Università
Ca'Foscari
Venezia

Universitá Cá Foscari Venezia

Scuola Dottorale di Ateneo

Graduate School

Dottorato di Ricerca in Economia

XXVI Ciclo

Anno di Discussione 2014

Essays on Markov Switching Models with Applications in Economics and Finance

Settore Scientifico Disciplinare di Afferenza: SECS-P/05

Tesi di Dottorato di Maddalena Cavicchioli (matr. 955787)

Coordinatore del Dottorato

Prof. Michele Bernasconi

Tutore del Dottorato

Prof. Monica Billio

The undersigned Maddalena Cavicchioli, in her quality of doctoral candidate for a Ph.D. degree in Economics granted by the *Universitá Cá Foscari Venezia* and the *Scuola Superiore di Economia* attests that the research exposed in this dissertation is original and that it has not been and it will not be used to pursue or attain any other academic degree of any level at any other academic institution, be it foreign or Italian.

Abstract

In this thesis we discuss problems emerging in the application of Markov Switching (MS) models both in Economics and Finance. The aim of the study is to propose solutions for model selection and estimation of multiple time series subject to regime shifts. In Chapter 1 we review the literature about dynamic systems for modeling time series with changes in regimes. In the second Chapter we investigate the problem of determining the number of regimes in MS-VARMA models and describe methods for model selection based on the autocovariance function and on stable representation of the system. Application to business cycle analysis is conducted. In Chapter 3 we introduce MS models for volatility of financial data and propose a unified framework for estimating MS-GARCH and MS-Stochastic Volatility models (duality result). In the fourth Chapter we explore other questions concerning with MS models as estimation and spectral representation. With regards to the first, we obtain simple matrix formulae for maximum likelihood estimates of parameters in the class of MS-VAR and conditional heteroskedastic models. This allows us to determine explicitly the asymptotic variance-covariance matrix of the estimators, thus giving a concrete possibility for the use of classical testing procedure. Concerning the second, we study the properties of spectral density function for MS-VAR models and derive close-form formulae for the spectral density. Several simulation exercises and applications to macroeconomic and financial data complete the work.

To my family and my husband

Acknowledgements

Firstly, I would like to acknowledge my Supervisor, Prof. Monica Billio, who introduced me into the fascinating world of non-linear dynamics since our first meeting and supported my research in these three years of PhD.

I want to thank Prof. Agar Brugiavini, who served as School Director in my first years at School and who was ready to support me when moving my first steps as PhD student, and Prof. Michele Bernasconi, the current Coordinator of the Program.

Moreover, I want to mention Prof. Mario Forni who was the Supervisor of my first research project in Econometrics during M.Sc. and who provided important inspiration for my actual works.

A big thank goes also to all the collegues met at conferences in those years that spent some of their time to discuss theoretical and empirical issues with me.

In my entire life, I couldn't achive any results without the constant and unvaluable support of my family, to whom this thesis is dedicated: many thanks are due to my dad, my mum, my sister and my grandad.

Last but certainly not least, I also dedicate this thesis to you, who became my husband last April. Any personal success is a shared success: this makes life worth living.

Contents

List of Figures **vii**

List of Tables **ix**

Introduction **xiii**

1 Some of Representations of Dynamic Systems.
Modeling Time Series with Changes in Regimes **1**

1.1 **Introduction** 2
1.2 **ARMA and ARIMA Representations** 2
1.3 **Characterizing ARMA Representations** 7
1.3.1 **Markov Coefficients** 7
1.3.2 **Hankel Matrix Rank** 8
1.4 **State-Space Representation** 9
1.5 **Markov Chains** 13
1.5.1 **The Definition** 14
1.5.2 **Representing a Markov chain by an Autoregression** 14
1.5.3 **Forecasts** 15
1.5.4 **Reducible Markov chain** 16
1.5.5 **Ergodic Markov chains** 17
1.5.6 **Periodic Markov chains** 18
1.6 **Time Series Models of Changes in Regime** 19
1.6.1 **The model** 19
1.6.2 **Optimal Inference for the Regime** 21

1.6.3 **Forecasts and Smoothed Inferences for the Regime** 24
1.6.4 **Forecasts for the Observed Variables** 28
1.6.5 **Maximum Likelihood Estimation of Parameters** . . 29
1.7 **EM algorithm and Likelihood function** 30
1.7.1 **EM algorithm: general principles** 31
1.7.2 **First characterization of EM algorithm** 33
1.7.3 **Second characterization of EM algorithm** 34
1.7.4 **Explicit form of the EM algorithm** 35
1.8 **State-Space Models with Markov Switching** 39
1.8.1 **Specification of the Model** 40
1.8.2 **Basic Filtering and Estimation** 40
1.8.3 **Smoothing** 46
1.9 **Determination of the Number of Regimes** 51

2 Markov-Switching VARMA Models **57**
2.1 **Determining the Number of Regimes in Markov-Switching VAR and VMA Models** 57
2.1.1 **Introduction** 58
2.1.2 **VARMA Representations** 60
2.1.3 **Markov Switching Moving Average Models** 67
2.1.4 **Markov Switching Autoregressive Models** 73
2.1.5 **Data Simulation** 78
2.1.6 **Application on foreign exchange rates** 81
2.1.7 **Conclusion** 83
2.1.8 **Appendix** 85
2.2 **Business Cycle and Markov Switching Models with Distributed Lags: a Comparison between US and Euro Area** 89
2.2.1 **Introduction** 90
2.2.2 **Markov Switching Models and Business Cycle** . . . 92
2.2.3 **The MSI(M, r) - VAR(0) Model** 94
2.2.4 **The MSI(M, r)-VARMA(p, q) Model** 97
2.2.5 **Business Cycle Models** 99
2.2.6 **Empirical Application** 105

2.2.7 **Conclusion** 110
2.2.8 **Appendix** 113

3 Markov Switching Models for Volatility: Filtering, Approximation and Duality **117**
3.1 **Introduction** 118
3.2 **Markov Switching GARCH** 121
3.3 **Auxiliary Models for MS-GARCH** 122
3.4 **State Space Representation and Filtering** 124
3.5 **Duality Results** 127
3.6 **Markov Switching Stochastic Volatility** 129
3.7 **Numerical and Empirical Applications** 131
3.8 **Conclusion** 137
3.9 **Appendix** 140

4 Estimation and Spectral Representation **145**
4.1 **Analysis of the Likelihood Function for Markov Switching VAR(CH) Models** 145
4.1.1 **Introduction** 146
4.1.2 **Time Series Models of Changes in Regime** 147
4.1.3 **The Basic Markov Switching Model** 152
4.1.4 **State-dependent Autoregressive Dynamics** 158
4.1.5 **State-dependent Multivariate ARCH Models** 163
4.1.6 **Conclusion** 168
4.2 **Spectral Density of Regime Switching VAR Models** 170
4.2.1 **Introduction** 170
4.2.2 **Markov-switching VAR(0) process** 171
4.2.3 **Markov-switching VAR(p) process** 176
4.2.4 **A numerical example** 177
4.2.5 **Long memory or Regime Switching?** 178
4.2.6 **Conclusion** 181

List of Figures

2.1 US quarterly growth rate of real GDP for the period 1951:1 - 2012:4. Data are taken from Fred website. 107

2.2 European (EU12) quarterly growth rate of real GDP for the period 1973:1 - 2012:4. Data are taken from Eurostat website. 107

2.3 Smoothed probabilities from the estimated MSI(4,0)-AR(0) model for EU quarterly growth rate of real GDP for the period 1973:1 - 2012:4. Data are taken from Eurostat. Regime 1 corresponds to strong recession, Regime 2 to contraction, Regime 3 to moderate/normal growth and Regime 4 to high growth. 109

3.1 The panel contains a time series plot of monthly one-month US Treasury bill rates (in annualized percentage term). The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained from FRED database. 133

3.2 The top panel refers to MS-GARCH model and bottom panel to the MS-SV model. They represent smoothed probabilities being in high-volatility regime. Parameters estimates are based on a data set of one-month Treasury Bill rates, reported in annualized percentage terms. The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained from FRED database. 136

4.1 Plot of spectral density functions of MS(2)-AR(1) 178

4.2 Spectra of S&P500 daily absolute log-return 179

4.3 Spectra MS(2)-AR(0) with $p_{00} = p_{11} = 0.95$ 180

4.4 Spectra MS(2)-AR(0) with $p_{00} = p_{11} = 0.9995$ 181

List of Tables

2.1 First experiment: Poisson Markov regime switching models with two states. In table we report the transition matrices Ps and Poisson Means λ for cases A–C. 79

2.2 Second experiment: Markov switching process with two states and autoregressive dynamic with one lag. In table we report the transition matrices Ps and the parameters of the process (means μs, standard deviation σs and autoregressive coefficients ϕs) for cases A–D. 80

2.3 Third experiment: Markov switching process with two states and autoregressive dynamic with two lags. In table we report the transition matrices Ps and the parameters of the process (means μs, standard deviation σs and autoregressive coefficients ϕ_1s and ϕ_2s) for cases A–C. 80

2.4 Simulation results of the first experiment: Poisson Markov regime switching models with two states for cases A–C. We report the number of regimes chosen by the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and Hannan-Quinn Criterion (HQC). TPM denotes the number of regimes chosen in accordance to the bounds presented in the present paper. 81

2.5 Simulation results of the second experiment: Markov switching process with two states and autoregressive dynamic with one lag (in short, MS(2)-AR(1) process) for cases A–D. We report the number of regimes chosen by the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and Hannan-Quinn Criterion (HQC). TPM denotes the number of regimes chosen in accordance to the bounds presented in the present paper. 82

2.6 Simulation results of the third experiment: Markov switching process with two states and autoregressive dynamic with two lags (in short, MS(2)-AR(2) process) for cases A–C. We report the number of regimes chosen by using the three-pattern method and applying different bounds either those obtained in Zhang and Stine (2001) (ZS), Francq and Zakoïan (2001) (FZ) or in the present paper (CAV). 82
2.7 Estimates of the number of regimes for quarterly data on French franc, British pound and German mark for the period from 1973:Q3 to 1988:Q4 based on penalized likelihood criteria (Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan-Quinn Criterion (HQC)) and based on three-pattern method computing the lower bounds as in Zhang and Stine (2001) (ZS), Francq and Zakoïan (2001) (FZ) or in the present paper (CAV). 83
2.8 Estimated number of regimes and distributed lags in the intercept for US and EU GDP. The time lags are from 1951:1 to 2012:4 for US GDP and from 1973:1 to 2012:4 for EU GDP. The procedure uses the bounds obtained in Section 3. 106
2.9 Estimated coefficients of the MSI(4,0)-AR(0) model for EU quarterly growth rate of real GDP for the period 1973:1 - 2012:4. Standard errors are in parenthesis. The log-likelihood value is 268.4987. 108
2.10 Transition probability matrix of the estimated MSI(4,0)-AR(0) model for EU quarterly growth rate of real GDP for the period 1973:1 - 2012:4. Standard errors are in parenthesis. 108

3.1 Estimation results of simulated data from model (11). Robust standard errors in parenthesis. [*]: Minutes for single iteration. The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained by the FRED database. 134
3.2 Estimation of the parameters in model (1) MS(2)-GARCH. Robust standard errors in parenthesis. The observables are one-month US Treasury bill rates (in annualized percentage term). The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained from FRED database. 135

3.3 Estimation of the parameters in model (11) MS(2)-SV. Robust standard errors in parenthesis. The observables are one-month US Treasury bill rates (in annualized percentage term). The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained from FRED database. 135

Introduction

This thesis discusses problems emerging in the study on non linear econometrics models that embed changes in regime: Markov Switching (MS) models. The study presents a statistical analysis of some MS models, such as MS-VARMA, MS-GARCH and MS-SV models, and it is completed with several applications in Economics and Finance. The content of the thesis consists of a review on existing literature and five original papers, which are presented in details hereafter.

In Chapter 1 we review the literature about dynamic systems for modeling time series with change in regimes. In particular, we develop some arguments from Gourieroux and Monfort (1997), Hamilton (1990, 1993, 1994), Kim (1994) and Krolzig (1997). These contents are convenient for next arguments since we will study Markov switching (MS) models taking advantage of suitable state space representations of them.

Chapter 2 is composed by two original papers. In the first we obtain stable VARMA representations of certain Autoregressive (AR) and Moving Average (MA) models with MS parameters, and examine how these representations may be exploited in order to determine empirically a lower bound for the number of Markov regimes. The results given in this work provide an improvement on the bounds found in earlier literature on the subject. In particular, the upper bounds for the stable VARMA orders are elementary functions of the dimension of the process, the number of regimes, the AR and MA orders of the initial model. If there is no cancellation, the bounds become equalities, and this solves the identification problem. Moreover, we provide an algorithm for model selection based on the results obtained in the paper where the orders of the switching model and the number of regimes are all unknown; the performance is evaluated with Monte Carlo experiments. An exercise on real data is also provided. This paper is forthcoming in the *Journal of Time Series Analysis* with the title *"Determining the Number of Regimes in Markov Switching VAR and VMA Models"*. A working paper version can be found in the *University of Venice Working Paper Series* (n. 3/2013) and it coincides with the text included in this thesis. The results of this paper have been presented at the PhD Workshop - CEMFI (Madrid, Spain, 2012), 6th CSDA International Conference on Computational and Financial Econometrics (Oviedo, Spain, 2012), QED Jamboree (Vienna, Austria, 2013) and 1st CIDE Workshop for PhD students in Econometrics and Empirical Economics - WEEE (Perugia, Italy, 2013).

In the second paper of Chapter 2 we consider a generalization of MS-VARMA models. In particular, we deal with the case in which the intercept term depends not only on the actual regime, but also on the last r regimes. We show that the most used models for business cycle analysis can be comprised into these latter models. Then we obtain their stable VARMA representation whose orders can be determined by evaluating the autocovariance function of the initial switching model. Secondly, related to the first result, we are able to propose a new and more rigorous way for the determination of the number of regimes. The application focuses on US and European business cycles and concludes that two regimes are sufficient for US economy, while the Euro area exhibits strong non-linearities and more regimes, particularly four, are necessary. The estimation of the model gives opportunity to identify regimes of this economic system. This paper is forthcoming in the journal *Rivista Italiana degli Economisti* with the title *"Business Cycle and Markov Switching Models with Distributed Lags: A Comparison Between US and Euro Area"* (joint with Monica Billio). It has been presented at 54th Riunione Scientifica Annuale della Societá Italiana degli Economisti (Bologna, Italy, 2013) and Economic Seminar Series (University of Modena and Reggio Emilia, Modena, Italy, 2013).

In Chapter 3 we study MS models for volatility: MS-GARCH and MS-SV models. More precisely, we can rewrite these models in suitable state space representations and propose approximated linear filters following the line of Kim and Nelson

(1999). Then we prove a duality theorem in the estimation of MS-GARCH by Kalman filter and various auxiliary models previously proposed in the literature. Finally, we apply these results to a simulation study and treasury bill rates showing that the proposed methods have the advantage of avoiding fine-tuning procedures implemented in most Bayesian estimation techniques. This work is currently submitted for publication with the title *"Markov Switching Models for Volatility: Filtering, Approximation and Duality"*.

Chapter 4 is composed by two papers. In the first paper we provide simple matrix expressions for the maximum likelihood estimates (MLE) and compute explicitly their corresponding limiting covariance matrices for three classes of MS-VAR models, namely: (i) Markov switching vector independent and identically distributed (i.i.d.) process (MS-VAR(0)), (ii) Markov switching vector autoregressions driven by i.i.d. innovations (MS-VAR(p)) and (iii) Markov switching autoregressions driven by Markov-switching vector ARCH(q) models (MS-VARCH(p, q)). Here, the MS-VARCH model is in the framework of the Baba, Engle, Kraft and Kroner (BEKK) formulation (see also Engle and Kroner, 1995). Furthermore, we prove consistency and asymptotic normality of MLE for such models. This paper entitled *"Analysis of the Likelihood Function for Markov Switching VAR(CH) Models"* is currently in its working paper version.

