models are more precise than the linear ones. Among the model averaging methods, BMA provides the best performance. The statistical significance of the differences between MSPEs or MAPEs has to be investigated. Following the approach of Diebold and Mariano (1995), we test for all pairs of models $H_0 : E(d_{t,k,l}) = 0$ where $d_{t,k,l} = L(y_t - \tilde{y}_{t,k}) - L(y_t - \tilde{y}_{t,l})$ is the loss differential.[6] The implementation of this test is done in two steps. We first regress $d_{t,k,l}$ on a constant $\phi_{k,l}$. Then, the nullity of $\phi_{k,l}$ is tested with a two-sided t-test using a heteroscedasticity and autocorrelation consistent (HAC) variance estimate. The p-values resulting from this test are shown in tables 5.3 and 5.4. The loss function $L(y_t - \tilde{y}_{t,k})$ is equal to $(y_t - \bar{y}_{t,k})^2$ in the first table and to $|y_t - y_{t,k}^{med}|$ in the second. We observe that, whatever the loss structure, the gains in accuracy of the BMA method over the AR models are significant at about the 5% level. We also see that under quadratic loss the model averaging methods and the AR(4) model are significantly more precise at roughly the 5% level than the RW model. However, no model significantly improves over the RW model under linear loss. The Diebold-Mariano test does not provide further discrimination between our models, suggesting a lack of power as already mentioned by

[6] Note that testing $H_0 : E(d_{t,k,l}) = 0$ is equivalent to the test of equal predictive accuracy $H_0 : E[L(y_t - \tilde{y}_{t,k})] = E[L(y_t - \tilde{y}_{t,l})]$ between two competing models.

Table 5.3: Diebold-Mariano test p-values when using quadratic loss

	BMA	OP	EWMA	LSTAR(4)	LSTAR(3)	AR(4)	AR(6)	RW
BMA	–	0.3435	0.4638	0.5900	0.3214	**0.0428**	**0.0219**	**0.0427**
OP		–	0.9428	0.7399	0.3326	0.2588	0.1029	**0.0646**
EWMA			–	0.7249	0.2732	0.2216	0.1634	**0.0546**
LSTAR(4)				–	0.1665	0.6481	0.6113	0.1153
LSTAR(3)					–	0.9647	0.9565	0.1445
AR(4)						–	0.9858	**0.0586**
AR(6)							–	0.1028
RW								–

Table 5.4: Diebold-Mariano test p-values when using linear loss

	BMA	OP	EWMA	LSTAR(4)	LSTAR(3)	AR(4)	AR(6)	RW
BMA	–	0.7281	0.3604	0.5009	0.2746	**0.0543**	**0.0349**	0.4961
OP		–	0.3192	0.4365	0.1816	0.1613	**0.0654**	0.5396
EWMA			–	0.7043	0.2724	0.2920	0.1789	0.6431
LSTAR(4)				–	0.3316	0.6951	0.5973	0.7921
LSTAR(3)					–	0.9930	0.9176	0.9790
AR(4)						–	0.8157	0.9792
AR(6)							–	0.9646
RW								–

Table 5.5: Efficiency test p-values

	Using pred. means	Using pred. medians
BMA	0.016072	0.010171
OP	0.012757	0.008844
EWMA	0.004977	0.002310
LSTAR(4)	0.001683	0.001124
LSTAR(3)	0.000102	0.000080
AR(6)	0.000099	0.000046
AR(4)	0.000066	0.000031
RW	0.000000	0.000000

Deschamps (2008, p. 456) in a similar context.

A second test procedure, proposed by West and McCracken (1998), will now be conducted. We begin by estimating the model:

$$y_t = \phi_0 + \phi_1 \tilde{y}_{t,1} + \ldots + \phi_8 \tilde{y}_{t,8} + \epsilon_t \tag{5.2}$$

where the point predictions $\tilde{y}_{t,1}, \ldots, \tilde{y}_{t,8}$ are provided by the models in $\mathcal{M}_3$. Then, F-tests of $y_t = \tilde{y}_{t,k} + \epsilon_t$ against the unrestricted model are performed for $k = 1, ..., 8$ using a HAC covariance matrix estimate. A model that passes the test is called efficient relative to the others. Table 5.5 displays the p-values of this test. To produce the first column of the table, we used predictive means as regressors in model (5.2). For the second column, we used predictive medians. In the first column, only the BMA and OP methods pass the test at the 1% level. In the second column, only the BMA p-value is larger than 1% although the OP p-value is close to this level.

The significance level should, according to Leamer (1978, ch. 4), be a decreasing function of sample size. As the sample used for the Diebold-Mariano and efficiency tests is large, a level of 1% rather than 5% can be appropriate for both tests. In this case, the results of the Diebold-Mariano test are not significant, reinforcing the suggestion that this test lacks power.

Table 5.6: Log scores over the forecasting period

	Log score
LSTAR(4)	713.2529
OP	711.8163
BMA	711.0672
EWMA	710.1152
LSTAR(3)	710.1128
AR(6)	708.4054
AR(4)	706.0151
RW	684.1699

However, with regard to the efficiency test, the BMA and OP methods show predictive superiority.

The predictive performance of our models can also be compared using the log score approach described in section 4.2. First, we evaluate the predictive likelihoods corresponding to the predictive densities produced by the individual models over the forecasting period. Then, the predictive likelihoods of each model averaging method are formed by averaging with appropriate weights over the predictive likelihoods of individual models. Of course, each model averaging method considers a specific set of models as explained in section 5.2. Lastly, we compute (4.4) over the forecasting period for each model in $\mathcal{M}_3$. Table 5.6 reports the numerical magnitudes obtained with this approach. Remarkably, the LSTAR(4) model provides the highest log score. The rest of the ranking is roughly similar to what we obtained in table 5.2. Note that the log score reached by the OP method is slightly superior to that of the BMA method. This is not surprising since the OP weights are chosen such as to maximize the past log score of the mixture as shown in problem (4.7).

Assuming that the smallest estimation window is a training sample, we can use the log scores of table 5.6 to form BFs which can be interpreted with the help of the Jeffreys scale presented in table 2.1. By comparing the

LSTAR(4) with other models, we obtain substantial evidence against OP and BMA, strong evidence against EWMA and LSTAR(3) and decisive evidence against the remaining models. Later in this section, we will seek to identify the observations that provide such an advantage to the LSTAR(4) model in terms of log score. Among the model averaging methods, the only BF that deserves to be mentioned is the one comparing OP and EWMA. It gives substantial evidence against EWMA. It is also noteworthy that the evidence for the OP or BMA method against the AR(6) model is strong, while it is decisive against the AR(4) model. This tends to demonstrate the predictive superiority of the OP and BMA methods over the AR models. Indeed, we already obtained similar results with the Diebold-Mariano test, although they were less conclusive with regard to the OP method. Finally, the evidence against the RW model is always decisive, even when it is compared with the AR(4) model.