In the second paper of Chapter 4 we derive close-form formulae for the spectral density function of MS-VAR models and use these findings to investigate via spectral analysis whether S&P500 stock market returns suffer of structural changes rather than long memory. This paper entitled *"Spectral Density of Regime Switching VAR Models"* was presented at the Conference *Complex Data Modelling and Computationally Intensive Statistical Methods for Estimation and Prediction - S.Co.2013* (Politecnico di Milano, Italy, 2013).

This thesis contributes to the literature on Markov Switching models both in Economics and Finance, gives new methods for the statistical and econometric analysis of such models. The proposed methods are of service for practioners in model selection and estimation and offer new insights that can be developed for future research.

Chapter 1

Some of Representations of Dynamic Systems. Modeling Time Series with Changes in Regimes

Abstract. *In this work we study the connections between various representations of dynamic systems, with major attention to ARMA and State-Space representations. Then we illustrate the main results and basic properties of time series subject to Markovian changes in regime. An EM algorithm for obtaining maximum likelihood estimates of parameters for such processes is also presented. Our discussion is based on the fundamental work developed by Hamilton, Kim, and Krolzig. All these arguments are then extended to general state-space models with Markov switching, for which basic filtering, smoothing, and forecasting algorithms are completely described. Finally, a strategy for simultaneously selecting the number of regimes and the order of the autoregression completes the paper. [JEL Classification: C01, C05, C32]*

Keywords: Time series, ARMA, State-Space models, Markov chains, changes in regime, filtering, smoothing, forecasting, EM algorithm, test on regime number.

1.1 Introduction

In this chapter we review some literature that is going to be convenient for arguments developed in the present work. Firstly, we introduce ARMA and ARIMA processes and their state space representations, following mainly Gourieroux-Monfort's book [1]. This is valuable since in Chapters 2 and 4 we study VARMA models with change in regime taking advantage from suitable state space representations of them. Then we illustrate in details the econometric tools used to statistically analize time series with change in regimes. In this context, we completely describe an EM algorithm to obtain maximum likelihood estimates using different sources, especially the seminal paper of Hamilton [2] and Krolzig's book [6]. Moreover, we present Markov switching state space systems following Kim [5] and we give extensive proofs of the filtering and smoothing algorithms to make the reading self-contained. Finally, we discuss some results obtained by Krolzig [6] on regime number's determination for various classes of Markov switching VARMA models. In the next chapter of the thesis we will extend this kind of results for more general Markov switching processes, together with new algebraic findings. This will be completed with numerical and empirical applications, with some focus on business cycle analysis.

1.2 ARMA and ARIMA Representations

Let $\mathbf{y} = (\mathbf{y}_t)$ be an n-dimensional process, i.e., $\mathbf{y}_t = (y_{1t} \dots y_{nt})'$ is the n-vector at time t. A *white noise* $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_t)$ is a n-dimensional process with zero mean and nonsingular variance-covariance matrix $\boldsymbol{\Omega}$ [1]. The process $\mathbf{y}$ admits an ARMA(p, q) *representation* (or it is called an *Autoregressive-Moving Average of orders* p *and* q) if it satisfies a difference equation

$$\boldsymbol{\Phi}(L)\mathbf{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{\epsilon}_t \tag{1}$$

where $\boldsymbol{\Phi}(L) = \sum_{i=0}^{p} \boldsymbol{\Phi}_i L^i$ and $\boldsymbol{\Theta}(L) = \sum_{j=0}^{q} \boldsymbol{\Theta}_j L^j$ are $(n \times n)$-matrix polynomials in the lag operator L, $\boldsymbol{\Phi}_0 = \boldsymbol{\Theta}_0 = \mathbf{I}$ (identity matrix), $\boldsymbol{\Phi}_p \neq \mathbf{0}$, $\boldsymbol{\Theta}_q \neq \mathbf{0}$. Here the variables

[1] The present and the next two Sections are based on the arguments treated in Chapters 2-8 of the Gourieroux-Monfort book [1]

$\mathbf{y}_{-1}, ..., \mathbf{y}_{-p}, \boldsymbol{\epsilon}_{-1}, ..., \boldsymbol{\epsilon}_{-q}$ are assumed to be uncorrelated with $\boldsymbol{\epsilon}_t$ for every $t \geq 0$. If $q = 0$ (resp. $p = 0$), the process can be written as $\boldsymbol{\Phi}(L)\mathbf{y}_t = \boldsymbol{\epsilon}_t$ (resp. $\mathbf{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{\epsilon}_t$) and it is called an *autoregressive of order* p, AR(p) (resp. a *moving average of order* q, MA(q)). The above equation defines the process $\mathbf{y}$ without ambiguity. In fact, the long right-division of $\mathbf{I}$ by $\boldsymbol{\Phi}(L)$ gives

$$\mathbf{I} = \mathbf{Q}_t(L)\boldsymbol{\Phi}(L) + L^{t+1}\mathbf{R}_t(L)$$

with degree $\mathbf{Q}_t(L) = t$ and degree $\mathbf{R}_t(L) \leq p-1$. This division may be carried out since $\boldsymbol{\Phi}_0 = \mathbf{I}$ is invertible. Premultiplying (1) by $\mathbf{Q}_t(L)$, we get

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Q}_t(L)\boldsymbol{\Theta}(L)\boldsymbol{\epsilon}_t + \mathbf{R}_t(L)\mathbf{y}_{-1} \\ &= \sum_{j=0}^{t} \mathbf{H}_j \boldsymbol{\epsilon}_{t-j} + \sum_{j=1}^{q} \widetilde{\mathbf{h}}_j(t)\boldsymbol{\epsilon}_{-j} + \sum_{j=1}^{p} \mathbf{h}^*_j(t)\mathbf{y}_{-j} \\ &= \sum_{j=0}^{t} \mathbf{H}_j \boldsymbol{\epsilon}_{t-j} + \widetilde{\mathbf{h}}(t)\mathbf{z}_{-1} \end{aligned} \tag{2}$$

Thus $\mathbf{y}$ admits a *linear representation* with a sequence of coefficients given by $\mathbf{H}(z) = \boldsymbol{\Phi}^{-1}(z)\boldsymbol{\Theta}(z)$ and by initial condition

$$\mathbf{z}_{-1} = (\boldsymbol{\epsilon}_{-1} \ldots \boldsymbol{\epsilon}_{-q}\, \mathbf{y}_{-1} \ldots \mathbf{y}_{-p})'.$$

The coefficient matrix $\widetilde{\mathbf{h}}(t) = (\widetilde{\mathbf{h}}_1(t) \, \ldots \, \widetilde{\mathbf{h}}_q(t)\, \mathbf{h}^*_1(t) \, \ldots \, \mathbf{h}^*_p(t))$ is deterministic and depends on the time. The matrices $\mathbf{H}_j$ are called the *Markov coefficients*. Setting $\widetilde{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_t$ for every $t \geq 0$, and $\widetilde{\boldsymbol{\epsilon}}_t = 0$ for $t < 0$, Equation (2) becomes

$$\mathbf{y}_t = \mathbf{H}(L)\widetilde{\boldsymbol{\epsilon}} + \widetilde{\mathbf{h}}(t)\mathbf{z}_{-1} = \sum_{j=0}^{+\infty} \mathbf{H}_j \widetilde{\boldsymbol{\epsilon}}_{t-j} + \widetilde{\mathbf{h}}(t)\mathbf{z}_{-1} \tag{3}$$

with $\mathbf{H}(L) = \sum_{j=0}^{+\infty} \mathbf{H}_j L^j$.

The distribution of the process $\mathbf{y}$ can often be summarized by the first two moments (which we will assume to exist). For this, we impose that the sequence of the Markov coefficients $(\mathbf{H}_j)_{j\geq 0}$ is absolutely summable, i.e.,

$$\sum_{j=0}^{+\infty} ||\mathbf{H}_j|| < +\infty$$

The *mean* of $\mathbf{y}_t$ is given by

$$\mathbf{m}_t = E(\mathbf{y}_t) = \widetilde{\mathbf{h}}(t)E(\mathbf{z}_{-1})$$

The *autocovariances* can be obtained from the assumption of uncorrelation

$$\begin{aligned}\mathbf{\Gamma}(t,h) &= cov(\mathbf{y}_t, \mathbf{y}_{t-h}) \\ &= cov(\sum_{j=0}^{t} \mathbf{H}_j \boldsymbol{\epsilon}_{t-j}, \sum_{\ell=0}^{t-h} \mathbf{H}_\ell \boldsymbol{\epsilon}_{t-h-\ell}) + cov(\widetilde{\mathbf{h}}(t)\mathbf{z}_{-1}, \widetilde{\mathbf{h}}(t-h)\mathbf{z}_{-1}) \\ &= \sum_{j=0}^{t} \mathbf{H}_j \mathbf{\Omega} \mathbf{H}_{j-h}^{'} + \widetilde{\mathbf{h}}(t) var(\mathbf{z}_{-1}) \widetilde{\mathbf{h}}(t-h)^{'}\end{aligned}$$

where we used the convention $\mathbf{H}_j = \mathbf{0}$ if $j < 0$.

The process $\mathbf{y} = (\mathbf{y}_t)$ as in (3) with $\mathbf{H}_0 = \mathbf{I}$ admits an autoregressive form. As $\mathbf{H}_0 = \mathbf{I}$ is invertible, the long right division of $\mathbf{I}$ by $\mathbf{H}(L)$ gives

$$\mathbf{I} = \mathbf{P}_t(L)\mathbf{H}(L) + L^{t+1}\mathbf{N}_t(L)$$

where $\mathbf{P}_t(L)$ has degree t. Premultiplying (3) by $\mathbf{P}_t(L)$, we get

$$\mathbf{P}_t(L)\mathbf{y}_t = \widetilde{\boldsymbol{\epsilon}_t} - \mathbf{N}_t(L)\widetilde{\boldsymbol{\epsilon}}_{-1} + \mathbf{P}_t(L)\widetilde{\mathbf{h}}(t)\mathbf{z}_{-1} = \boldsymbol{\epsilon}_t + \mathbf{P}_t(L)\widetilde{\mathbf{h}}(t)\mathbf{z}_{-1}$$

as $\widetilde{\boldsymbol{\epsilon}_t} = \boldsymbol{\epsilon}_t$ if $t \geq 0$ and $\widetilde{\boldsymbol{\epsilon}}_{-1} = \mathbf{0}$. We see that

$$\mathbf{P}_t(L) = \mathbf{I} + \mathbf{\Pi}_1 L + \mathbf{\Pi}_t L^t$$

with

$$\mathbf{\Pi}(z) = \sum_{j=0}^{+\infty} \mathbf{\Pi}_j z^j = \mathbf{H}(z)^{-1}$$

Setting $\widetilde{\mathbf{y}_t} = \mathbf{y}_t$ if $t \geq 0$, $\widetilde{\mathbf{y}_t} = \mathbf{0}$ if $t < 0$, $\mathbf{h}^*(t) = \widetilde{\mathbf{h}}(t)$ if $t \geq 0$, and $\mathbf{h}^*(t) = \mathbf{0}$ if $t < 0$, we obtain

$$\mathbf{\Pi}(L)\widetilde{\mathbf{y}_t} = \boldsymbol{\epsilon}_t + \mathbf{\Pi}(L)\mathbf{h}^*(t)\mathbf{z}_{-1} \tag{4}$$

A process $\mathbf{y} = (\mathbf{y}_t)$ is called *stationary* if the mean $\mathbf{m}_t$ and the autocovariances $\mathbf{\Gamma}(t,h)$ are independent of t. In this case, we set $\mathbf{m}_t = \mathbf{m}$ (constant) and $\mathbf{\Gamma}(t,h) = \mathbf{\Gamma}(h)$.

Theorem 2.1 *Let $\mathbf{y} = (\mathbf{y}_t)$ be a stationary process.*

i) Then $\mathbf{y}$ admits an infinite moving average representation $MA(\infty)$:

$$\mathbf{y}_t = \sum_{j=0}^{+\infty} \mathbf{H}_j \boldsymbol{\epsilon}_{t-j} + \mathbf{m}$$

ii) Suppose that $\mathbf{y}$ has an ARMA representation (1) and that the polynomial $\det \Theta(z), z \in \mathbb{C}$, *has all its roots strictly outside the unit circle. Then $\mathbf{y}$ admits an infinite autoregressive representation $AR(\infty)$*

$$\boldsymbol{\epsilon}_t = \sum_{j=0}^{+\infty} \mathbf{\Pi}_j \mathbf{y}_{t-j}$$

where $\mathbf{\Pi}(L) = \mathbf{\Theta}(L)^{-1}\mathbf{\Phi}(L)$.

Proof. (i) follows immediately from (3) as

$$\mathbf{m} = \mathbf{m}_t = \widetilde{\mathbf{h}}(t)\mathbf{z}_{-1},$$

for every t.

(ii) The condition on the roots of det $\mathbf{\Theta}(z)$ implies the invertibility of the operator $\mathbf{\Theta}(L)$ defined on the values of a stationary process. □

A process $\widetilde{\mathbf{y}} = (\widetilde{\mathbf{y}}_t)$ is said to be *asymptotically stationary* if there exists a stationary process $\mathbf{y} = (\mathbf{y}_t)$ such that

$$lim_{t\to+\infty} E||\mathbf{y}_t - \widetilde{\mathbf{y}}_t||^2 = 0$$

In this case, we say that $\mathbf{y}$ (*asymptotically*) *approximates* $\widetilde{\mathbf{y}}$.

Theorem 2.2

(i) If either $\mathbf{z}_{-1} = \mathbf{0}$ or $lim_{t\to+\infty}\widetilde{\mathbf{h}}(t) = \mathbf{0}$ and $\sum_{j=0}^{+\infty} ||\mathbf{H}_j|| < +\infty$, then the process

$$\widetilde{\mathbf{y}}_t = \mathbf{H}(L)\widetilde{\boldsymbol{\epsilon}} + \widetilde{\mathbf{h}}(t)\mathbf{z}_{-1}$$

is asymptotically stationary.

(ii) If either $\mathbf{z}_{-1} = \mathbf{0}$ *or* $lim_{t\to+\infty}\mathbf{h}^*(t) = \mathbf{0}$ *and* $\sum_{j=0}^{+\infty} ||\mathbf{\Pi}_j|| < +\infty$, *then the process* $\widetilde{\mathbf{y}} = (\widetilde{\mathbf{y}}_t)$ *given by*

$$\mathbf{\Pi}(L)\widetilde{\mathbf{y}}_t = \boldsymbol{\epsilon}_t + \mathbf{\Pi}(L)\mathbf{h}^*(t)\mathbf{z}_{-1}$$

is asymptotically stationary.

To end this section we describe a generalization of ARMA processes, called ARIMA processes. A process $\mathbf{y} = (\mathbf{y}_t)$ admits an ARIMA *representation* (or it is called an *Autoregressive Integrated Moving Average*) if it satisfies a difference equation of type $\mathbf{\Phi}(L)\mathbf{y}_t = \mathbf{\Theta}(L)\boldsymbol{\epsilon}_t$ as above, where det $\mathbf{\Theta}(z)$, $z \in \mathbb{C}$, has all its roots outside the unit circle and det $\mathbf{\Phi}(z)$ has all its roots outside the unit circle but some of them equal to one.