We will now look at the evolution of cumulative log predictive BFs over the forecasting period for some model pairs in order to evaluate the support provided by individual observations to the considered models. This investigation technique has already been implemented by Geweke and Amisano (2010), Deschamps (2012) and Durham and Geweke (2014) in financial econometrics. The cumulative log predictive BFs we consider are in favor of the following models: LSTAR(4), OP and BMA. The comparison is always made with respect to the AR(6) model. We chose this reference model since it is the best linear model in table 5.6. For these model pairs, we compute (4.5) over expanding samples always starting at the beginning of the forecasting period. These calculations are presented in the top panel of figure 5.5. In the bottom panel of this figure, we plot the corresponding observations of the transformed US unemployment rate. By comparing the two panels, we can examine the contribution of individual observations to the accumulation of evidence in favor of the different models.

Let us describe figure 5.5. In the early 1980s, as unemployment is rising sharply, we observe a strong decrease of the evidence for the LSTAR(4) model

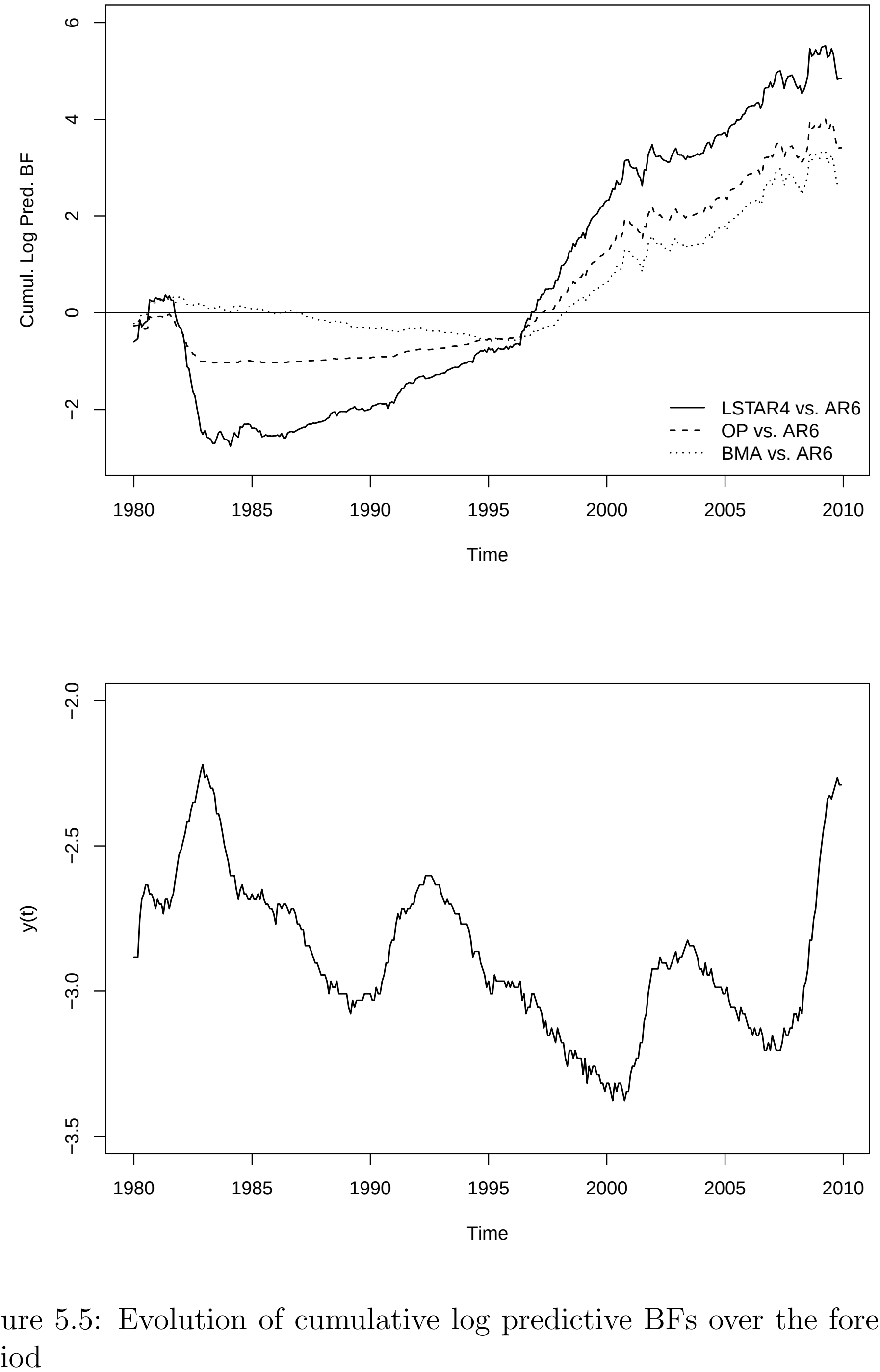


Figure 5.5: Evolution of cumulative log predictive BFs over the forecasting period

relative to other models and a moderate decrease of the evidence for the OP method relative to BMA or AR(6). These decreases allow BMA and AR(6) to be the prevailing models during at least two decades. Nevertheless, we see that in 1995 all models finally performed equivalently on the first half of the forecasting period, although the AR(6) still keeps a small advantage. In 1996, a break occurs and the cumulative log predictive BFs start to behave differently. The volatile decrease of unemployment that occurs from 1996 to 2000 contributes to seriously reduce the evidence for the AR(6) model relative to other models. This leads the LSTAR(4) to become the prevailing model and the model averaging methods to present better forecasting performance than the AR(6) model. After 2000, the evidence for the LSTAR(4) model or for the model averaging methods against the AR(6) model continues to increase until the end of the forecasting period.

Throughout this section, we saw that the BMA and OP methods are valuable tools to predict the US unemployment rate in the short-term. Although they are outperformed by the LSTAR(4) model in the log score approach (table 5.6), these methods provide superior predictive performance with regard to the statistical tests. Moreover, they provide results that are somewhat better than those of their naive benchmark, the EWMA method. On the other hand, discriminating between them is a difficult task; the statistical tests slightly favor the BMA method, while the log scores of table 5.6 marginally favor the OP method. In this section, we also saw that the AR and RW models provide poor predictive performance. Nevertheless, the cumulative log predictive BFs reveal that the AR formulation is sometimes very attractive. Indeed, for some observations in the early 1980s the AR(6) model strongly outperforms the LSTAR(4) model.