Such processes are introduced especially for the following case. Given the *first-difference operator*

$$\Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1} = (1-L)\mathbf{y}_t$$

set

$$\mathbf{\Phi}(L) = \phi(L)\Delta^d = \phi(L)(1-L)^d$$

where det $\phi(z)$ has all its roots outside the unit circle. The model becomes

$$\phi(L)(1-L)^d\mathbf{y}_t = \mathbf{\Theta}(L)\boldsymbol{\epsilon}_t \tag{5}$$

If degree $\phi(L) = p$ and degree $\mathbf{\Theta}(L) = q$, then $\mathbf{y} = (\mathbf{y}_t)$ is said to have an ARIMA(p, d, q) *representation.* In this case, Relation (5) can be written in the form

$$\Delta^d\mathbf{y}_t = \phi(L)^{-1}\mathbf{\Theta}(L)\widetilde{\boldsymbol{\epsilon}_t} + \mathbf{R}_t(L)\Delta^d\mathbf{y}_{-1} + \sum_{j=1}^{q}\widetilde{\mathbf{h}}_j(t)\boldsymbol{\epsilon}_{-j} \tag{6}$$

where $\mathbf{R}_t(L)$ is the remainder of the long division of $\mathbf{I}$ by $\phi(L)$ up to order t and $\widetilde{\mathbf{h}}_j(t)$, $j = 1, ..., q$, are linear combinations of the coefficients of order $t, t-1, ..., t-q-1$ of the quotient of $\mathbf{I}$ by $\phi(L)$. When t goes to infinity, the second and third summands in (6) vanish, so $(\Delta^d\mathbf{y}_t)$ tends to the stationary process $\phi(L)^{-1}\mathbf{\Theta}(L)\widetilde{\boldsymbol{\epsilon}_t}$. Finally, we mention that the integer d can be interpreted as the order of the nonstationarity of an ARIMA process.

1.3 Characterizing ARMA Representations

1.3.1 Markov Coefficients

A process $\mathbf{y} = (\mathbf{y}_t)$ defined by $\mathbf{y}_t = \mathbf{H}(L)\boldsymbol{\epsilon}_t$ admits an ARMA(p, q) representation if and only if there exist two $(n \times n)$-matrix polynomials $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Theta}(L)$ such that

$$\boldsymbol{\Phi}(L)\mathbf{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{\epsilon}_t \Longleftrightarrow \boldsymbol{\Phi}(L)\mathbf{H}(L)\boldsymbol{\epsilon}_t = \boldsymbol{\Theta}(L)\boldsymbol{\epsilon}_t$$

with degree $\boldsymbol{\Phi}(L) = p$ and degree $\boldsymbol{\Theta}(L) = q$. Since $var(\boldsymbol{\epsilon}_t) = \boldsymbol{\Omega}$ is nonsingular, this is equivalent to

$$\boldsymbol{\Phi}(L)\mathbf{H}(L) = \boldsymbol{\Theta}(L)$$

which implies element by element

$$\sum_{j=0}^{min(\ell,p)} \boldsymbol{\Phi}_j \mathbf{H}_{\ell-j} = \begin{cases} \boldsymbol{\Theta}_\ell & \text{if} \quad \ell \leq q \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{7}$$

Theorem 3.1 *A process $\mathbf{y} = (\mathbf{y}_t)$, with $\mathbf{y}_t = \mathbf{H}(L)\boldsymbol{\epsilon}_t$, admits an ARMA representation if and only if the Markov coefficients sequence $(\mathbf{H}_j)$ satisfies a homogeneous linear difference equation starting from a certain index.*

Proof. Necessary Condition. From (7) we get

$$\sum_{j=0}^{p} \boldsymbol{\Phi}_j \mathbf{H}_{\ell-j} = \mathbf{0}$$

for all $\ell \geq max(p, q + 1)$.

Sufficient Condition. Let us assume that the sequence $(\mathbf{H}_j)$ satisfies

$$\sum_{j=0}^{p} \boldsymbol{\Phi}_j \mathbf{H}_{\ell-j} = \mathbf{0}$$

for all $\ell \geq r \geq p$, and $\mathbf{\Phi}_0 = \mathbf{I}$. The term $\sum_{j=0}^{p} \mathbf{\Phi}_j \mathbf{H}_{\ell-j}$ is, by definition, the ℓ-th coefficient of the convolution product $\mathbf{\Phi} * \mathbf{H}$ where $\mathbf{\Phi} = (\mathbf{\Phi}_0, \mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p, \mathbf{0}, \mathbf{0}, \ldots)$. Thus $(\mathbf{\Phi} * \mathbf{H})_\ell = \mathbf{0}$ for every $\ell \geq r$. This means that $\mathbf{\Phi}(z)\mathbf{H}(z)$ is a power series which does not contain terms of type z^ℓ for $\ell \geq r$, i.e., it is a polynomial. Then the process $\mathbf{y}$ satisfies $\mathbf{\Phi}(L)\mathbf{y}_t = \mathbf{\Phi}(L)\mathbf{H}(L)\boldsymbol{\epsilon}_t = \mathbf{\Theta}(L)\boldsymbol{\epsilon}_t$, where $\mathbf{\Theta}(L) = \mathbf{\Phi}(L)\mathbf{H}(L)$ has finite degree. □

1.3.2 Hankel Matrix Rank

The condition on the Markov coefficients can be rewritten in terms of the infinite Hankel matrix

$$\mathcal{H} = \begin{pmatrix} \mathbf{H}_0 & \mathbf{H}_1 & \mathbf{H}_2 & \cdots \\ \mathbf{H}_1 & \mathbf{H}_2 & \mathbf{H}_3 & \cdots \\ \mathbf{H}_2 & \mathbf{H}_3 & \mathbf{H}_4 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

The *column* (resp. *row*) *rank* of $\mathcal{H}$ is the largest number of linearly independent column (resp. row) sequences of $\mathcal{H}$. The *rank* of $\mathcal{H}$ is defined as

$$\rho(\mathcal{H}) = \sup_{N,M} \quad \rho(\mathcal{H}_{N,M})$$

where $\rho(\mathcal{H}_{N,M})$ is the usual rank of the finite size matrix

$$\mathcal{H}_{N,M} = \begin{pmatrix} \mathbf{H}_0 & \mathbf{H}_1 & \cdots & \mathbf{H}_M \\ \mathbf{H}_1 & \mathbf{H}_2 & \cdots & \mathbf{H}_{M+1} \\ \vdots & \vdots & & \vdots \\ \mathbf{H}_N & \mathbf{H}_{N+1} & \cdots & \mathbf{H}_{M+N} \end{pmatrix}$$

Of course, the above definitions coincide for finite size matrices. It turns out that the same is true for infinite Hankel matrices.

Theorem 3.2 *A process $\mathbf{y} = (\mathbf{y}_t)$, with $\mathbf{y}_t = \mathbf{H}(L)\boldsymbol{\epsilon}_t$, admits an ARMA representation if and only if the rank of the Hankel matrix $\mathcal{H}$ is finite.*

Proof. Let us assume that $\mathcal{H}$ has a finite row rank r_0. Then there exists an index r such that the rows of the submatrix

$$\begin{pmatrix} \mathbf{H}_0 & \mathbf{H}_1 & \cdots \\ \mathbf{H}_1 & \mathbf{H}_2 & \cdots \\ \vdots & \vdots & \\ \mathbf{H}_{r-1} & \mathbf{H}_r & \cdots \end{pmatrix}$$

contain a complete subsystem of rank r_0. This implies that the rows $(\mathbf{H}_r\, \mathbf{H}_{r+1}\, \cdots)$ are linear combinations of the preceding ones, i.e., there exist matrices $\mathbf{\Phi}_1, \cdots, \mathbf{\Phi}_r$ such that

$$(\mathbf{H}_r\, \mathbf{H}_{r+1}\, \cdots) = -\mathbf{\Phi}_1(\mathbf{H}_{r-1}\, \mathbf{H}_r\, \cdots) - \cdots - \mathbf{\Phi}_r(\mathbf{H}_0\, \mathbf{H}_1\, \cdots).$$

In other words, there exist matrices $\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_r$ such that

$$\sum_{j=0}^{r} \mathbf{\Phi}_j \mathbf{H}_{\ell-j} = \mathbf{0}$$

for all $\ell \geq r$, with $\mathbf{\Phi}_0 = \mathbf{I}$. This reasoning can be reversed. Now the result follows from Theorem 3.1. □

The results of this section hold for the nonstationary case as in (3).

1.4 State-Space Representation

Let us consider a system S

$$\begin{cases} \mathbf{z}_{t+1} = \mathbf{A}\mathbf{z}_t + \mathbf{B}\mathbf{u}_t \\ \mathbf{y}_t = \mathbf{C}\mathbf{z}_t + \mathbf{D}\mathbf{u}_t \end{cases} \tag{8}$$

for $t \geq 0$, where $\mathbf{u}_t$ is $m \times 1$, $\mathbf{z}_t$ is $K \times 1$, $\mathbf{y}_t$ is $n \times 1$ and the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ are $K \times K$, $K \times m$, $n \times K$, $n \times m$, respectively. The variables $\mathbf{u}$ (resp. $\mathbf{y}$) are called *inputs* (resp. *outputs or observations*) while $\mathbf{z}$ are called *state variables*. The first equation in (8), called *state equation*, explains how the state variable evolves according to the input. The second equation, called *measurement equation*, determines the output as a function of the state of the system. The system S can be seen as a linear map from the space of m-dimensional sequences $(\mathbf{u}_t)$ to the space of n-dimensional sequences $(\mathbf{y}_t)$. This map, also denoted by S, is characterized by the given matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ and the initial condition $\mathbf{z}_0$. We write $S = S(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{z}_0)$, and call $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{z}_0)$ a *representation* of S. A system S admits many different representations. In fact, for any nonsingular matrix $\mathbf{Q}$ of order K, we have

$$S(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{z}_0) = S(\mathbf{QAQ^{-1}}, \mathbf{QB}, \mathbf{CQ^{-1}}, \mathbf{D}, \mathbf{Qz}_0)$$

A representation $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{z}_0)$ of S is said to be *minimal* if $\mathbf{A}$ is of order $\underline{K}$, where $\underline{K}$ is the smallest possible size of the state vector. By successive substitutions in the state equation of S, we get

$$\mathbf{z}_{t+1} = \mathbf{A}^{t+1}\mathbf{z}_0 + \sum_{j=0}^{t} \mathbf{A}^j \mathbf{B} \mathbf{u}_{t-j} \tag{9}$$

Substituting (9) in the measurement equation yields

$$\mathbf{y}_t = \mathbf{C}(\mathbf{A}^t \mathbf{z}_0 + \sum_{j=0}^{t-1} \mathbf{A}^j \mathbf{B} \mathbf{u}_{t-1-j}) + \mathbf{D}\mathbf{u}_t = \mathbf{D}\mathbf{u}_t + \sum_{j=1}^{t} \mathbf{C}\mathbf{A}^{j-1}\mathbf{B}\mathbf{u}_{t-j} + \mathbf{C}\mathbf{A}^t \mathbf{z}_0$$

When the input $\mathbf{u}$ is an m-dimensional white noise $\boldsymbol{\epsilon}$, we see that $\mathbf{y}_t$ has a moving average representation

$$\mathbf{y}_t = \sum_{j=0}^{t} \mathbf{H}_j \boldsymbol{\epsilon}_{t-j} + \widetilde{\mathbf{h}}(t)\mathbf{z}_0$$

where $\mathbf{H}_0 = \mathbf{D}$, $\mathbf{H}_j = \mathbf{C}\mathbf{A}^{j-1}\mathbf{B}$ for every $j \geq 1$, and $\widetilde{\mathbf{h}}(t) = \mathbf{C}\mathbf{A}^t$. If we write down the moving average expansions for successive values $\mathbf{y}_t, \mathbf{y}_{t+1}, \cdots$, we get

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{y}_{t+1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{H}_1 & \mathbf{H}_2 & \mathbf{H}_3 & \cdots \\ \mathbf{H}_2 & \mathbf{H}_3 & \mathbf{H}_4 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{\epsilon}}_{t-1} \\ \widetilde{\boldsymbol{\epsilon}}_{t-2} \\ \vdots \end{pmatrix} + \begin{pmatrix} \mathbf{H}_0 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{H}_1 & \mathbf{H}_0 & \mathbf{0} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{\epsilon}}_t \\ \widetilde{\boldsymbol{\epsilon}}_{t+1} \\ \vdots \end{pmatrix} + \begin{pmatrix} \widetilde{\mathbf{h}}(t) \\ \widetilde{\mathbf{h}}(t+1) \\ \vdots \end{pmatrix} \mathbf{z}_0$$

$$= \overline{\mathcal{H}} \begin{pmatrix} \widetilde{\boldsymbol{\epsilon}}_{t-1} \\ \widetilde{\boldsymbol{\epsilon}}_{t-2} \\ \vdots \end{pmatrix} + \mathcal{T} \begin{pmatrix} \widetilde{\boldsymbol{\epsilon}}_{t} \\ \widetilde{\boldsymbol{\epsilon}}_{t+1} \\ \vdots \end{pmatrix} + \begin{pmatrix} \widetilde{\mathbf{h}}(t) \\ \widetilde{\mathbf{h}}(t+1) \\ \vdots \end{pmatrix} \mathbf{z}_0$$

Here $\mathcal{T}$ and $\overline{\mathcal{H}}$ are infinite Toeplitz and Hankel matrices (the latter scaled down by one element). Moreover, we have

$$\overline{\mathcal{H}} = \begin{pmatrix} \mathbf{CB} & \mathbf{CAB} & \mathbf{CA}^2\mathbf{B} & \cdots \\ \mathbf{CAB} & \mathbf{CA}^2\mathbf{B} & \mathbf{CA}^3\mathbf{B} & \cdots \\ \mathbf{CA}^2\mathbf{B} & \mathbf{CA}^3\mathbf{B} & \mathbf{CA}^4\mathbf{B} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} =$$

$$= \begin{pmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \end{pmatrix} \begin{pmatrix} \mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} & \cdots \end{pmatrix} = \mathcal{O}\mathcal{C}$$

where $\mathcal{O}$ is $\infty \times K$ and $\mathcal{C}$ is $K \times \infty$. A representation $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{z}_0)$ of the system S, where $\mathbf{A}$ is $K \times K$, is said to be *observable* (resp. *controllable*) if $\rho(\mathcal{O}) = K$ (resp. $\rho(\mathcal{C}) = K$).

Theorem 4.1 *A representation of a system S is minimal if and only if it is observable and controllable.*

Theorem 4.2 *The Hankel matrix $\overline{\mathcal{H}}$ of a state-space representation has a finite rank $\leq K$.*

Proof. We have $\rho(\overline{\mathcal{H}}) = sup_{N,M}\rho(\overline{\mathcal{H}}_{N,M}) \leq sup_N\, \rho(\mathcal{O}_N) \leq K$, where $\mathcal{O}_N$ is made of the first N rows of $\mathcal{O}$. □

Given an n-dimensional process $\mathbf{y} = (\mathbf{y}_t)$, with $\mathbf{y}_t = \sum_{j=0}^{+\infty} \mathbf{H}_j \boldsymbol{\epsilon}_{t-j}$, it admits an ARMA representation if and only if $\rho(\mathcal{H})$ is finite. Furthermore, it admits a state-space representation

$$\begin{cases} \mathbf{z}_{t+1} = \mathbf{A}\mathbf{z}_t + \mathbf{B}\boldsymbol{\epsilon}_t \\ \mathbf{y}_t = \mathbf{C}\mathbf{z}_t + \boldsymbol{\epsilon}_t \end{cases}$$

for $t \geq 0$, $\mathbf{D} = \mathbf{H}_0 = \mathbf{I}$ if and only if $\rho(\overline{\mathcal{H}})$ is finite. The matrices $\mathcal{H}$ and $\overline{\mathcal{H}}$ are different by n columns, so that $\rho(\mathcal{H})$ is finite if and only if $\rho(\overline{\mathcal{H}})$ is finite.