5.5 The Probability Integral Transformation

In this section, we will use the probability integral transformation (PIT) advocated by Rosenblatt (1952) to investigate whether the predictive densities generated by our models depart too much from those of the DGP. This approach to evaluate density forecasts was proposed by Diebold et al. (1998) and extended by Berkowitz (2001). It is argued in Clements and Smith (2000) that this technique may be relevant to discriminate between linear and nonlinear models. We think that it could also be interesting to apply it to mixtures of predictive densities arising from linear and nonlinear models. For a given model (which can be a model average), let us consider the sequence of normalized PITs:

$$z_t = \Phi^{-1}\left[\int_{-\infty}^{y_t} p(x|y_{1:t-1}, M_k)dx\right] \tag{5.3}$$

that can be formed over the forecasting period where $\Phi(\bullet)$ is the standard normal distribution function. If the predictive densities generated by M_k coincide with those of the DGP, then the z_t are i.i.d. $N(0,1)$ ex ante. On the basis of the normalized PITs evaluated at the realized y_t, statistical tests can be implemented in order to detect deviations from the independent standard normal. Such deviations suggest that the model is misspecified (Deschamps, 2012, p. 3043).

Given that we hold samples from the predictive densities of our models, we can easily compute their normalized PITs using Monte Carlo integration as is described in Deschamps (2008, p. 455). Of course, this procedure is not necessary in the case of the RW model since the analytical form of its predictive density is available (see theorem 2 in appendix A).

For each model in $\mathcal{M}_3$, we evaluate the behavior of the sequence of normalized PITs computed over the forecasting period with the battery of tests proposed by Deschamps (2012, p. 3044). To assess independence, we perform a F-test of the nullity of the coefficients (intercept excluded) in the regression of z_t on a constant and on $z_{t-1}, \ldots, z_{t-12}$ as well as in the regression of z_t^2 on a constant and on $z_{t-1}^2, \ldots, z_{t-12}^2$. The first test is labeled as the AR test and the second as the autoregressive conditional heteroscedasticity (ARCH) test. To assess normality, we implement the Bera-Jarque (BJ) test

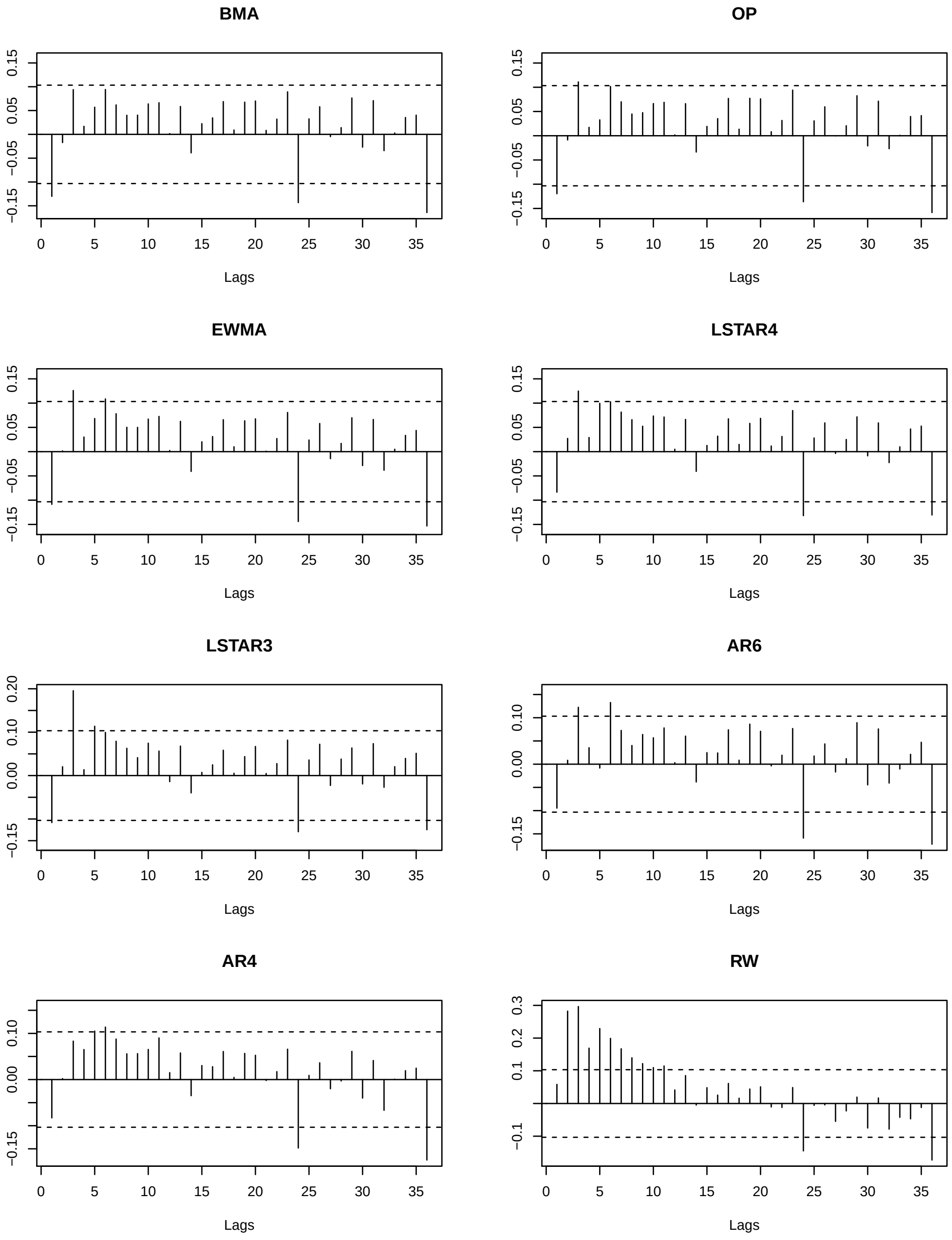


Figure 5.6: Sample ACFs for normalized PITs of models in $\mathcal{M}_3$ with 95% confidence intervals around zero

Table 5.7: p-values of PIT diagnostics

	AR	ARCH	BJ	LR
BMA	0.0026	0.3734	0.0131	0.0000
OP	0.0048	0.4985	0.0066	0.0000
EWMA	0.0023	0.3661	0.0189	0.0000
LSTAR(4)	0.0047	0.3228	0.0311	0.0000
LSTAR(3)	0.0005	0.1238	0.0201	0.0000
AR(6)	0.0062	0.4517	0.0083	0.0000
AR(4)	0.0008	0.0631	0.0186	0.0000
RW	0.0000	0.0000	0.0000	0.0000

as well as a likelihood ratio (LR) test of the $N(0,1)$ null hypothesis against a normal alternative with unconstrained moments. The p-values obtained from these diagnostics are reported in table 5.7. The AR test provides evidence of autocorrelation in the z_t of each model. This test seems nevertheless to be somewhat severe because when we consider the sample ACFs of normalized PITs presented in figure 5.6, we only observe significant autocorrelation in the case of the RW model. However, the AR test is obviously a more formal and powerful procedure. Regarding the ARCH test, there only is evidence of conditional heteroscedasticity for the RW model. Concerning the BJ test, it rejects normality at the 1% level for the OP, AR(6) and RW models. Clear evidence of misspecification is provided by the LR test; its null hypothesis is strongly rejected for all models. In order to know whether the problem comes from the mean or the variance, we carry out for every model a separate LR test on each moment (considering the other parameter as unknown). The p-values of these tests are reported in table 5.8. Furthermore, we also compute the sample mean and variance of the z_t of each model. We obtain values between 0.04 and 0.13 for the sample means and between 0.39 and 0.42 for the sample variances. These figures enable us to understand that, for each model, the misspecification revealed by the initial LR test is mainly due to the variance of the z_t. This suggests that the predictive densities generated

Table 5.8: p-values of LR tests on individual moments

	Zero mean	Unit var.
BMA	0.0107	0.0000
OP	0.0016	0.0000
EWMA	0.0024	0.0000
LSTAR(4)	0.0001	0.0000
LSTAR(3)	0.0003	0.0000
AR(6)	0.0119	0.0000
AR(4)	0.0355	0.0000
RW	0.2305	0.0000

by our models are generally too dispersed. This issue could be overcome in further research by making more informative prior choices or by using rolling rather than expanding estimation windows.