Theorem 4.3 *A process* $\mathbf{y} = (\mathbf{y}_t)$, *with* $\mathbf{y}_t = \sum_{j=0}^{+\infty} \mathbf{H}_j \boldsymbol{\epsilon}_{t-j}$ *and* $\mathbf{H}_0 = \mathbf{I}$, *admits an ARMA representation if and only if it has a state-space representation.*

This equivalence can be expressed in terms of a transfer function. The *transfer function* of an ARMA process $\boldsymbol{\Phi}(L)\mathbf{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{\epsilon}_t$ is given by

$$\boldsymbol{\Phi}(z)^{-1}\boldsymbol{\Theta}(z) = \sum_{j=0}^{+\infty} \mathbf{H}_j z^j$$

The transfer function associated to a state-space representation is

$$\sum_{j=0}^{+\infty} \mathbf{H}_j z^j = \mathbf{I} + \sum_{j=0}^{+\infty} \mathbf{C}\mathbf{A}^{j-1}\mathbf{B}z^j = \mathbf{I} + \mathbf{C}\sum_{j=0}^{+\infty} \mathbf{A}^{j-1} z^{j-1}\mathbf{B}z = \mathbf{I} + \mathbf{C}(\mathbf{I} - \mathbf{A}z)^{-1}\mathbf{B}z$$

Corollary 4.4 *Any rational transfer function* $\boldsymbol{\Phi}(z)^{-1}\boldsymbol{\Theta}(z)$, *where* $\boldsymbol{\Phi}_i$ *and* $\boldsymbol{\Theta}_j$ *are square matrices of order* n *can be written as*

$$\mathbf{I} + \mathbf{C}(\mathbf{I} - \mathbf{A}z)^{-1}\mathbf{B}z$$

where $\mathbf{C}, \mathbf{A}$ *and* $\mathbf{B}$ *are* $n \times K$, $K \times K$, *and* $K \times n$, *respectively. The reverse is also true.*

To end the section we illustrate a state-space representation of an ARMA(p, q) model

$$\mathbf{y}_t = -\boldsymbol{\phi}_1 \mathbf{y}_{t-1} - \cdots - \boldsymbol{\phi}_p \mathbf{y}_{t-p} + \boldsymbol{\theta}_1 \boldsymbol{\epsilon}_{t-1} + \cdots + \boldsymbol{\theta}_q \boldsymbol{\epsilon}_{t-q} + \boldsymbol{\epsilon}_t$$

where $\boldsymbol{\phi}_i$ and $\boldsymbol{\theta}_j$ are $(n \times n)$-matrices, and $\mathbf{y}_t$ and $\boldsymbol{\epsilon}_t$ are $(n \times 1)$-vectors.

Let

$$\mathbf{z}_t = (\mathbf{y}_{t-1}^{'} \ldots \mathbf{y}_{t-p}^{'} \boldsymbol{\epsilon}_{t-1}^{'} \ldots \boldsymbol{\epsilon}_{t-q}^{'})^{'}$$

be the lag-vector. It is easily seen that $\mathbf{y}_t$ and $\mathbf{z}_t$ evolve according to the following state-space model:

$$\begin{cases} \mathbf{z}_{t+1} = \mathbf{A}\mathbf{z}_t + \mathbf{B}\boldsymbol{\epsilon}_t \\ \mathbf{y}_t = \mathbf{C}\mathbf{z}_t + \boldsymbol{\epsilon}_t \end{cases}$$

where:

$$\mathbf{A} = \begin{pmatrix} -\boldsymbol{\phi}_1 & -\boldsymbol{\phi}_2 & \cdots & -\boldsymbol{\phi}_{p-1} & -\boldsymbol{\phi}_p & \boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_{q-1} & \boldsymbol{\theta}_q \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{I} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} -\boldsymbol{\phi}_1 & \cdots & -\boldsymbol{\phi}_p & \boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_q \end{pmatrix}$$

1.5 Markov Chains

Many random variables undergo episodes in which the behavior of the time series process changes quite dramatically [1]. These apparent changes can result from events such as wars, financial panics, or significant modifications in government policies. To take in account changes

[1]Sections 1.5, 1.6 and 1.7 are based on the arguments treated in the Hamilton book [4], in the Krolzig book [6], and in the Hamilton papers [2] and [3]

in the process, we might consider it to be influenced by an unobserved random variable, called *state* or *regime*, which is discrete-valued. The simplest time series model for a discrete-valued random variable is a Markov chain. This section is devoted to review the basic definitions and results about Markov chains.

1.5.1 The Definition

Let s_t be a random variable which can assume only an integer value in the set $\{1, 2, \dots, M\}$. Suppose that the probability that s_t equals some particular value j depends on the past only through the most recent value s_{t-1}, that is,

$$P(s_t = j | s_{t-1} = i, s_{t-2} = k, \dots) = P(s_t = j | s_{t-1} = i) = p_{ij}$$

Such a process $(s_t)_{t\geq 0}$ is called an *M-state Markov chain* with transition matrix

$$\mathbf{P} = (p_{ij})_{i,j=1,\dots,M} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2M} \\ \vdots & \vdots & & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MM} \end{pmatrix}$$

The term p_{ij} of $\mathbf{P}$ gives the probability that the state i will be followed by state j. Note that

$$p_{i1} + p_{i2} + \cdots + p_{iM} = 1$$

for every $i = 1, \dots, M$.

1.5.2 Representing a Markov chain by an Autoregression

An useful representation for a Markov chain is obtained by letting $\boldsymbol{\xi}_t$ denote a random $(M \times 1)$ vector whose jth element is equal to unity if $s_t = j$ and it is zero otherwise. Thus, when $s_t = 1$, the vector $\boldsymbol{\xi}_t$ coincides with the first column of the $(M \times M)$ identity matrix $\mathbf{I}_M$; when $s_t = 2$, the vector $\boldsymbol{\xi}_t$ is the second column of $\mathbf{I}_M$; and so on. If $s_t = i$, then the jth element of $\boldsymbol{\xi}_{t+1}$

is a random variable that takes on the value unity with probability p_{ij} and takes on the value zero otherwise. Such a random variable has expectation p_{ij}. Thus the conditional expectation of $\boldsymbol{\xi}_{t+1}$ given $s_t = i$ is given by

$$E(\boldsymbol{\xi}_{t+1}|s_t = i) = \begin{pmatrix} p_{i1} & p_{i2} & \cdots & p_{iM} \end{pmatrix}^{'}$$

which is the ith column of $\mathbf{P}^{'}$. Moreover, when $s_t = i$, the vector $\boldsymbol{\xi}_t$ corresponds to the ith column of $\mathbf{I}_M$, hence

$$E(\boldsymbol{\xi}_{t+1}|s_t = i) = E(\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t) = \mathbf{P}^{'}\boldsymbol{\xi}_t.$$

It follows further that

$$E(\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \dots) = E(\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t) = \mathbf{P}^{'}\boldsymbol{\xi}_t$$

This result implies that it is possible to express a Markov chain in the AR(1) form

$$\boldsymbol{\xi}_{t+1} = \mathbf{P}^{'}\boldsymbol{\xi}_t + \mathbf{v}_{t+1} \tag{10}$$

where

$$\mathbf{v}_{t+1} = \boldsymbol{\xi}_{t+1} - E(\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \cdots).$$

In particular, the innovation $\mathbf{v}_t$ is a martingale difference sequence. Although, the vector $\mathbf{v}_t$ can take on only a finite set of values, on average $\mathbf{v}_t$ is zero. Moreover, the value of $\mathbf{v}_t$ is impossible to forecast on the basis of previous states of the process.

1.5.3 Forecasts

Formula (10) implies that

$$\boldsymbol{\xi}_{t+n} = \mathbf{v}_{t+n} + \mathbf{P}^{'}\mathbf{v}_{t+n-1} + (\mathbf{P}^{'})^2\mathbf{v}_{t+n-2} + \cdots + (\mathbf{P}^{'})^{n-1}\mathbf{v}_{t+1} + (\mathbf{P}^{'})^n\boldsymbol{\xi}_t$$

where $\mathbf{P}^i$ indicates the transition matrix multiplied by itself i times. Then the n-period-ahead forecasts for a Markov chain are given by

$$E(\boldsymbol{\xi}_{t+n}|\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \dots) = (\mathbf{P}^{'})^n \boldsymbol{\xi}_t \tag{11}$$

Again, since the jth element of $\boldsymbol{\xi}_{t+n}$ will be unity if $s_{t+n} = j$ and zero otherwise, the jth element of the $(M \times 1)$ vector $E(\boldsymbol{\xi}_{t+n}|\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \dots)$ indicates the probability that s_{t+n} takes on the value j, conditional on the state of the system at date t. For example, if the process is in state i at date t, then (11) asserts that

$$\begin{pmatrix} P(s_{t+n} = 1|s_t = i) \\ P(s_{t+n} = 2|s_t = i) \\ \vdots \\ P(s_{t+n} = M|s_t = i) \end{pmatrix} = (\mathbf{P}^{'})^n \mathbf{e}_i$$

where $\mathbf{e}_i$ denotes the ith column of $\mathbf{I}_M$. The matrix formula indicates that the n-period-ahead transition probabilities for a Markov chain can be calculate by $(\mathbf{P}^{'})^n$. Specifically, the probability that an observation from regime i will be followed n periods later by an observation from regime j, that is, $P(s_{t+n} = j|s_t = i)$, is given by the row j, column i element of $(\mathbf{P}^{'})^n$.

1.5.4 Reducible Markov chain

An M-state Markov chain is said to be *reducible* if there exists a way to label the states such that the transpose of the transition matrix is upper block-triangular

$$\mathbf{P}^{'} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}$$

where $\mathbf{B}$ denotes a $(K \times K)$ matrix for some $1 \leq K < M$. In this case, $(\mathbf{P}^{'})^n$ is also upper block-triangular for any m. Hence, once such a process enters a state j such that $j \leq K$, there is no possibility of ever returning to one of the states $K+1, K+2, \dots, M$. A Markov chain that is not reducible is called *irreducible*. For example, for a two-state $(M = 2)$ Markov chain, the transpose of the transition matrix is

$$\mathbf{P}^{'} = \begin{pmatrix} p_{11} & 1-p_{22} \\ 1-p_{11} & p_{22} \end{pmatrix}$$

hence this process is irreducible if $p_{11} < 1$ and $p_{22} < 1$.

1.5.5 Ergodic Markov chains

Since every row of $\mathbf{P}$ sums to unity, we get

$$\mathbf{Pi} = \mathbf{i}$$

where $\mathbf{i} = (1, 1, \ldots, 1)^{'}$. Then unity is an eigenvalue of $\mathbf{P}$ (and hence, of $\mathbf{P}^{'}$) and the $(M \times 1)$ vector $\mathbf{i}$ is the associated eigenvector. If one of the eigenvalues of $\mathbf{P}$ is unity and all other eigenvalues of it are inside the unit circle, then the Markov chain is said to be *ergodic*. The $(M \times 1)$ vector of *ergodic probabilities* for an ergodic Markov chain is denoted by $\boldsymbol{\pi}$. This vector satisfies $\mathbf{P}^{'}\boldsymbol{\pi} = \boldsymbol{\pi}$, and it is normalized so that its elements sum to unity, i.e., $\mathbf{i}^{'}\boldsymbol{\pi} = 1$. If $\mathbf{P}$ is the transition matrix for an ergodic Markov chain, then

$$P^{'}{}_{\infty} = lim_{n\to+\infty}(\mathbf{P}^{'})^{n} = \boldsymbol{\pi}\mathbf{i}^{'}$$

This implies that the long-run forecast for an ergodic Markov chain is independent of the current state since

$$E(\boldsymbol{\xi}_{t+n}|\boldsymbol{\xi}_{t}, \boldsymbol{\xi}_{t-1}, \cdots) = (\mathbf{P}^{'})^{n}\boldsymbol{\xi}_{t} \underset{p}{\longrightarrow} \boldsymbol{\pi}\mathbf{i}^{'}\boldsymbol{\xi}_{t} = \boldsymbol{\pi}$$

Here we use the fact $\mathbf{i}^{'}\boldsymbol{\xi}_{t} = 1$ regardless of the value of $\boldsymbol{\xi}_{t}$. The vector of ergodic probabilities can also be viewed as indicating the unconditional probability of each of the M different states. To see this, suppose that we had used the symbol π_{j} to indicate the unconditional probability $P(s_{t} = j) = \pi_{j}$. Then the vector $\boldsymbol{\pi} = (\pi_{1}, \pi_{2}, \ldots, \pi_{M})^{'}$ could be described as the unconditional expectation of $\boldsymbol{\xi}_{t}$, that is, $\boldsymbol{\pi} = E(\boldsymbol{\xi}_{t})$. Taking unconditional expectations of both sides of (10) and assuming stationarity, we get

$$E(\boldsymbol{\xi}_{t+1}) = \mathbf{P}' E(\boldsymbol{\xi}_t)$$

hence

$$\boldsymbol{\pi} = \mathbf{P}' \boldsymbol{\pi}$$

which characterizes again $\boldsymbol{\pi}$ as the eigenvector of $\mathbf{P}'$ associated with the unit eigenvalue. For an ergodic Markov chain, this eigenvector is unique, and so the vector $\boldsymbol{\pi}$ of ergodic probabilities can be interpreted as the vector of unconditional probabilities. Finally, notice that an ergodic Markov chain is a covariance-stationary process. To determine explicitly a vector $\boldsymbol{\pi}$ satisfying $\mathbf{P}'\boldsymbol{\pi} = \boldsymbol{\pi}$ and $\mathbf{i}'\boldsymbol{\pi} = 1$, we seek a vector $\boldsymbol{\pi}$ such that

$$\mathbf{A}\boldsymbol{\pi} = \mathbf{e}_{M+1} \tag{12}$$

where $\mathbf{e}_{M+1}$ denotes the $(M+1)$th column of $\mathbf{I}_{M+1}$ and $\mathbf{A}$ is the $(M+1) \times M$ matrix given by the block form

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_M - \mathbf{P}' \\ \mathbf{i}' \end{pmatrix}$$

Such a solution can be found by premultiplying (12) by $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$, that is,

$$\boldsymbol{\pi} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{e}_{M+1}$$

In other words, $\boldsymbol{\pi}$ is the $(M+1)$th column of the matrix $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$.

1.5.6 Periodic Markov chains

It is possible to show that for any irreducible M-state Markov chain, all the eigenvalues of the transition matrix will be on or inside the unit circle. If there are K eigenvalues strictly on the unit circle with $K > 1$, then the chain is said to be *periodic* with period K. Such chains have the property that the states can be classified into K distinct classes, such that if the state at date t is from class α, then the state at date $t+1$ is certain to be from class $\alpha+1$. Thus, there is zero probability of returning to the original state s_t, and indeed zero probability of returning to any member of the original class α, except at horizons that are integer multiples of the period, such as dates $t+K$, $t+2K$, $t+3K$, and so on.