To conclude, the PIT technique was not particularly helpful for discriminating between our models. However, it suggested us new ways of investigation to improve the predictive densities generated by our models.

Conclusion

Results

The main empirical results obtained in part I are the following. The specification search conducted in section 3.1 led us to select the AR(6) as the best AR model for the transformed US unemployment series and the LSTAR(4) as the best LSTAR model for the same series. However, it was difficult to discriminate between the two best models with the help of the BIC and the marginal likelihood because the former encouraged the use of the AR(6) while the latter favored the LSTAR(4). We interpreted these contradictory results as evidence of model uncertainty on the whole data set. In section 3.2, we presented the estimation results for the AR(6) and LSTAR(4) models. The MCMC diagnostics showed that our posterior simulators are reliable. Nevertheless, we noted that the mixing for the parameters of the transition function in the LSTAR model could be slightly improved. Finally, we observed in this section that a well-chosen LSTAR model is able to successfully identify the two asymmetric regimes of the US unemployment rate.

In part II, we presented the BMA and OP methods and used them to combine predictive densities of linear and nonlinear models. The exact composition of these model averages was determined through specific procedures in section 5.2. We retained the AR(4), AR(6), LSTAR(3) and LSTAR(4) models for the BMA method and the same models together with a RW model for the OP method. The principal findings of part II relate to the model average weights and the predictive performance of the models and model averages in competition. Regarding the real-time weights produced by the BMA and OP methods, we pointed out in section 5.3 that they exhibited a similar pattern. The AR models with nonzero weights received more weights than the LSTAR models on roughly the first half of the forecasting period, while the contrary occurred on the second half of the forecasting period. This outcome was surprising since both model averaging methods are fundamentally different as explained in chapter 4. Another interesting outcome was that the OP method attributed nonzero weights to a smaller number of models than the BMA method.

Concerning the evaluation of predictive performance, the statistical tests used in section 5.4 indicated that the BMA and OP methods performed better than the other models considered. Note that this evidence was mainly provided by the efficiency test. On the other hand, we saw in the same section that these methods were outperformed by the LSTAR(4) model when predictive performance was assessed with the log scoring rule. It is also noteworthy that it was difficult to discriminate between the BMA and OP methods and that these sophisticated methods were more accurate than the naive EWMA benchmark. In this section, we also investigated the support provided by individual observations to the different models with the evolution over time of cumulative log predictive BFs and observed that the AR model was highly attractive at the beginning of the forecasting period. Lastly, the PIT framework implemented in section 5.5 turned out not to be very useful for comparing the predictive performance of our models. Instead, it enabled us to identify how we could improve the predictive densities produced by our models in further research.

In the Introduction, we suggested that linear and nonlinear models could describe the US unemployment process in a complementary fashion. The behavior of the BMA and OP weights as well as that of the cumulative log predictive BFs seemed to support this presumption. Moreover, the good predictive performance exhibited by our sophisticated averages of linear and nonlinear models also argued in favor of this assertion. Regarding the BMA method, we explained in section 4.1 that it is difficult to know a priori whether it will perform well in a particular context or not. Therefore, it was instructive to see in chapter 5 that a BMA of AR and LSTAR models is able to improve short-term predictions of the US unemployment rate. The OP method was indeed a close competitor, however it does not provide a formal treatment of model uncertainty, unlike BMA.

Further Research

Suggestions for further research that we are going to make here can be classified into two groups. The purpose of the first group of suggestions is to overcome some issues encountered in this thesis, while that of the second is to expand the scope of our research.

The propositions in the first group are the following. As the large-scale forecasting experiment conducted in chapter 5 was very time-consuming, we performed parallel computations on multiple central processing units (CPUs). Another way to sharply speed up computations would be to carry out estimation of the AR and LSTAR models with the sequential posterior simulator recently developed by Durham and Geweke (2013) within a graphics processing unit (GPU) environment. This algorithm is especially devised for GPU massive parallelization and can provide predictive and marginal likelihoods as well as PITs as byproducts. Another issue was revealed by the PIT framework in section 5.5. The PIT diagnostics suggested that our models produced predictive densities that are often too uncertain. We mentioned in this section that this phenomenon may come from the use of expanding rather than rolling estimation windows. It could be interesting in further research to determine which kind of estimation window is the most appropriate for predicting the US unemployment rate.

The second group consists of the following suggestions. Golan and Perloff (2004) propose a nonlinear nonparametric method that predicts remarkably well the US unemployment rate. Comparing the predictive performance of our averages of linear and nonlinear parametric models with that of their nonparametric approach might be an exercise for future research. In this thesis, we focused on the statistical performance of forecasting models for the US unemployment rate. Nevertheless, evaluating the economic performance of these models as for instance in Hoogerheide et al. (2010) would also be an avenue for further research. Finally, the OP method could be extended to other scoring rules than the logarithmic one. Some candidate scoring rules are proposed in Gneiting and Raftery (2007) or Diks et al. (2011).

Appendix A

Some Theorems

Some theorems are presented in this appendix in order to complete the analysis of the main text. As unknown quantities are random variables in the Bayesian paradigm, we will use many probability distributions in these theorems. A summary of important probability distributions and their properties can usually be found in the main Bayesian textbooks (See e.g. Koop, 2003 or Carlin and Louis, 2009). To save on notation, we will in general focus on the kernels of densities instead of their full formulae. Bauwens et al. (1999, pp. 43-44) define the notion of kernel as follows:

Definition 1. *The kernel of a density $p(x)$ is a function $k(x)$ such that:*

$$p(x) = \frac{k(x)}{\int k(x)dx}.$$

To simplify, we can write $p(x) \propto k(x)$. Note that a kernel is not unique. If we multiply $k(x)$ by a constant, definition 1 remains satisfied. However, it is conventional to remove all the factors that do not depend on x in order to form the kernel of a density.

Theorem 1 presents the analytical results that allow to implement a Gibbs sampler for the AR model.