1.6 Time Series Models of Changes in Regime

1.6.1 The model

Here we consider time series models in which the parameters can change as a result of a regime-shift variable, described as the outcome of an unobserved Markov chain. Let $\mathbf{y}_t$ be an $(K \times 1)$ vector of observed endogenous variables and $\mathbf{x}_t$ a $(R \times 1)$ vector of explanatory observed variables. Let $\mathbf{Y}_t$ denote a vector containing all observations obtained through date t, i.e., $\mathbf{Y}_t = (\mathbf{y}_t^{'}, \mathbf{y}_{t-1}^{'}, \ldots, \mathbf{y}_{t-s}^{'}, \mathbf{x}_t^{'}, \mathbf{x}_{t-1}^{'}, \ldots, \mathbf{x}_{t-s}^{'})^{'}$ for some natural number s. If the process is governed by regime $s_t = j$ at date t, then the *conditional density* of $\mathbf{y}_t$ is assumed to be given by

$$f(\mathbf{y}_t | s_t = j, \mathbf{Y}_{t-1}; \boldsymbol{\alpha}) \tag{13}$$

where $\boldsymbol{\alpha}$ is a vector of parameters characterizing the conditional density. If there are M different regimes $s_t \in \{1, \ldots, M\}$, then there are M different densities represented by (13) for $j = 1, \ldots, M$, which will be collected in an $M \times 1$ vector denoted by $\boldsymbol{\eta}_t$, i.e.,

$$\boldsymbol{\eta}_t = \begin{pmatrix} f(\mathbf{y}_t | s_t = 1, \mathbf{Y}_{t-1}; \boldsymbol{\alpha}) \\ \vdots \\ f(\mathbf{y}_t | s_t = M, \mathbf{Y}_{t-1}; \boldsymbol{\alpha}) \end{pmatrix}$$

In particular, we have

$$\begin{aligned} f(\mathbf{y}_t | s_{t-1} = i, \mathbf{Y}_{t-1}; \boldsymbol{\alpha}) &= \sum_{j=1}^{M} f(\mathbf{y}_t | s_t = j, \mathbf{Y}_{t-1}; \boldsymbol{\alpha}) P(s_t = j | s_{t-1} = i) \\ &= \sum_{j=1}^{M} p_{ij} \eta_{jt} \end{aligned}$$

$$= \boldsymbol{\eta}_t^{'} \begin{pmatrix} p_{i1} \\ p_{i2} \\ \vdots \\ p_{iM} \end{pmatrix}$$

$$= \boldsymbol{\eta}_t^{'} \mathbf{P}^{'} \boldsymbol{\xi}_t$$

$$= \boldsymbol{\eta}_t^{'} E(\boldsymbol{\xi}_{t+1} | s_t = i)$$

As an illustration, consider a first-order autoregression in which both the constant term and the autoregressive coefficient might be different for different subsamples

$$y_t = c_{s_t} + \phi_{s_t} y_{t-1} + \epsilon_t$$

where $\epsilon_t \sim i.i.d.N(0, \sigma^2)$. Then y_t is a scalar ($K = 1$), the exogenous variables consist only of a constant term ($x_t = 1$), and the unknown parameters in $\boldsymbol{\alpha}$ consist of $c_1, \ldots, c_M, \phi_1, \ldots, \phi_M$, and σ^2. The vector of conditional densities is given by

$$\boldsymbol{\eta}_t = \begin{pmatrix} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y_t - c_1 - \phi_1 y_{t-1})^2}{2\sigma^2}) \\ \vdots \\ \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y_t - c_M - \phi_M y_{t-1})^2}{2\sigma^2}) \end{pmatrix}$$

Assumption A1. The conditional density in (13) is assumed to depend only on the current regime s_t and not on past regimes, i.e.,

$$f(\mathbf{y}_t | \mathbf{x}_t, \mathbf{Y}_{t-1}, s_t = j; \boldsymbol{\alpha}) = f(\mathbf{y}_t | \mathbf{x}_t, \mathbf{Y}_{t-1}, s_t = j, s_{t-1} = i, s_{t-2} = k, \ldots; \boldsymbol{\alpha})$$

though this is not really restrictive (see Hamilton [4], Ch.22).

Assumption A2. The random variable s_t evolves according a Markov chain that is independent of past observations on $\mathbf{y}_t$, or current or past $\mathbf{x}_t$, i.e.,

$$P(s_t = j | s_{t-1} = i, s_{t-2} = k, \ldots, \mathbf{x}_t, \mathbf{Y}_{t-1}) = P(s_t = j | s_{t-1} = i) = p_{ij}$$

Let $\mathbf{p} = (p_{1,1}, p_{1,2}, \ldots, p_{M,M})'$ denote the $(M^2 \times 1)$ vector of Markov transition probabilities. The population parameters that describe a time series governed by A1 and A2 consist of $\boldsymbol{\alpha}$ and $\mathbf{p}$. We collect the unknown parameters to be estimated in a single vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \mathbf{p}')'$. One important objective is to maximize the likelihood function of the observed data $f(\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_1; \boldsymbol{\theta})$ by choice of the population parameters vector $\boldsymbol{\theta}$.

1.6.2 Optimal Inference for the Regime

Another objective will be to estimate the value of $\boldsymbol{\theta}$ based on observation of Y_T. Let us nevertheless put this objective on hold for the moment and suppose that the value of $\boldsymbol{\theta}$ is somehow known with certainty to the analyst. Even if we know the value of $\boldsymbol{\theta}$, we will not know which regime the process was in at every date in the sample. Instead the best we can do is to provide inference for $\boldsymbol{\xi}_t$ given a specified observation set $\mathbf{Y}_\tau$, $\tau \leq T$. The statistical tools are the *filter* and *smoother* recursions which reconstruct the time path of the regime $(\boldsymbol{\xi}_t)$ under alternative information sets. Let $P(s_t = j|\mathbf{Y}_t; \boldsymbol{\theta})$ denote the analyst's inference about the value of s_t based on data obtained through date t and based on knowledge of the population parameter $\boldsymbol{\theta}$. This inference takes the form of a conditional probability that the analyst assigns to the probability that the tth observation was generated by regime j. Collect these conditional probabilities $P(s_t = j|\mathbf{Y}_\tau; \boldsymbol{\theta})$ for $j = 1, \ldots, M$ in an $(M \times 1)$ vector denoted by

$$\widehat{\boldsymbol{\xi}}_{t|\tau} = \begin{pmatrix} P(s_t = 1|\mathbf{Y}_\tau; \boldsymbol{\theta}) \\ \vdots \\ P(s_t = M|\mathbf{Y}_\tau; \boldsymbol{\theta}) \end{pmatrix} \tag{14}$$

which allows two different interpretations. First, $\widehat{\boldsymbol{\xi}}_{t|\tau}$ denotes the discrete conditional probability distribution of $\boldsymbol{\xi}_t$ given $\mathbf{Y}_\tau$. Secondly, $\widehat{\boldsymbol{\xi}}_{t|\tau}$ is equivalent to the conditional mean of $\boldsymbol{\xi}_t$ given $\mathbf{Y}_\tau$. This is due to the binarity of the elements of $\boldsymbol{\xi}_t$ which implies that

$$E(\boldsymbol{\xi}_{jt}) = P(\boldsymbol{\xi}_t = \mathbf{e}_j) = P(s_t = j) = \pi_j$$

for every $j = 1, \ldots, M$, i.e., $E(\boldsymbol{\xi}_t) = \boldsymbol{\pi}$ and

$$\widehat{\boldsymbol{\xi}}_{t|\tau} = E(\boldsymbol{\xi}_t|\mathbf{Y}_\tau;\boldsymbol{\theta}) \tag{15}$$

where $\mathbf{e}_j$ is the jth column of the identity matrix I_M.

Theorem 6.1 *(Hamilton [4], Chp. 22) The optimal inference and forecast for each date t in the sample can be found by iterating on the following pair of recursive formulae*

$$\begin{aligned} \widehat{\boldsymbol{\xi}}_{t|t} &= \frac{\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t}{\mathbf{i}'(\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t)} \\ \widehat{\boldsymbol{\xi}}_{t+1|t} &= \mathbf{P}'\widehat{\boldsymbol{\xi}}_{t|t} \end{aligned} \tag{16}$$

Here $\boldsymbol{\eta}_t$ is the $(M \times 1)$ vector defined in (1.6.1), $\mathbf{P}$ is the $(M \times M)$ transition matrix, $\mathbf{i}$ represents the $(M \times 1)$ vector of 1s, and the symbol $\odot$ denotes the element-by-element multiplication. Furthermore, the conditional probability density of $\mathbf{y}_t$ based upon $\mathbf{Y}_{t-1}$ is given by

$$f(\mathbf{y}_t|\mathbf{Y}_{t-1};\boldsymbol{\alpha}) = \mathbf{i}'(\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t) \tag{17}$$

Remark. Given a starting value $\widehat{\boldsymbol{\xi}}_{1|0} = E(\boldsymbol{\xi}_1|\mathbf{Y}_0;\boldsymbol{\theta})$ and an assumed value for the population parameter vector $\boldsymbol{\theta}$, one can iterate on (16) for $t = 1, \ldots, T$ to calculate the values of $\widehat{\boldsymbol{\xi}}_{t|t}$ and $\widehat{\boldsymbol{\xi}}_{t+1|t}$ for each date t in the sample. This gives the filtered regime probabilities $\widehat{\boldsymbol{\xi}}_{t|\tau}$, $t = \tau$ (*filtering*) and the predicted regime probabilities $\widehat{\boldsymbol{\xi}}_{t|\tau}$, $\tau < t$ (*forecasting*).

Proof of Theorem 6.1. The jth element of $\widehat{\boldsymbol{\xi}}_{t|t-1} = E(\boldsymbol{\xi}_t|\mathbf{Y}_{t-1};\boldsymbol{\theta})$ can also be described as

$$\widehat{\boldsymbol{\xi}}_{j,t|t-1} = E(\boldsymbol{\xi}_{jt}|\mathbf{Y}_{t-1};\boldsymbol{\theta}) = P(s_t = j|\mathbf{x}_t, \mathbf{Y}_{t-1};\boldsymbol{\theta})$$

This follows since we have assumed that $\mathbf{x}_t$ contains no information about s_t beyond that contained in $\mathbf{Y}_{t-1}$. The jth element of $\boldsymbol{\eta}_t$ is

$$f(\mathbf{y}_t|s_t = j, \mathbf{x}_t, \mathbf{Y}_{t-1};\boldsymbol{\theta})$$

The jth element of the $(M \times 1)$ vector $\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t$ is the product of these two magnitudes, which can be interpreted as the conditional joint density-distribution of $\mathbf{y}_t$ and $\mathbf{x}_t$:

$$P(s_t = j|\mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) f(\mathbf{y}_t|s_t = j, \mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) = P(\mathbf{y}_t, s_t = j|\mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) \tag{18}$$

The density of the observed vector $\mathbf{y}_t$, conditioned on past observables, is the sum of the M magnitudes in (18) for $j = 1, \ldots, M$. This sum can be written in vector notation as

$$\begin{aligned} f(\mathbf{y}_t|\mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) &= \sum_{j=1}^{M} P(\mathbf{y}_t, s_t = j|\mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) \\ &= \sum_{j=1}^{M} f(\mathbf{y}_t|s_t = j, \mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) P(s_t = j|\mathbf{Y}_{t-1}; \boldsymbol{\theta}) \\ &= \sum_{j=1}^{M} \widehat{\xi}_{j,t|t-1} \eta_{jt} \\ &= \boldsymbol{\eta}_t' \widehat{\boldsymbol{\xi}}_{t|t-1} \\ &= \mathbf{i}' (\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t) \end{aligned} \tag{19}$$

which proves the last formula in the statement of Theorem 6.1. If the joint density-distribution in (18) is divided by the density of $\mathbf{y}_t$ in (19), the result is the conditional distribution of s_t:

$$\frac{P(\mathbf{y}_t, s_t = j|\mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta})}{f(\mathbf{y}_t|\mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta})} = P(s_t = j|\mathbf{x}_t, \mathbf{y}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) = P(s_t = j|\mathbf{Y}_t; \boldsymbol{\theta})$$

Hence from (19)

$$P(s_t = j|\mathbf{Y}_t; \boldsymbol{\theta}) = \frac{P(\mathbf{y}_t, s_t = j|\mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta})}{\mathbf{i}' (\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t)} \tag{20}$$

But from (18) the numerator in the expression on the right side of (20) is the jth element of the vector $\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t$, while the left side of (20) is the jth element of the vector $\widehat{\boldsymbol{\xi}}_{t|t}$. Thus, collecting the equations in (20) for $j = 1, \ldots, M$ into an $(M \times 1)$ vector produces

$$\widehat{\boldsymbol{\xi}}_{t|t} = \frac{\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t}{\mathbf{i}' (\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t)}$$

as claimed in (16). To prove the second relation in (16), take expectations of $\boldsymbol{\xi}_{t+1} = \mathbf{P}^{'}\boldsymbol{\xi}_t + \mathbf{v}_{t+1}$ conditional on $\mathbf{Y}_t$, that is,

$$E(\boldsymbol{\xi}_{t+1}|\mathbf{Y}_t) = \mathbf{P}^{'}E(\boldsymbol{\xi}_t|\mathbf{Y}_t) + E(\mathbf{v}_{t+1}|\mathbf{Y}_t)$$

Note that $\mathbf{v}_{t+1}$ is a martingale difference sequence with respect to $\mathbf{Y}_t$, so the last formula becomes

$$\widehat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{P}^{'}\widehat{\boldsymbol{\xi}}_{t|t}$$

as $E(\mathbf{v}_{t+1}|\mathbf{Y}_t) = \mathbf{0}$ (use also formula (15)). $\square$

Remark (Starting the algorithm). Given a starting value $\widehat{\boldsymbol{\xi}}_{1|0}$ one can use (16) to calculate $\widehat{\boldsymbol{\xi}}_{t|t}$ for any t. Several options are available for choosing the starting value. One approach is to set $\widehat{\boldsymbol{\xi}}_{1|0}$ equal to the vector $\boldsymbol{\pi}$ of unconditional probabilities. Another option is to set $\widehat{\boldsymbol{\xi}}_{1|0} = \boldsymbol{\rho}$, where $\boldsymbol{\rho}$ is a fixed $(M \times 1)$ vector of nonnegative constants summing to unity, such as $\boldsymbol{\rho} = M^{-1}\mathbf{i}$. Alternatively, $\boldsymbol{\rho}$ could be estimated by maximum likelihood along with $\boldsymbol{\theta}$ subject to the constraint that $\mathbf{i}^{'}\boldsymbol{\rho} = 1$ and $\rho_j \geq 0$ for $j = 1, \ldots, M$.

1.6.3 Forecasts and Smoothed Inferences for the Regime

The $(M \times 1)$ vector $\widehat{\boldsymbol{\xi}}_{t|\tau}$ for $t > \tau$ represents a *forecast* about the regime for some future period t, whereas for $t < \tau$ it represents the smoothed inference about the regime as noted above.

Theorem 6.2 *(Hamilton [4], Chp. 22) (i) The optimal h-period-ahead forecast of* $\boldsymbol{\xi}_{t+h}$ *is given by*

$$\widehat{\boldsymbol{\xi}}_{t+h|t} = (\mathbf{P}^{'})^h\widehat{\boldsymbol{\xi}}_{t|t} \tag{21}$$

where $\widehat{\boldsymbol{\xi}}_{t|t}$ *is calculated from (16).*

(ii) Smoothed inferences can be calculated using an algorithm which can be written, in vector form, as

$$\widehat{\boldsymbol{\xi}}_{t|T} = \widehat{\boldsymbol{\xi}}_{t|t} \odot \{\mathbf{P}[\widehat{\boldsymbol{\xi}}_{t+1|T}(\div)\widehat{\boldsymbol{\xi}}_{t+1|t}]\} \tag{22}$$

where the symbol $(\div)$ *denotes element-by-element division.*

Remark. The smoothed probabilities $\widehat{\boldsymbol{\xi}}_{t|T}$ are found by iterating (22) backward for $t = T-1, T-2, \ldots, 1$. This iteration is started with $\widehat{\boldsymbol{\xi}}_{T|T}$ which is obtained from (16) for $t = T$. This algorithm is valid only when s_t follows a first-order Markov chain as in A2, when the conditional density (13) depends on $s_t, s_{t-1}, \ldots$ only through the current state s_t, and when $\mathbf{x}_t$ (the vector of explanatory variables other than the lagged values of $\mathbf{y}$) is strictly exogenous, meaning that $\mathbf{x}_t$ is independent of s_τ for all t and τ.

Proof of Theorem 6.2. (i) follows by taking expectation of both sides of

$$\boldsymbol{\xi}_{t+h} = \mathbf{v}_{t+h} + \mathbf{P}^{'}\mathbf{v}_{t+h-1} + \cdots + (\mathbf{P}^{'})^{h-1}\mathbf{v}_{t+1} + (\mathbf{P}^{'})^{h}\boldsymbol{\xi}_t$$

conditional on information available at date t:

$$E(\boldsymbol{\xi}_{t+h}|\mathbf{Y}_t) = (\mathbf{P}^{'})^{h}E(\boldsymbol{\xi}_t|\mathbf{Y}_t)$$

or

$$\widehat{\boldsymbol{\xi}}_{t+h|t} = (\mathbf{P}^{'})^{h}\widehat{\boldsymbol{\xi}}_{t|t}$$

(ii) Recall first that under the maintained assumptions, the regime s_t depends on past observations $\mathbf{Y}_{t-1}$ only through the value of s_{t-1}. Similarly, s_t depends on future observations only through the value s_{t+1}:

$$P(s_t = j|s_{t+1} = i, \mathbf{Y}_T; \boldsymbol{\theta}) = P(s_t = j|s_{t+1} = i; \mathbf{Y}_t; \boldsymbol{\theta}) \tag{23}$$

The validity of (23) is formally established as follows (the implicit dependence on $\boldsymbol{\theta}$ will be suppressed to simplify notation). Observe that

$$\begin{aligned} P(s_t = j|s_{t+1} = i, \mathbf{Y}_{t+1}) &= P(s_t = j|s_{t+1} = i, \mathbf{y}_{t+1}, \mathbf{x}_{t+1}, \mathbf{Y}_t) \\ &= \frac{P(\mathbf{y}_{t+1}, s_t = j|s_{t+1} = i, \mathbf{x}_{t+1}, \mathbf{Y}_t)}{f(\mathbf{y}_{t+1}|s_{t+1} = i, \mathbf{x}_{t+1}, \mathbf{Y}_t)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{f(\mathbf{y}_{t+1}|s_t = j, s_{t+1} = i, \mathbf{x}_{t+1}, \mathbf{Y}_t) P(s_t = j|s_{t+1} = i, \mathbf{x}_{t+1}, \mathbf{Y}_t)}{f(\mathbf{y}_{t+1}|s_{t+1} = i, \mathbf{x}_{t+1}, \mathbf{Y}_t)} \\
&= P(s_t = j|s_{t+1} = i, \mathbf{x}_{t+1}, \mathbf{Y}_t)
\end{aligned} \tag{24}
$$

as $f(\mathbf{y}_{t+1}|s_t = j, s_{t+1} = i, \mathbf{x}_{t+1}, \mathbf{Y}_t) = f(\mathbf{y}_{t+1}|s_{t+1} = i, \mathbf{x}_{t+1}, \mathbf{Y}_t)$. In fact, $\mathbf{y}_{t+1}$ depends on $s_{t+1}, s_t, \ldots$ only through the current value s_{t+1}. Since $\mathbf{x}$ is exogenous, (24) further implies that

$$
P(s_t = j|s_{t+1} = i, \mathbf{Y}_{t+1}) = P(s_t = j|s_{t+1} = i, \mathbf{Y}_t)
$$

By induction, the same argument gives

$$
P(s_t = j|s_{t+1} = i, \mathbf{Y}_{t+h}) = P(s_t = j|s_{t+1} = i, \mathbf{Y}_t)
$$

for $h = 1, 2, \ldots$ from which (23) follows. Next note that

$$
\begin{aligned}
P(s_t = j|s_{t+1} = i, \mathbf{Y}_t) &= \frac{P(s_t = j, s_{t+1} = i|\mathbf{Y}_t)}{P(s_{t+1} = i|\mathbf{Y}_t)} \\
&= \frac{P(s_t = j|\mathbf{Y}_t) P(s_{t+1} = i|s_t = j)}{P(s_{t+1} = i|\mathbf{Y}_t)} \\
&= \frac{p_{ji} P(s_t = j|\mathbf{Y}_t)}{P(s_{t+1} = i|\mathbf{Y}_t)}
\end{aligned} \tag{25}
$$

Therefore we obtain

$$
\begin{aligned}
P(s_t = j, s_{t+1} = i|\mathbf{Y}_T) &= P(s_{t+1} = i|\mathbf{Y}_T) P(s_t = j|s_{t+1} = i, \mathbf{Y}_T) \\
&= P(s_{t+1} = i|\mathbf{Y}_T) P(s_t = j|s_{t+1} = i, \mathbf{Y}_t) \\
&= P(s_{t+1} = i|\mathbf{Y}_T) \frac{p_{ji}\, P(s_t = j|\mathbf{Y}_t)}{P(s_{t+1} = i|\mathbf{Y}_t)}
\end{aligned} \tag{26}
$$

where the second equality follows from (23) and the third follows from (25). The smoothed inference for date t is the sum of (26) over $i = 1, \ldots, M$.

$$
\begin{aligned}
\widehat{\xi}_{j,t|T} = P(s_t = j|\mathbf{Y}_T) &= \sum_{i=1}^{M} P(s_t = j, s_{t+1} = i|\mathbf{Y}_T) \\
&= \sum_{i=1}^{M} P(s_{t+1} = i|\mathbf{Y}_T) \frac{p_{ji} P(s_t = j|\mathbf{Y}_t)}{P(s_{t+1} = i|\mathbf{Y}_t)} \\
&= P(s_t = j|\mathbf{Y}_t) \sum_{i=1}^{M} \frac{p_{ji} P(s_{t+1} = i|\mathbf{Y}_T)}{P(s_{t+1} = i|\mathbf{Y}_t)}
\end{aligned}
$$

$$
\begin{aligned}
&= P(s_t = j|\mathbf{Y}_t)(p_{j1}\, p_{j2}\, \ldots p_{jM}) \begin{pmatrix} \frac{P(s_{t+1}=1|\mathbf{Y}_T)}{P(s_{t+1}=1|\mathbf{Y}_t)} \\ \vdots \\ \frac{P(s_{t+1}=M|\mathbf{Y}_T)}{P(s_{t+1}=M|\mathbf{Y}_t)} \end{pmatrix} \\
&= P(s_t = j|\mathbf{Y}_t)\mathbf{p}_j(\widehat{\boldsymbol{\xi}}_{t+1|T}(\div)\widehat{\boldsymbol{\xi}}_{t+1|t}) \\
&= \widehat{\boldsymbol{\xi}}_{j,t|t}\mathbf{p}_j(\widehat{\boldsymbol{\xi}}_{t+1|T}(\div)\widehat{\boldsymbol{\xi}}_{t+1|t})
\end{aligned}
$$

where the $(1 \times M)$ vector $\mathbf{p}_j$ denotes the jth row of the matrix $\mathbf{P}$ and the symbol $(\div)$ indicates element-by-element division. When the equations

$$\widehat{\xi}_{j,t|T} = \widehat{\xi}_{j,t|t}\mathbf{p}_j(\widehat{\boldsymbol{\xi}}_{t+1|T}(\div)\widehat{\boldsymbol{\xi}}_{t+1|t})$$

for $j = 1, \ldots, M$ are collected in an $(M \times 1)$ vector, the result is formula (22) as claimed. □

Assuming presample value $\mathbf{Y}_0$ is given, the density of the sample $\mathbf{Y} = \mathbf{Y}_T$ for given state $\boldsymbol{\xi}$ is determined by

$$f(\mathbf{Y}|\boldsymbol{\xi}) = \prod_{t=1}^{T} f(\mathbf{y}_t|\boldsymbol{\xi}_t, \mathbf{Y}_{t-1}).$$

Hence the joint probability distribution of observations and states can be calculated as

$$
\begin{aligned}
P(\mathbf{Y}, \boldsymbol{\xi}) &= f(\mathbf{Y}|\boldsymbol{\xi})P(\boldsymbol{\xi}) \\
&= \prod_{t=1}^{T} f(\mathbf{y}_t|\boldsymbol{\xi}_t, \mathbf{Y}_{t-1}) \prod_{t=2}^{T} P(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1})P(\boldsymbol{\xi}_{t-1})
\end{aligned}
$$

Thus the unconditional density of $\mathbf{Y}$ is given by the marginal density

(27) $$f(\mathbf{Y}) = \int P(\mathbf{Y}, \boldsymbol{\xi})\, d\boldsymbol{\xi}$$

where

$$\int P(\mathbf{x}, \boldsymbol{\xi})\, d\boldsymbol{\xi} = \sum_{i_1=1}^{M} \cdots \sum_{i_T=1}^{M} P(\mathbf{x}, \boldsymbol{\xi}_T = \mathbf{e}_{i_T}, \dots, \boldsymbol{\xi}_1 = \mathbf{e}_{i_1})$$
$$= \sum_{i_1=1}^{M} \cdots \sum_{i_T=1}^{M} P(\mathbf{x}, s_T = i_T, \dots, s_1 = i_1)$$

denotes the summation over all possible values of $\boldsymbol{\xi} = \boldsymbol{\xi}_T \otimes \boldsymbol{\xi}_{T-1} \otimes \cdots \otimes \boldsymbol{\xi}_1$ in equation (27). Setting $P(s_t = j; \boldsymbol{\theta}) = \pi_j$, $j = 1, \dots, M$, the joint density-distribution of $\mathbf{y}_t$ and s_t is given by

$$P(\mathbf{y}_t, s_t = j; \boldsymbol{\theta}) = f(\mathbf{y}_t | s_t = j; \boldsymbol{\theta}) P(s_t = j; \boldsymbol{\theta}) = \pi_j f(\mathbf{y}_t | s_t = j; \boldsymbol{\theta}) \tag{28}$$

The unconditional density of $\mathbf{y}_t$ can be found by summing up (28) over all possible states

$$f(\mathbf{y}_t; \boldsymbol{\theta}) = \sum_{j=1}^{M} \pi_j f(\mathbf{y}_t | s_t = j; \boldsymbol{\theta})$$
$$= (f(\mathbf{y}_t | s_1 = 1; \boldsymbol{\theta}) \dots f(\mathbf{y}_t | s_t = M; \boldsymbol{\theta}))\boldsymbol{\pi}$$

Finally, it follows by the definition of the conditional density that the conditional distribution of the total regime $\boldsymbol{\xi}$ is given by

$$P(\boldsymbol{\xi} | \mathbf{Y}) = \frac{P(\mathbf{Y}, \boldsymbol{\xi})}{f(\mathbf{Y})}$$

Thus the desired conditional regime probabilities $P(\boldsymbol{\xi}_t | \mathbf{Y})$ can be derived by marginalization of $P(\boldsymbol{\xi} | \mathbf{Y})$.

1.6.4 Forecasts for the Observed Variables

From the conditional density (13), it is easy to forecast $\mathbf{y}_{t+1}$ conditional on knowing $\mathbf{Y}_t, \mathbf{x}_{t+1}$ and s_{t+1}. For example, for the $AR(1)$ specification $y_{t+1} = c_{s_{t+1}} + \phi_{s_{t+1}} y_t + \epsilon_{t+1}$, where $\epsilon_{t+1} \sim NID(0, \sigma^2)$, such a forecast is given by

$$E(y_{t+1} | s_{t+1} = j, \mathbf{Y}_t; \boldsymbol{\theta}) = c_j + \phi_j y_t.$$

There are M different conditional forecasts associated with the M possible values for s_{t+1}. Note that the unconditional forecast based on all actual observable variables is related to these conditional forecasts by

$$E(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta}) = \int \mathbf{y}_{t+1} f(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta})\, d\mathbf{y}_{t+1}$$
$$= \int \mathbf{y}_{t+1} \{\sum_{j=1}^{M} P(\mathbf{y}_{t+1}, s_{t+1} = j|\mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta})\}\, d\mathbf{y}_{t+1}$$

$$= \int \mathbf{y}_{t+1} \{\sum_{j=1}^{M} [f(\mathbf{y}_{t+1}|s_{t+1} = j, \mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta}) P(s_{t+1} = j|\mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta})]\}\, d\mathbf{y}_{t+1}$$
$$= \sum_{j=1}^{M} P(s_{t+1} = j|\mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta}) \int \mathbf{y}_{t+1} f(\mathbf{y}_{t+1}|s_{t+1} = j, \mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta})\, d\mathbf{y}_{t+1}$$
$$= \sum_{j=1}^{M} P(s_{t+1} = j|\mathbf{Y}_t; \boldsymbol{\theta}) E(\mathbf{y}_{t+1}|s_{t+1} = j, \mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta})$$
$$= \mathbf{h}_t^{'} \widehat{\boldsymbol{\xi}}_{t+1|t}$$

where $\mathbf{h}_t$ is the $(M \times 1)$ vector whose jth element is $E(\mathbf{y}_{t+1}|s_{t+1} = j, \mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta})$. Thus

$$E(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \mathbf{Y}_t; \boldsymbol{\theta}) = \mathbf{h}_t^{'} \widehat{\boldsymbol{\xi}}_{t+1|t}$$

Note that although the Markov chain itself admits the linear representation (12), the *optimal forecast* of $\mathbf{y}_{t+1}$ is a non linear function of observables, since the inference $\widehat{\boldsymbol{\xi}}_{t+1|t}$ in (16) depends non linearly on $\mathbf{Y}_t$.

1.6.5 Maximum Likelihood Estimation of Parameters

In the iteration on (16) the parameter vector $\boldsymbol{\theta}$ was taken to be a fixed known vector. Once the iteration has been completed for $t = 1, \ldots, T$ for a given fixed $\boldsymbol{\theta}$, the value of the log likelihood implied by that value of $\boldsymbol{\theta}$ is then known as follows

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} log\, f(\mathbf{y}_t|\mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta})$$

where

$$f(\mathbf{y}_t|\mathbf{x}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) = \mathbf{i}^{'} (\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t) = \boldsymbol{\eta}_t^{'} \mathbf{P}^{'} \widehat{\boldsymbol{\xi}}_{t-1|t-1}.$$

If the transition probabilities are restricted only by the conditions that $p_{ij} \geq 0$ and $\sum_{j=1}^{M} p_{ij} = 1$ for all i and j, and if the initial probability $\widehat{\boldsymbol{\xi}}_{1|0}$ is taken to be a fixed value $\boldsymbol{\rho}$ unrelated to the other parameters, then it is shown in Hamilton (1990) [2] that the estimated transition probabilities can be expressed as follows.

Theorem 6.3 *(Hamilton 1990 [2]) The maximum likelihood estimates for the transition probabilities satisfy*

$$\widehat{p}_{ij} = \frac{\sum_{t=2}^{T} P(s_t = j, s_{t-1} = i|\mathbf{Y}_T; \widehat{\boldsymbol{\theta}})}{\sum_{t=2}^{T} P(s_{t-1} = i|\mathbf{Y}_T; \widehat{\boldsymbol{\theta}})}$$

where $\widehat{\boldsymbol{\theta}}$ denotes the full vector of maximum likelihood estimates.

Thus the estimated transition probabilities $\widehat{p}_{ij}$ is essentially the number of times state i seems to have been followed by state j divided by the number of times the process was in state i. These counts are estimated on the basis of the smoothed probabilities. If the vector of initial probabilities $\boldsymbol{\rho}$ is regarded as a separate vector of parameters constrained only by $\mathbf{i}^{'}\boldsymbol{\rho} = 1$ and $\rho_j \geq 0$, for every j, the maximum likelihood estimate of $\boldsymbol{\rho}$ turns out to be the smoothed inference about the initial state $\widehat{\boldsymbol{\rho}} = \widehat{\boldsymbol{\xi}}_{1|T}$. The maximum likelihood estimate of the vector $\boldsymbol{\alpha}$ that governs the conditional density (13) is characterized by

$$\sum_{t=1}^{T} (\frac{\partial log \boldsymbol{\eta}_t}{\partial \boldsymbol{\alpha}'})' \, \widehat{\boldsymbol{\xi}}_{t|T} = \mathbf{0}$$

Here $\boldsymbol{\eta}_t$ is the $(M \times 1)$ vector in (14) and $\frac{\partial log \boldsymbol{\eta}_t}{\partial \boldsymbol{\alpha}'}$ is the $(M \times K)$ matrix of derivatives of the logs of the densities in $\boldsymbol{\eta}_t$, where k represents the number of parameters in $\boldsymbol{\alpha}$. We postpone the proofs of Theorem 6.3 and the last formula in the next section.