Theorem 1. *Consider the AR model of order p of equation (1.2) in matrix notation $y = X\phi + \epsilon$ where $y = (y_1, \ldots, y_T)'$, $\phi = (\phi_0, \phi_1, \ldots, \phi_p)'$ and where* *the row t of the $T \times (p+1)$ matrix X is $(1, y_{t-1}, \dots, y_{t-p})$. If a multivariate normal prior with mean vector ϕ_a and covariance matrix V_a is assumed for ϕ:*

$$p(\phi) \propto \exp\left[-\frac{1}{2}(\phi - \phi_a)' V_a^{-1}(\phi - \phi_a)\right] \tag{A.1}$$

and an independent inverted gamma prior with hyperparameters a and b is assumed for σ^2:

$$p(\sigma^2) \propto \frac{1}{(\sigma^2)^{a+1}} \exp\left(-\frac{b}{\sigma^2}\right) \tag{A.2}$$

then the full conditional posteriors of ϕ and σ^2 are respectively multivariate normal and inverted gamma:

$$p(\phi|y, \sigma^2) \propto \exp\left[-\frac{1}{2}(\phi - \phi_\star)' V_\star^{-1}(\phi - \phi_\star)\right]$$

$$p(\sigma^2|y, \phi) \propto \frac{1}{(\sigma^2)^{a_\star+1}} \exp\left(-\frac{b_\star}{\sigma^2}\right)$$

where $\phi_\star = V_\star(X'y/\sigma^2 + V_a^{-1}\phi_a)$, $V_\star = (X'X/\sigma^2 + V_a^{-1})^{-1}$, $a_\star = a + T/2$ and $b_\star = b + (y - X\phi)'(y - X\phi)/2$.

Proof. The likelihood of the AR model of order p that was defined in equation (1.3) can also be written as:

$$p(y|\phi, \sigma^2) = \frac{1}{(2\pi)^{\frac{T}{2}}\sigma^T} \exp\left[-\frac{1}{2\sigma^2}(y - X\phi)'(y - X\phi)\right]. \tag{A.3}$$

We obtain the kernel of the posterior by multiplying (A.3), (A.1) and (A.2) and by considering only the factors that depend on ϕ or σ^2:

$$\begin{aligned} p(\phi, \sigma^2|y) \propto \frac{1}{\sigma^T(\sigma^2)^{a+1}} \exp\Bigg[&-\frac{1}{2\sigma^2}(y - X\phi)'(y - X\phi) \\ &-\frac{1}{2}(\phi - \phi_a)' V_a^{-1}(\phi - \phi_a) - \frac{b}{\sigma^2}\Bigg]. \end{aligned}$$

Define $\phi_\star \equiv V_\star(X'y/\sigma^2 + V_a^{-1}\phi_a)$ and $V_\star \equiv (X'X/\sigma^2 + V_a^{-1})^{-1}$. Since

$p(\phi|y,\sigma^2) \propto p(\phi,\sigma^2|y)$, we have:

$$\begin{aligned}
p(\phi|y,\sigma^2) &\propto \exp\left[-\frac{1}{2\sigma^2}(y-X\phi)'(y-X\phi) - \frac{1}{2}(\phi-\phi_a)'V_a^{-1}(\phi-\phi_a)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\frac{y'y}{\sigma^2} - \frac{2\phi'X'y}{\sigma^2} + \frac{\phi'X'X\phi}{\sigma^2}\right.\right. \\
&\qquad\qquad \left.\left. + \phi'V_a^{-1}\phi - 2\phi'V_a^{-1}\phi_a + \phi_a'V_a^{-1}\phi_a\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\phi'\left[\frac{X'X}{\sigma^2} + V_a^{-1}\right]\phi - 2\phi'V_\star^{-1}V_\star\left[\frac{X'y}{\sigma^2} + V_a^{-1}\phi_a\right]\right)\right] \\
&\propto \exp\left[-\frac{1}{2}(\phi'V_\star^{-1}\phi - 2\phi'V_\star^{-1}\phi_\star + \phi_\star'V_\star^{-1}\phi_\star - \phi_\star'V_\star^{-1}\phi_\star)\right] \\
&\propto \exp\left[-\frac{1}{2}(\phi-\phi_\star)'V_\star^{-1}(\phi-\phi_\star)\right].
\end{aligned}$$

Let $a_\star \equiv a+T/2$ and $b_\star \equiv b+(y-X\phi)'(y-X\phi)/2$. As $p(\sigma^2|y,\phi) \propto p(\phi,\sigma^2|y)$, we see that:

$$\begin{aligned}
p(\sigma^2|y,\phi) &\propto \frac{1}{\sigma^T(\sigma^2)^{a+1}} \exp\left[-\frac{1}{2\sigma^2}(y-X\phi)'(y-X\phi) - \frac{b}{\sigma^2}\right] \\
&\propto \frac{1}{(\sigma^2)^{T/2+a+1}} \exp\left[-\frac{1}{\sigma^2}\left(b + \frac{(y-X\phi)'(y-X\phi)}{2}\right)\right] \\
&\propto \frac{1}{(\sigma^2)^{a_\star+1}} \exp\left(-\frac{b_\star}{\sigma^2}\right). \qquad \square
\end{aligned}$$

Theorem 2 presents the main analytical results for the RW model.

Theorem 2. *Consider the RW model $y_t = y_{t-1} + \epsilon_t$ where the ϵ_t are i.i.d. $N(0,\sigma^2)$. If the prior of σ^2 is inverted gamma as in (A.2), then the marginal likelihood is multivariate Student with $2a$ degrees of freedom, mean vector $\mu = (y_0,\ldots,y_{T-1})'$ and scale matrix $(b/a)I_T$:*

$$p(y) \propto \left[1 + \frac{(y-\mu)'[(b/a)I_T]^{-1}(y-\mu)}{2a}\right]^{-\frac{2a+T}{2}} \tag{A.4}$$

where $y = (y_1, \ldots, y_T)'$. The posterior of σ^2 is inverted gamma:

$$p(\sigma^2|y) \propto \frac{1}{(\sigma^2)^{a_\star+1}} \exp\left(-\frac{b_\star}{\sigma^2}\right) \tag{A.5}$$

where $a_\star = a + T/2$ and $b_\star = b + (y-\mu)'(y-\mu)/2$. The one-step ahead predictive density is univariate Student with $2a_\star$ degrees of freedom, mean y_T and scale $b_\star/a_\star$:

$$p(y_{T+1}|y) \propto \left[1 + \frac{1}{2a_\star}\frac{(y_{T+1}-y_T)^2}{b_\star/a_\star}\right]^{-\frac{2a_\star+1}{2}}.$$

Proof. We implicitly condition on y_0 to form the likelihood function as:

$$p(y|\sigma^2) = \frac{1}{(2\pi)^{\frac{T}{2}}\sigma^T} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)'(y-\mu)\right]. \tag{A.6}$$

Define $a_\star \equiv a + T/2$ and $b_\star \equiv b + (y-\mu)'(y-\mu)/2$. With the help of (A.2) and (A.6), we can formulate the marginal likelihood as follows:

$$\begin{aligned} p(y) &= \int_0^\infty p(y|\sigma^2)p(\sigma^2)d\sigma^2 \\ &= \frac{b^a}{(2\pi)^{\frac{T}{2}}\Gamma(a)} \int_0^\infty \frac{1}{(\sigma^2)^{T/2+a+1}} \exp\left[-\frac{b+(y-\mu)'(y-\mu)/2}{\sigma^2}\right] d\sigma^2 \\ &= \frac{\Gamma(a_\star)b^a}{(2\pi)^{\frac{T}{2}}\Gamma(a)b_\star^{a_\star}} \end{aligned}$$

where $\Gamma(\bullet)$ is the gamma function. The last step comes from the fact that the inverted gamma density integrates to one. Some manipulations have still to be done:

$$\begin{aligned} p(y) &\propto \frac{b^{a+T/2}}{b_\star^{a+T/2}}\, b^{-T/2} \\ &\propto \left(\frac{b_\star}{b}\right)^{-\frac{2a+T}{2}} \end{aligned}$$

$$\propto \left[1 + \frac{(y-\mu)'[(b/a)I_T]^{-1}(y-\mu)}{2a}\right]^{-\frac{2a+T}{2}}.$$

The posterior is proportional to likelihood times prior:

$$\begin{aligned} p(\sigma^2|y) &\propto p(y|\sigma^2)p(\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{T/2+a+1}} \exp\left[-\frac{b+(y-\mu)'(y-\mu)/2}{\sigma^2}\right] \\ &\propto \frac{1}{(\sigma^2)^{a_\star+1}} \exp\left(-\frac{b_\star}{\sigma^2}\right). \end{aligned}$$

Finally, the predictive density can be written as follows:

$$\begin{aligned} p(y_{T+1}|y) &= \int_0^\infty p(y_{T+1}|y,\sigma^2)p(\sigma^2|y)d\sigma^2 \\ &= \frac{b_\star^{a_\star}}{\sqrt{2\pi}\Gamma(a_\star)} \int_0^\infty \frac{1}{(\sigma^2)^{1/2+a_\star+1}} \exp\left[-\frac{b_\star+(y_{T+1}-y_T)^2/2}{\sigma^2}\right] d\sigma^2 \\ &= \frac{\Gamma(1/2+a_\star)b_\star^{a_\star}}{\sqrt{2\pi}\Gamma(a_\star)[b_\star+(y_{T+1}-y_T)^2/2]^{1/2+a_\star}} \end{aligned}$$

since the inverted gamma is a valid density. Then, we perform the following steps:

$$\begin{aligned} p(y_{T+1}|y) &\propto \frac{b_\star^{1/2+a_\star}}{[b_\star+(y_{T+1}-y_T)^2/2]^{1/2+a_\star}} b_\star^{-1/2} \\ &\propto \left[\frac{b_\star+(y_{T+1}-y_T)^2/2}{b_\star}\right]^{-\frac{2a_\star+1}{2}} \\ &\propto \left[1+\frac{1}{2a_\star}\frac{(y_{T+1}-y_T)^2}{b_\star/a_\star}\right]^{-\frac{2a_\star+1}{2}}. \end{aligned}$$

□

Various point predictions can be obtained from a predictive density. Depending on the form of the loss function, theorem 3 shows which point prediction should be used to predict a future outcome. Note that this theorem can be viewed as a particular case of proposition 5.2 in Bernardo and Smith

(1994, pp. 257-258).

Theorem 3. *Let $p(y_{T+1}|y)$ be a predictive density for the future outcome y_{T+1} given the sample $y = (y_1, \ldots, y_T)'$ and let $L(y_{T+1}, \tilde{y}_{T+1})$ be the loss associated with the point prediction $\tilde{y}_{T+1}$. If the loss is quadratic $L(y_{T+1}, \tilde{y}_{T+1}) = (y_{T+1} - \tilde{y}_{T+1})^2$, then the optimal point prediction, i.e. the one that minimizes expected loss where the expectation is with respect to $p(y_{T+1}|y)$, is the predictive mean. If the loss is linear $L(y_{T+1}, \tilde{y}_{T+1}) = |y_{T+1} - \tilde{y}_{T+1}|$, then the optimal point prediction is the predictive median.*

Proof. We start with the quadratic loss. The optimal point prediction can be found by solving:

$$\min_{\tilde{y}_{T+1}} \int_{-\infty}^{\infty} (y_{T+1} - \tilde{y}_{T+1})^2 p(y_{T+1}|y) dy_{T+1}. \tag{A.7}$$

The objective function in (A.7) can be rewritten as follows:

$$E[(y_{T+1} - \tilde{y}_{T+1})^2|y] = E(y_{T+1}^2|y) - 2\tilde{y}_{T+1}E(y_{T+1}|y) + \tilde{y}_{T+1}^2.$$

The first-order condition is then given by:

$$-2E(y_{T+1}|y) + 2\tilde{y}_{T+1} = 0.$$

Therefore, we find $\tilde{y}_{T+1} = E(y_{T+1}|y)$ which is a minimum since the second derivative of the objective function is equal to 2. In the case of the linear loss, the problem to be solved is given by:

$$\min_{\tilde{y}_{T+1}} \int_{-\infty}^{\infty} |y_{T+1} - \tilde{y}_{T+1}| p(y_{T+1}|y) dy_{T+1}. \tag{A.8}$$

The objective function in (A.8) is equal to:

$$\int_{-\infty}^{\tilde{y}_{T+1}} (\tilde{y}_{T+1} - y_{T+1}) p(y_{T+1}|y) dy_{T+1} + \int_{\tilde{y}_{T+1}}^{\infty} (y_{T+1} - \tilde{y}_{T+1}) p(y_{T+1}|y) dy_{T+1}.$$

Using the Leibniz's formula (see Sydsæter et al., 2005, sec. 4.2), we obtain the following first-order condition:

$$\int_{-\infty}^{\tilde{y}_{T+1}} p(y_{T+1}|y)dy_{T+1} - \int_{\tilde{y}_{T+1}}^{\infty} p(y_{T+1}|y)dy_{T+1} = 0$$

and find that:

$$\int_{-\infty}^{\tilde{y}_{T+1}} p(y_{T+1}|y)dy_{T+1} = 1/2.$$

Moreover, as the second derivative of the objective function is equal to $2p(\tilde{y}_{T+1}|y)$, the predictive median is indeed a minimum. □

Appendix B

Some Empirical Results

Bayesian Estimation of the RW Model

The RW model $y_t = y_{t-1} + \epsilon_t$ where the ϵ_t are i.i.d. $N(0, \sigma^2)$ is a nonstationary time series model that contains only one parameter. The main analytical results concerning this simple model are derived in theorem 2 of appendix A. We will now use some of these results to estimate the RW model and to compare it to the AR and LSTAR models. As usual, the dependent variable corresponds to transformation (1.1) and goes from 2 : 1949 to 3 : 2011 (746 observations). The prior of σ^2 is assumed to be inverted gamma as in theorem 2 and the prior hyperparameters a and b are both set to 10^{-6}.[1] Figure B.1 presents the posterior density of σ^2 which is also inverted gamma as shown in equation (A.5). The posterior mean is equal to 0.001962, a slightly higher magnitude than those obtained for the AR(6) and LSTAR(4) in section 3.2.