1.7 EM algorithm and Likelihood function

Following Hamilton (1990) [2], we illustrate an EM (*Expectation Maximization*) algorithm for obtaining maximum likelihood estimates of parameters for time series subject to discrete Markovian shifts. Our task is to maximize the likelihood function of the observed data

$$f(\mathbf{Y}; \boldsymbol{\theta}) = f(\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_1; \boldsymbol{\theta})$$

by choice of the population parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}' \ \mathbf{p}')'$. It usually proves computationally simplest in time series autoregressions to replace the exact likelihood function $f(\mathbf{Y};\boldsymbol{\theta})$ with the likelihood conditional on the first m observations $f(\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_{m+1} | \mathbf{y}_m, \mathbf{y}_{m-1}, \ldots, \mathbf{y}_1; \boldsymbol{\theta})$. The same turns out to be true for the EM algorithm illustrated here. To form this conditional likelihood, one also needs to make assumptions about the probability law governing the initial unobserved states. It turns out to be computationally simplest to assume that the initial states were drawn from a separate probability distribution, whose parameters are unrelated to $\boldsymbol{\theta}$:

$$\rho_{s_m, s_{m-1}, \ldots, s_1} = P(s_m, s_{m-1}, \ldots, s_1 | \mathbf{y}_m, \mathbf{y}_{m-1}, \ldots, \mathbf{y}_1) \tag{29}$$

These population probabilities are collected in a $(M^m \times 1)$ vector

$$\boldsymbol{\rho} = (\rho_{1,1,\ldots,1}, \rho_{1,1,\ldots,2}, \ldots, \rho_{M,M,\ldots,M})'.$$

The elements of $\boldsymbol{\rho}$ sum to unity and are to be estimated by maximum likelihood along with $\boldsymbol{\theta}$. We collect these parameters in a single vector $\boldsymbol{\lambda} = (\boldsymbol{\alpha}' \ \mathbf{p}' \ \boldsymbol{\rho}')' = (\boldsymbol{\theta}' \ \boldsymbol{\rho}')'$.

1.7.1 EM algorithm: general principles

Given a vector $\mathbf{Y} = (\mathbf{y}_T', \mathbf{y}_{T-1}', \ldots, \mathbf{y}_1')'$, our objective is to choose the vector $\boldsymbol{\lambda}$ so as to maximize the conditional likelihood

$$f(\mathbf{Y};\boldsymbol{\lambda}) = f(\mathbf{y}_T, \ldots, \mathbf{y}_{m+1} | \mathbf{y}_m, \ldots, \mathbf{y}_1; \boldsymbol{\lambda}).$$

We can most easily characterize the structure of this maximum likelihood estimation if we consider the hypothetical joint likelihood function for unobserved states (s_t) and observed data $(\mathbf{y}_t)$. Define the $(T \times 1)$ vector $\mathbf{S}$ to be the realization of the unobserved states for the entire sample, i.e., $\mathbf{S} = (s_T, s_{T-1}, \ldots, s_1)'$. Though $\mathbf{S}$ is not observed by the econometrician, it is easy to characterize what the joint distribution of $\mathbf{Y}$ and $\mathbf{S}$ would look like if $\mathbf{S}$ were observed:

$$P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda}) = P(\mathbf{y}_T, \ldots, \mathbf{y}_{m+1}, s_T, \ldots, s_1 | \mathbf{y}_m, \ldots, \mathbf{y}_1; \boldsymbol{\lambda}) \tag{30}$$

Notice that one does not need to calculate (30) in order to use the EM algorithm. Expression (30) is only considered as a theoretical construct for expositing what the EM algorithm is and why it works. Given this, we can regard the marginal likelihood function $f(\mathbf{Y};\boldsymbol{\lambda})$ as the summation of the joint likelihood $P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda})$ over all possible values of $\mathbf{S}$:

$$f(\mathbf{Y};\boldsymbol{\lambda}) = \int_{\mathbf{S}} P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}) \tag{31}$$

where the notation $\int_{\mathbf{S}}$ denotes summation over all possible values of all the elements of $\mathbf{S}$:

$$\int_{\mathbf{S}} g(\mathbf{S}) = \sum_{s_T=1}^{M} \sum_{s_{T-1}=1}^{M} \cdots \sum_{s_1=1}^{M} g(s_T, s_{T-1}, \ldots, s_1)$$

Again (31) is not the expression one would use to evaluate the actual likelihood, but is a representation of the sample likelihood in terms of the theoretical construct $P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda})$. Let $\mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_{\ell},\mathbf{Y})$ denote the expected log-likelihood, where the log-likelihood is parameterized by $\boldsymbol{\lambda}_{\ell+1}$ and the expectation is taken with respect to a second distribution parameterized by $\boldsymbol{\lambda}_{\ell}$, i.e.,

$$\mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_{\ell},\mathbf{Y}) = \int_{\mathbf{S}} [logP(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}_{\ell+1})]P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}_{\ell}) \tag{32}$$

There are two ways to characterize the EM algorithm for arriving at the MLE $\widehat{\boldsymbol{\lambda}}$.

The *first characterization* is based on a sequence of optimization problems (indexed by $\ell = 1, 2, \ldots$), each of whose analytic solution $\widehat{\boldsymbol{\lambda}}_{\ell}$ is found exactly. The solution $\widehat{\boldsymbol{\lambda}}_{\ell+1}$ to optimization problem $\ell + 1$ increases the value of the likelihood function relative to the value for $\widehat{\boldsymbol{\lambda}}_{\ell}$. The limit of this sequence of estimators achieves a local maximum of the likelihood function:

$$\lim_{\ell \to +\infty} \widehat{\boldsymbol{\lambda}}_{\ell} = \widehat{\boldsymbol{\lambda}}_{MLE}$$

This follows from Theorem 7.1.

An *alternative characterization* of the EM algorithm is as follows. Imagine that $\mathbf{S}$ were observed directly. The first-order conditions for calculating the MLE for $\boldsymbol{\lambda}$ would in this case be quite easy. These conditions (one for each possible realization of $\mathbf{S}$) can be weighted by the probability that the unobserved state variables indeed took on the particular values represented by $\mathbf{S}$. These probabilities in turn can be evaluated, using the previous iteration's estimate $\widehat{\boldsymbol{\lambda}}_\ell$ as $P(\mathbf{S}|\mathbf{Y};\widehat{\boldsymbol{\lambda}}_\ell)$. The sum of the weighted conditions over all possible states then characterizes the EM algorithm's choice for $\widehat{\boldsymbol{\lambda}}_{\ell+1}$. Thus the EM algorithm replaces the unobserved scores by their expectation given the previous iteration's estimated parameter vector. This follows from Theorem 7.2.

1.7.2 First characterization of EM algorithm

Let $\widehat{\boldsymbol{\lambda}}_\ell$ denote the estimate of the parameters vector resulting from our previous iteration, with $\widehat{\boldsymbol{\lambda}}_0$ an arbitrary initial guess. We choose for $\widehat{\boldsymbol{\lambda}}_{\ell+1}$ the value of $\boldsymbol{\lambda}_{\ell+1}$ that maximizes $\mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\widehat{\boldsymbol{\lambda}}_\ell,\mathbf{Y})$ given in (32), i.e., $\widehat{\boldsymbol{\lambda}}_{\ell+1}$ satisfies

$$\int_{\mathbf{S}} \frac{\partial log P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}_{\ell+1})}{\partial \boldsymbol{\lambda}_{\ell+1}}\bigg|_{\boldsymbol{\lambda}_{\ell+1}=\widehat{\boldsymbol{\lambda}}_{\ell+1}} P(\mathbf{Y},\mathbf{S};\widehat{\boldsymbol{\lambda}}_\ell)=\mathbf{0} \tag{33}$$

We show in Subsection (1.7.4) how (33) can be solved analytically for $\widehat{\boldsymbol{\lambda}}_{\ell+1}$ as a function of $\mathbf{Y}$ and $\widehat{\boldsymbol{\lambda}}_\ell$.

Theorem 7.1

i) The estimate $\widehat{\boldsymbol{\lambda}}_{\ell+1}$ is associated with a higher value of the likelihood function than is $\widehat{\boldsymbol{\lambda}}_\ell$, that is, $f(\mathbf{Y};\widehat{\boldsymbol{\lambda}}_{\ell+1})\geq f(\mathbf{Y};\widehat{\boldsymbol{\lambda}}_\ell)$ with equality only if $\widehat{\boldsymbol{\lambda}}_{\ell+1}=\widehat{\boldsymbol{\lambda}}_\ell$;

ii) If

$$\frac{\partial \mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\widehat{\boldsymbol{\lambda}}_\ell,\mathbf{Y})}{\partial \boldsymbol{\lambda}_{\ell+1}}\bigg|_{\boldsymbol{\lambda}_{\ell+1}=\widehat{\boldsymbol{\lambda}}_\ell}=\mathbf{0}\,,$$

then

$$\frac{\partial f(\mathbf{Y};\boldsymbol{\lambda}_{\ell+1})}{\partial \boldsymbol{\lambda}_{\ell+1}}\bigg|_{\boldsymbol{\lambda}_{\ell+1}=\widehat{\boldsymbol{\lambda}}_\ell}=\mathbf{0}.$$

Proof. i) By construction, $\widehat{\boldsymbol{\lambda}}_{\ell+1}$ maximizes $\mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\widehat{\boldsymbol{\lambda}}_\ell,\mathbf{Y})$, so in particular $\mathcal{Q}(\widehat{\boldsymbol{\lambda}}_{\ell+1};\widehat{\boldsymbol{\lambda}}_\ell,\mathbf{Y})\geq \mathcal{Q}(\widehat{\boldsymbol{\lambda}}_\ell;\widehat{\boldsymbol{\lambda}}_\ell,\mathbf{Y})$ with equality only if $\widehat{\boldsymbol{\lambda}}_{\ell+1}=\widehat{\boldsymbol{\lambda}}_\ell$. Recall that for any positive real number x, we have $log(x) \leq x - 1$ with equality only if $x = 1$. Then we have

$$\begin{aligned}
\mathcal{Q}(\widehat{\boldsymbol{\lambda}}_{\ell+1}; \widehat{\boldsymbol{\lambda}}_{\ell}, \mathbf{Y}) - \mathcal{Q}(\widehat{\boldsymbol{\lambda}}_{\ell}; \widehat{\boldsymbol{\lambda}}_{\ell}, \mathbf{Y}) &= \int_{\mathbf{S}} log[\frac{P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell+1})}{P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell})}] P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell}) \\
&\leq \int_{\mathbf{S}} [\frac{P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell+1})}{P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell})} - 1] P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell}) \\
&= \int_{\mathbf{S}} [P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell+1}) - P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell})] \\
&= f(\mathbf{Y}; \widehat{\boldsymbol{\lambda}}_{\ell+1}) - f(\mathbf{Y}; \widehat{\boldsymbol{\lambda}}_{\ell})
\end{aligned}$$

Thus, if $\mathcal{Q}(\widehat{\boldsymbol{\lambda}}_{\ell+1}; \widehat{\boldsymbol{\lambda}}_{\ell}, \mathbf{Y}) > \mathcal{Q}(\widehat{\boldsymbol{\lambda}}_{\ell}; \widehat{\boldsymbol{\lambda}}_{\ell}, \mathbf{Y})$, then $f(\mathbf{Y}; \widehat{\boldsymbol{\lambda}}_{\ell+1}) > f(\mathbf{Y}; \widehat{\boldsymbol{\lambda}}_{\ell})$. This proves (i).

ii) By definition of $\mathcal{Q}$, we have

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\boldsymbol{\lambda}_{\ell+1}; \widehat{\boldsymbol{\lambda}}_{\ell}, \mathbf{Y})}{\partial \boldsymbol{\lambda}_{\ell+1}}\Big|_{\boldsymbol{\lambda}_{\ell+1}=\widehat{\boldsymbol{\lambda}}_{\ell}} &= \int_{\mathbf{S}} \{\frac{1}{P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda}_{\ell+1})} \frac{\partial P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda}_{\ell+1})}{\partial \boldsymbol{\lambda}_{\ell+1}}\}_{\boldsymbol{\lambda}_{\ell+1}=\widehat{\boldsymbol{\lambda}}_{\ell}} P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell}) \\
&= \int_{\mathbf{S}} \frac{\partial P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda}_{\ell+1})}{\partial \boldsymbol{\lambda}_{\ell+1}}\Big|_{\boldsymbol{\lambda}_{\ell+1}=\widehat{\boldsymbol{\lambda}}_{\ell}} \\
&= \frac{\partial f(\mathbf{Y}; \boldsymbol{\lambda}_{\ell+1})}{\partial \boldsymbol{\lambda}_{\ell+1}}\Big|_{\boldsymbol{\lambda}_{\ell+1}=\widehat{\boldsymbol{\lambda}}_{\ell}}
\end{aligned}$$

Thus, if the left-hand-side is zero, so it must be the right-hand-side as well. □

Theorem 7.1 implies that the sequence $(\widehat{\boldsymbol{\lambda}}_{\ell})_{\ell=1}^{\infty}$ converges to the (local) MLE $\widehat{\boldsymbol{\lambda}}$.

1.7.3 Second characterization of EM algorithm

Suppose that the vector of regimes $\mathbf{S}$ were observed directly. Then the MLE $\widehat{\boldsymbol{\lambda}}(\mathbf{S})$ would be characterized by the first-order conditions

(34)
$$\frac{\partial log P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\Big|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}(\mathbf{S})} = \mathbf{0}$$

Now, though the econometrician does not have data directly on $\mathbf{S}$, after iteration ℓ we have an inference about $\mathbf{S}$ based on our parameter estimate $\widehat{\boldsymbol{\lambda}}_{\ell}$ and the observed data $\mathbf{Y}$:

(35)
$$P(\mathbf{S}|\mathbf{Y}; \widehat{\boldsymbol{\lambda}}_{\ell}) = \frac{P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_{\ell})}{f(\mathbf{Y}; \widehat{\boldsymbol{\lambda}}_{\ell})}$$

For each of the M^T possible values for $\mathbf{S}$, there is a corresponding particular first-order condition (34). If we weight each of these first-order conditions by probability (35) that $\mathbf{S}$ took on that particular value, we would be chosing $\boldsymbol{\lambda}$ so as to satisfy

$$\int_{\mathbf{S}} \frac{\partial log P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \frac{P(\mathbf{Y}, \mathbf{S}; \widehat{\boldsymbol{\lambda}}_\ell)}{f(\mathbf{Y}; \widehat{\boldsymbol{\lambda}}_\ell)} = \mathbf{0}$$

or equivalently

$$\frac{1}{f(\mathbf{Y}; \widehat{\boldsymbol{\lambda}}_\ell)} \frac{\partial \mathcal{Q}(\boldsymbol{\lambda}; \widehat{\boldsymbol{\lambda}}_\ell, \mathbf{Y})}{\partial \boldsymbol{\lambda}} = \mathbf{0}$$

that is

$$\frac{\partial \mathcal{Q}(\boldsymbol{\lambda}; \widehat{\boldsymbol{\lambda}}_\ell, \mathbf{Y})}{\partial \boldsymbol{\lambda}} = \mathbf{0}.$$

So the following result holds.

Theorem 7.2 *The estimate $\widehat{\boldsymbol{\lambda}}_{\ell+1}$ in Theorem 7.1 (i) coincides with the estimate that would result if we weighted the first-order conditions (33) associated with direct observation of $\mathbf{S}$ by the probability that $\mathbf{S}$ took on each of its feasible values.*

1.7.4 Explicit form of the EM algorithm

Theorem 7.3 *For a Markov switching time series satisfying Assumptions A1 and A2 in Subsection 1.6.1, the expected log-likelihood function $\mathcal{Q}(\boldsymbol{\lambda}_{\ell+1}; \boldsymbol{\lambda}_\ell, \mathbf{Y})$ is maximized by chosing $\boldsymbol{\lambda}_{\ell+1} = (\boldsymbol{\alpha}'_{\ell+1}, \mathbf{p}'_{\ell+1}, \boldsymbol{\rho}'_{\ell+1})'$ to satisfy:*

$$p_{ij}^{(\ell+1)} = \frac{\sum_{t=m+1}^{T} P(s_t = j, s_{t-1} = i | \mathbf{Y}; \boldsymbol{\lambda}_\ell)}{\sum_{t=m+1}^{T} P(s_{t-1} = i | \mathbf{Y}; \boldsymbol{\lambda}_\ell)} \tag{36}$$

for $i, j = 1, \dots, M$.

$$\sum_{t=m+1}^{T} \sum_{s_t=1}^{M} \cdots \sum_{s_{t-m}=1}^{M} \frac{\partial log f(\mathbf{y}_t | \mathbf{z}_t; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \bigg|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{\ell+1}} P(s_t, \dots, s_{t-m} | \mathbf{Y}; \boldsymbol{\lambda}_\ell) = \mathbf{0} \tag{37}$$

where $\mathbf{z}_t = (s_t, \dots, s_{t-m}, \mathbf{y}'_{t-1}, \dots, \mathbf{y}'_{t-m})'$, and

$$\boldsymbol{\rho}_{i_m, i_{m-1}, \dots, i_1}^{(\ell+1)} = P(s_m = i_m, s_{m-1} = i_{m-1}, \dots, s_1 = i_1 | \mathbf{Y}; \boldsymbol{\lambda}_\ell) \tag{38}$$

for $i_1, \ldots, i_m \in \{1, \ldots, M\}$.