We will now compute the BIC and the log marginal likelihood for the RW model. The BIC of formula (2.13) is evaluated at posterior mean and gives a value of 2528.7161. This result is lower than almost all the BIC values obtained in tables 3.1 and 3.2 for the AR and LSTAR models. The log of the marginal likelihood given by (A.4) is equal to 1251.8091. The BFs comparing

[1] We use here the same sample as in chapter 3. The prior choices for σ^2 are also the same. This allows us to compare the RW model to the AR and LSTAR models.

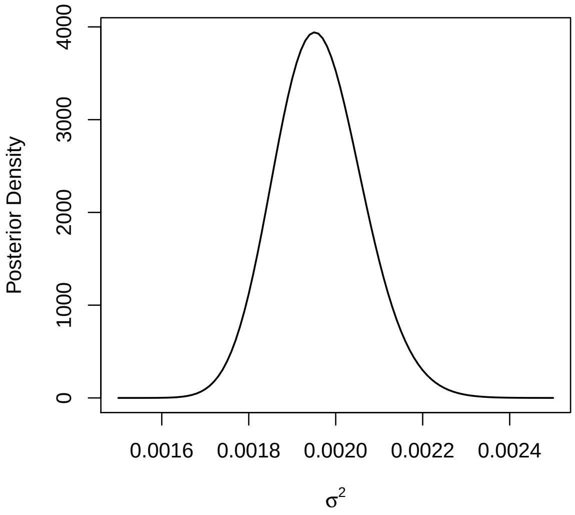


Figure B.1: The posterior for the innovation variance of the RW model

the RW model to the AR and LSTAR specifications of tables 3.1 and 3.2 give nearly always a decisive evidence against the RW model.

List of Acronyms

ACF	autocorrelation function
AR	autoregression
ARCH	autoregressive conditional heteroscedasticity
BF	Bayes factor
BIC	Bayesian information criterion
BJ	Bera-Jarque
BMA	Bayesian model averaging
CD	convergence diagnostic
CPS	current population survey
CPU	central processing unit
DGP	data generating process
DSGE	dynamic stochastic general equilibrium
ESTAR	exponential smooth transition autoregression
EWMA	equally-weighted model averaging
GPU	graphics processing unit
HAC	heteroscedasticity and autocorrelation consistent
i.i.d.	independent and identically distributed
KLIC	Kullback-Leibler information criterion
LR	likelihood ratio
LSTAR	logistic smooth transition autoregression
MAPE	mean absolute prediction error
MCMC	Markov chain Monte Carlo
MSAR	Markov switching autoregression

MSPE	mean squared prediction error
NSE	numerical standard error
OP	optimal pooling
PIT	probability integral transformation
PMP	posterior model probability
RNE	relative numerical efficiency
RW	random walk
SETAR	self-exciting threshold autoregression
STAR	smooth transition autoregression
TAR	threshold autoregression
UK	United Kingdom
US	United States

List of Figures

List of Tables

Computational Details

In this thesis, all the computations and graphs were performed with the R statistical language and environment (R Core Team, 2013). Some packages were helpful in this research. The package `coda` was used to carry out MCMC diagnostics. The package `Rsolnp` was very convenient for computing the OP weights. The package `ggplot2` helped us with its graphical power. The package `multicore` allowed us to run parallel computations on several CPUs and the package `sandwich` enabled us to estimate robust covariance matrices. Finally, many computations became feasible within a reasonable time thanks to the high performance cluster of the University of Fribourg (Switzerland).

Bibliography

J. M. Bates and C. W. J. Granger. The Combination of Forecasts. *Operational Research Quarterly*, 20(4):451–468, 1969.

L. Bauwens, M. Lubrano, and J.-F. Richard. *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, 1999.

J. Berkowitz. Testing Density Forecasts, with Applications to Risk Management. *Journal of Business and Economic Statistics*, 19(4):465–474, 2001.

J. M. Bernardo. Expected Information as Expected Utility. *The Annals of Statistics*, 7(3):686–690, 1979.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, 1994.

G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis*. CRC Press, 2009.

C. W. S. Chen and J. C. Lee. Bayesian Inference of Threshold Autoregressive Models. *Journal of Time Series Analysis*, 16(5):483–492, 1995.

C. L. Chua, S. Suardi, and S. Tsiaplias. Predicting Short-Term Interest Rates Using Bayesian Model Averaging: Evidence from Weekly and High Frequency Data. *International Journal of Forecasting*, 29(3):442–455, 2013.

M. P. Clements and J. Smith. Evaluating the Forecast Densities of Linear and Nonlinear Models: Applications to Output Growth and Unemployment. *Journal of Forecasting*, 19(4):255–276, 2000.

J. Cornfield. Bayes Theorem. *Review of the International Statistical Institute*, 35(1):34–49, 1967.

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, 1974.

M. Del Negro and F. Schorfheide. DSGE Model-Based Forecasting. In G. Elliott and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 2A, pages 57–140. North-Holland, 2013.

P. J. Deschamps. Comparing Smooth Transition and Markov Switching Autoregressive Models of US Unemployment. *Journal of Applied Econometrics*, 23(4):435–462, 2008.

P. J. Deschamps. Bayesian Estimation of Generalized Hyperbolic Skewed Student GARCH Models. *Computational Statistics and Data Analysis*, 56 (11):3035–3054, 2012.

F. X. Diebold. A Note on Bayesian Forecast Combination Procedures. In P. Hackl and A. H. Westlund, editors, *Economic Structural Change: Analysis and Forecasting*, pages 225–232. Springer-Verlag, 1991.

F. X. Diebold. The Past, Present, and Future of Macroeconomic Forecasting. *Journal of Economic Perspectives*, 12(2):175–192, 1998.

F. X. Diebold and R. S. Mariano. Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263, 1995.

F. X. Diebold, T. A. Gunther, and A. S. Tay. Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, 39(4):863–883, 1998.

C. Diks, V. Panchenko, and D. van Dijk. Likelihood-Based Scoring Rules for Comparing Density Forecasts in Tails. *Journal of Econometrics*, 163(2): 215–230, 2011.

G. Durham and J. Geweke. Adaptive Sequential Posterior Simulators for Massively Parallel Computing Environments. Unpublished Manuscript, 2013.

G. Durham and J. Geweke. Improving Asset Price Prediction When All Models Are False. *Journal of Financial Econometrics*, 12(2):278–306, 2014.

C. Fernández, E. Ley, and M. F. J. Steel. Model Uncertainty in Cross-Country Growth Regressions. *Journal of Applied Econometrics*, 16(5): 563–576, 2001.

S. Frühwirth-Schnatter. Estimating Marginal Likelihoods for Mixture and Markov Switching Models Using Bridge Sampling Techniques. *Econometrics Journal*, 7(1):143–167, 2004.

J. Geweke. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 169–193. Oxford University Press, 1992.

J. Geweke. *Contemporary Bayesian Econometrics and Statistics.* Wiley, 2005.

J. Geweke and G. Amisano. Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns. *International Journal of Forecasting*, 26 (2):216–230, 2010.

J. Geweke and G. Amisano. Optimal Prediction Pools. *Journal of Econometrics*, 164(1):130–141, 2011.

J. Geweke and G. Amisano. Prediction with Misspecified Models. *American Economic Review*, 102(3):482–486, 2012.

J. Geweke and N. Terui. Bayesian Threshold Autoregressive Models for Nonlinear Time Series. *Journal of Time Series Analysis*, 14(5):441–454, 1993.

J. Geweke and C. Whiteman. Bayesian Forecasting. In G. Elliott, C. W. J. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, pages 3–80. North-Holland, 2006.

T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477): 359–378, 2007.

A. Golan and J. M. Perloff. Superior Forecasts of the US Unemployment Rate Using a Nonparametric Method. *Review of Economics and Statistics*, 86 (1):433–438, 2004.

I. J. Good. Rational Decisions. *Journal of the Royal Statistical Society Series B*, 14(1):107–114, 1952.

I. J. Good. Weight of Evidence: A Brief Survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2*, pages 249–270. Elsevier Science Publishers, 1985.

S. G. Hall and J. Mitchell. Combining Density Forecasts. *International Journal of Forecasting*, 23(1):1–13, 2007.

J. D. Hamilton. *Time Series Analysis.* Princeton University Press, 1994.

B. E. Hansen. Inference in TAR Models. *Studies in Nonlinear Dynamics and Econometrics*, 2(1):1–14, 1997.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–401, 1999.

L. Hoogerheide, R. Kleijn, F. Ravazzolo, H. K. van Dijk, and M. Verbeek. Forecast Accuracy and Economic Gains from Bayesian Model Averaging

Using Time-Varying Weights. *Journal of Forecasting*, 29(1-2):251–269, 2010.

H. Jeffreys. *Theory of Probability.* Oxford University Press, 1961.

C. Kascha and F. Ravazzolo. Combining Inflation Density Forecasts. *Journal of Forecasting*, 29(1-2):231–250, 2010.

R. E. Kass. Bayes Factors in Practice. *The Statistician*, 42(5):551–560, 1993.

R. E. Kass and A. E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

D. E. Knuth. *The Art of Computer Programming: Fundamental Algorithms.* Addison-Wesley, 1968.

G. Koop. *Bayesian Econometrics.* Wiley, 2003.

G. Koop and S. M. Potter. Dynamic Asymmetries in US Unemployment. *Journal of Business and Economic Statistics*, 17(3):298–312, 1999.

E. E. Leamer. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Wiley, 1978.

D. V. Lindley. The Future of Statistics: A Bayesian 21st Century. *Advances in Applied Probability*, 7:106–115, 1975.

S. B. McGrayne. *The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy.* Yale University Press, 2011.

X.-L. Meng and W. H. Wong. Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration. *Statistica Sinica*, 6(4): 831–860, 1996.

C.-k. Min and A. Zellner. Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates. *Journal of Econometrics*, 56(1-2):89–118, 1993.

A. L. Montgomery, V. Zarnowitz, R. S. Tsay, and G. C. Tiao. Forecasting the US Unemployment Rate. *Journal of the American Statistical Association*, 93(442):478–493, 1998.

D. T. Mortensen and C. A. Pissarides. Job Creation and Job Destruction in the Theory of Unemployment. *The Review of Economic Studies*, 61(3): 397–415, 1994.

D. Parkinson and A. R. Liddle. Bayesian Model Averaging in Astrophysics: A Review. *Statistical Analysis and Data Mining*, 6(1):3–14, 2013.

S. M. Potter. Nonlinear Time Series Modelling: An Introduction. *Journal of Economic Surveys*, 13(5):505–528, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2013. URL `http://www.R-project.org`.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

H. V. Roberts. Probabilistic Prediction. *Journal of the American Statistical Association*, 60(309):50–62, 1965.

D. Romer. *Advanced Macroeconomics*. McGraw-Hill, 2006.

M. Rosenblatt. Remarks on a Multivariate Transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

P. Rothman. Forecasting Asymmetric Unemployment Rates. *The Review of Economics and Statistics*, 80(1):164–168, 1998.

N. Sarantis. Modeling Nonlinearities in Real Effective Exchange Rates. *Journal of International Money and Finance*, 18(1):27–45, 1999.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

J. M. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley. Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, 135(9):3209–3220, 2007.

J. H. Stock and M. W. Watson. Phillips Curve Inflation Forecasts. In J. Fuhrer, Y. K. Kodrzycki, J. S. Little, and G. P. Olivei, editors, *Understanding Inflation and the Implications for Monetary Policy: A Phillips Curve Retrospective*, pages 99–204. MIT Press, 2009.

K. Sydsæter, P. Hammond, A. Seierstad, and A. Strøm. *Further Mathematics for Economic Analysis*. FT Prentice Hall, 2005.

M. P. Taylor, D. A. Peel, and L. Sarno. Nonlinear Mean-Reversion in Real Exchange Rates: Toward a Solution to the Purchasing Power Parity Puzzles. *International Economic Review*, 42(4):1015–1042, 2001.

T. Teräsvirta. Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models. *Journal of the American Statistical Association*, 89(425):208–218, 1994.

H. Tong. On a Threshold Model. In C. H. Chen, editor, *Pattern Recognition and Signal Processing*, pages 101–141. Sijhoff and Noordhoff, 1978.

US Bureau of Labor Statistics. How the Government Measures Unemployment. Technical report, US Bureau of Labor Statistics, 2009.

US Bureau of Labor Statistics. *BLS Handbook of Methods*, 2013. URL `http://www.bls.gov/opub/hom`. accessed January 12, 2013.

D. van Dijk, T. Teräsvirta, and P. H. Franses. Smooth Transition Autoregressive Models: A Survey of Recent Developments. *Econometric Reviews*, 21(1):1–47, 2002.

K. F. Wallis. Time Series Analysis of Bounded Economic Variables. *Journal of Time Series Analysis*, 8(1):115–123, 1987.

K. D. West and M. W. McCracken. Regression-Based Tests of Predictive Ability. *International Economic Review*, 39(4):817–840, 1998.