Proof. To prove formula (36), we show that it maximizes $\mathcal{Q}(\boldsymbol{\lambda}_{\ell+1}; \boldsymbol{\lambda}_\ell, \mathbf{Y})$ with respect to $\mathbf{p}_{\ell+1}$. Note first that

$$\begin{aligned} P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda}) = & P(\mathbf{y}_T|\mathbf{z}_T; \boldsymbol{\theta}) P(s_T|s_{T-1}; \mathbf{p}) \\ & P(\mathbf{y}_{T-1}|\mathbf{z}_{T-1}; \boldsymbol{\theta}) P(s_{T-1}|s_{T-2}; \mathbf{p}) \\ & \vdots \\ & P(\mathbf{y}_{m+1}|\mathbf{z}_{m+1}; \boldsymbol{\theta}) P(s_{m+1}|s_m; \mathbf{p}) \rho_{s_m, s_{m-1}, \ldots, s_1} \end{aligned}$$

Differentiating $P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda})$ with respect to p_{ij} (a representative element of $\mathbf{p}$), we obtain

$$\frac{\partial P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda})}{\partial p_{ij}} = \sum_{t=m+1}^{T} \frac{\partial log P(s_t|s_{t-1}; \mathbf{p})}{\partial p_{ij}} P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda}).$$

But recall that

$$\frac{\partial log P(s_t|s_{t-1}; \mathbf{p})}{\partial p_{ij}} = \begin{cases} \frac{1}{p_{ij}} & if \quad s_t = j \quad and \quad s_{t-1} = i \\ 0 & otherwise \end{cases}$$

Using notation $\delta_{[]}$ for the Kronecker delta (that is, $\delta_{[A]} = 1$ when the event A occurs and zero otherwise), we get

$$\frac{\partial P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda})}{\partial p_{ij}} = p_{ij}^{-1} \{ \sum_{t=m+1}^{T} \delta_{[s_t=j, s_{t-1}=i]} \} P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda})$$

hence

$$\frac{\partial log P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda})}{\partial p_{ij}} = p_{ij}^{-1} \{ \sum_{t=m+1}^{T} \delta_{[s_t=j, s_{t-1}=i]} \}.$$

By definition of $\mathcal{Q}(\cdot)$, it follows that

$$\begin{aligned} \frac{\partial \mathcal{Q}(\boldsymbol{\lambda}_{\ell+1}; \boldsymbol{\lambda}_\ell, \mathbf{Y})}{\partial p_{ij}^{(\ell+1)}} &= \int_{\mathbf{S}} \frac{\partial log P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda}_{\ell+1})}{\partial p_{ij}^{(\ell+1)}} P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda}_\ell) \\ &= \int_{\mathbf{S}} [p_{ij}^{(\ell+1)}]^{-1} \{ \sum_{t=m+1}^{T} \delta_{[s_t=j, s_{t-1}=i]} \} P(\mathbf{Y}, \mathbf{S}; \boldsymbol{\lambda}_\ell) \end{aligned}$$

Noting that

$$\int_{\mathbf{S}} \delta_{[s_t=j,s_{t-1}=i]} P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}_\ell) = P(s_t = j, s_{t-1} = i|\mathbf{Y};\boldsymbol{\lambda}_\ell) f(\mathbf{Y};\boldsymbol{\lambda}_\ell)$$

we get

$$\text{(39)} \qquad \frac{\partial \mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_\ell,\mathbf{Y})}{\partial p_{ij}^{(\ell+1)}} = [p_{ij}^{(\ell+1)}]^{-1} \sum_{t=m+1}^{T} P(s_t = j, s_{t-1} = i|\mathbf{Y};\boldsymbol{\lambda}_\ell) f(\mathbf{Y};\boldsymbol{\lambda}_\ell)$$

Now our task in the EM algorithm was to find the value of $\mathbf{p}_{\ell+1}$ for which $\mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_\ell,\mathbf{Y})$ was maximized. Imposing the constraint $\sum_{j=1}^{M} p_{ij}^{(\ell+1)} = 1$, we form the Lagrangean

$$\mathcal{L}(\boldsymbol{\lambda}_{\ell+1}) = \mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_\ell,\mathbf{Y}) - \mu_i(\sum_{j=1}^{M} p_{ij}^{(\ell+1)} - 1)$$

from which the first-order conditions are

$$\text{(40)} \qquad \frac{\partial \mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_\ell,\mathbf{Y})}{\partial p_{ij}^{(\ell+1)}} = \mu_i$$

for $j = 1,\ldots,M$. Substituting (39) in (40) yields

$$\text{(41)} \qquad \sum_{t=m+1}^{T} P(s_t = j, s_{t-1} = i|\mathbf{Y};\boldsymbol{\lambda}_\ell) = \frac{p_{ij}^{(\ell+1)}\mu_i}{f(\mathbf{Y};\boldsymbol{\lambda}_\ell)}$$

Summing up the last relations for $j = 1,\ldots,M$ we obtain

$$\sum_{t=m+1}^{T}\sum_{j=1}^{M} P(s_t = j, s_{t-1} = i|\mathbf{Y};\boldsymbol{\lambda}_\ell) = \sum_{j=1}^{M} \frac{p_{ij}^{(\ell+1)}\mu_i}{f(\mathbf{Y};\boldsymbol{\lambda}_\ell)} = \frac{\mu_i}{f(\mathbf{Y};\boldsymbol{\lambda}_\ell)}$$

or, equivalently,

$$\sum_{t=m+1}^{T} P(s_{t-1} = i|\mathbf{Y};\boldsymbol{\lambda}_\ell) = \frac{\mu_i}{f(\mathbf{Y};\boldsymbol{\lambda}_\ell)}$$

Substituting this formula in (41) gives (36). To prove formula (37) we differentiate $P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda})$

with respect to $\boldsymbol{\alpha}$. This gives

$$\frac{\partial P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda})}{\partial \boldsymbol{\alpha}} = \sum_{t=m+1}^{T} \frac{\partial logf(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda})$$

hence

$$\frac{\partial logP(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda})}{\partial \boldsymbol{\alpha}} = \sum_{t=m+1}^{T} \frac{\partial logf(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$$

The point to note here is that $f(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\alpha})$ depends on s (through $\mathbf{z}_t$) at most for dates $t, t-1, \ldots, t-m$. Thus we have

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_\ell,\mathbf{Y})}{\partial \boldsymbol{\alpha}_{\ell+1}} &= \int_{\mathbf{S}} \frac{\partial logP(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}_{\ell+1})}{\partial \boldsymbol{\alpha}_{\ell+1}} P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}_\ell) \\
&= \sum_{t=m+1}^{T} \int_{\mathbf{S}} \frac{\partial logf(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\alpha}_{\ell+1})}{\partial \boldsymbol{\alpha}_{\ell+1}} P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}_\ell) \\
&= \sum_{t=m+1}^{T} \sum_{s_t=1}^{M} \cdots \sum_{s_{t-m}=1}^{M} \{ \frac{\partial logf(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\alpha}_{\ell+1})}{\partial \boldsymbol{\alpha}_{\ell+1}} P(s_t,\ldots,s_{t-m}|\mathbf{Y};\boldsymbol{\lambda}_\ell) f(\mathbf{Y};\boldsymbol{\lambda}_\ell)\}
\end{aligned}$$

Noting that $f(\mathbf{Y};\boldsymbol{\lambda}_\ell)$ is not a function of the index t, it can be taken outside of the summation operators. Setting this derivative equal to zero, we get formula (37). To prove formula (38) we differentiate $logP(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda})$ with respect to $\rho_{i_m,\ldots,i_1}$. This gives

$$\frac{\partial logP(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda})}{\partial \rho_{i_m,\ldots,i_1}} = [\rho_{i_m,\ldots,i_1}]^{-1} \delta_{[s_m=i_m,\ldots,s_1=i_1]}$$

hence

$$\frac{\partial \mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_\ell,\mathbf{Y})}{\partial \rho^{(\ell+1)}_{i_m,\ldots,i_1}} = \int_{\mathbf{S}} [\rho^{(\ell+1)}_{i_m,\ldots,i_1}]^{-1} \delta_{[s_m=i_m,\ldots,s_1=i_1]} P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}_\ell)$$

Maximizing $\mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_\ell,\mathbf{Y})$ subject to the constraint that the sum of the elements of $\boldsymbol{\rho}_{\ell+1}$ equals unity yields the first-order conditions

$$\frac{\partial \mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_\ell,\mathbf{Y})}{\partial \rho^{(\ell+1)}_{i_m,\ldots,i_1}} = \mu$$

for μ the Lagrange multiplier associated with the summation constraint. Thus we have

$$\int_{\mathbf{S}} \delta_{[s_m=i_m,\ldots,s_1=i_1]} P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda}_\ell) = \mu\, \rho^{(\ell+1)}_{i_m,\ldots,i_1}$$

or

$$P(s_m = i_m,\ldots,s_1 = i_1|\mathbf{Y};\boldsymbol{\lambda}_\ell) f(\mathbf{Y};\boldsymbol{\lambda}_\ell) = \mu\, \rho^{(\ell+1)}_{i_m,\ldots,i_1} \tag{42}$$

Summing up these relations over all possible values of $i_m,\ldots,i_1$, we obtain $\mu = f(\mathbf{Y};\boldsymbol{\lambda}_\ell)$. Substituting the last formula in (42) gives (38). □

Thus the EM algorithm begins at iteration $\ell = 0$ with an arbitrary guess for the parameter vector $\boldsymbol{\lambda}_\ell = \boldsymbol{\lambda}_0$. For this guess, we calculate the smoothed probabilities $P(s_t,\ldots,s_{t-m}|\mathbf{Y};\boldsymbol{\lambda}_0)$. Equations (36)-(38) are then solved for $\boldsymbol{\lambda}_{\ell+1} = \boldsymbol{\lambda}_1$. The next iteration ($\ell = 1$) takes $\boldsymbol{\lambda}_\ell$ to be the value $\boldsymbol{\lambda}_1$ calculated from the previous iteration, and solves equations (36)-(38) for $\boldsymbol{\lambda}_{\ell+1} = \boldsymbol{\lambda}_2$. The process continues until a fixed point $\boldsymbol{\lambda}_{\ell+1} = \boldsymbol{\lambda}_\ell$ is satisfactorily approximated. Calculation of $\boldsymbol{\lambda}_{\ell+1}$ as a function of $\boldsymbol{\lambda}_\ell$ is quite easy. Once one has calculated smoothed probabilities such as $P(s_t = j, s_{t-1} = i|\mathbf{Y};\boldsymbol{\lambda}_\ell)$, equations (36) and (38) allow calculations of $\mathbf{p}_{\ell+1}$ and $\boldsymbol{\rho}_{\ell+1}$. Then it should be clear that in order to implement the EM algorithm, it is not at all necessary to calculate such cumbersome expressions as $P(\mathbf{Y},\mathbf{S};\boldsymbol{\lambda})$ or $\mathcal{Q}(\boldsymbol{\lambda}_{\ell+1};\boldsymbol{\lambda}_\ell,\mathbf{Y})$. Rather, all one ever needs to evaluate are the smoothed inferences about the unobserved state: $P(s_t, s_{t-1},\ldots,s_{t-m}|\mathbf{Y};\boldsymbol{\lambda}_\ell)$, as done in Section 1.6.

1.8 State-Space Models with Markov Switching

In [5] Kim extended the Hamilton model with Markov switching [2][3] to a general state-space model. Here we illustrate the basic filtering and smoothing algorithm obtained by Kim, and give an explicit proof of his formulae by using arguments from [1], chp.15.

1.8.1 Specification of the Model

Let us consider the following state-space model with Markov switching

$$\begin{cases} \mathbf{y}_t = \mathbf{F}_{s_t}\mathbf{x}_t + \boldsymbol{\beta}_{s_t}\mathbf{z}_t + \mathbf{e}_t \\ \mathbf{x}_t = \mathbf{A}_{s_t}\mathbf{x}_{t-1} + \boldsymbol{\gamma}_{s_t}\mathbf{z}_t + \mathbf{G}_{s_t}\boldsymbol{\nu}_t \end{cases}$$

where

$$\begin{pmatrix} \mathbf{e}_t \\ \boldsymbol{\nu}_t \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{pmatrix}\right)$$

Here $\mathbf{x}_t$ is a $J \times 1$ unobserved state vector, $\mathbf{z}_t$ is a $K \times 1$ vector of weakly exogenous or lagged dependent variables, $\mathbf{e}_t$ and $\boldsymbol{\nu}_t$ are $n \times 1$ and $L \times 1$ vectors of disturbances, $\mathbf{y}_t$ is a $n \times 1$ vector of measurements, and the parameters $\mathbf{F}_{s_t}, \boldsymbol{\beta}_{s_t}, \mathbf{A}_{s_t}, \boldsymbol{\gamma}_{s_t}$ and $\mathbf{G}_{s_t}$ are $n \times J$, $n \times K$, $J \times J$, $J \times K$ and $J \times L$ matrices, respectively, depending upon an M-state first-order Markov process (s_t) with transition probability matrix $\mathbf{P} = (p_{ij})$. The parameter matrices may be known under different regimes, but in some circumstances a particular element of a parameter matrix takes on different values which are unknown. The model incorporates also the latter case. When a particular element of the matrix $\mathbf{F}_{s_t}$, for example, switches from one state to another, and when the value of that element are unknown under different states, it can be modeled in the following way. Assuming that the state variable s_t can take the values $1, \ldots, M$, the (i, j)th element of $\mathbf{F}_{s_t}$ can be written as

$$f_{i,j,s_t} = f_{i,j,1}s_{1t} + \cdots + f_{i,j,M}s_{Mt}$$

where $s_{mt} = 1$ if $s_t = m$, and zero otherwise. The elements $f_{i,j,m}$, $m = 1, \ldots, M$ are part of the parameters to be estimated.

1.8.2 Basic Filtering and Estimation

Suppose the parameters of the model are known. Let $\boldsymbol{\psi}_{t-1} = (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_1, \mathbf{z}'_t, \mathbf{z}'_{t-1}, \ldots, \mathbf{z}'_1)'$ denote the vector of observations received as of time $t-1$. In the usual derivation of the Kalman filter for a fixed-coefficient state-space model (see [1], chp.15), the goal is to form a forecast of the unobserved state vector $\mathbf{x}_t$ based on $\boldsymbol{\psi}_{t-1}$, denoted $\mathbf{x}_{t|t-1} = E(\mathbf{x}_t|\boldsymbol{\psi}_{t-1})$. Similarly, in the conventional fixed-coefficient case, the matrix $\mathbf{P}_{t|t-1}$ denotes the mean square error of the forecast, i.e.,

$$\mathbf{P}_{t|t-1} = E[(\mathbf{x}_t - \mathbf{x}_{t|t-1})(\mathbf{x}_t - \mathbf{x}_{t|t-1})^{'}|\boldsymbol{\psi}_{t-1}].$$

Here the goal will be to form a forecast of $\mathbf{x}_t$ based not just on $\boldsymbol{\psi}_{t-1}$ but also conditional on the random variable s_t taking on the value j and on s_{t-1} taking on the value i, denoted

$$\mathbf{x}_{t|t-1}^{(i,j)} = E(\mathbf{x}_t|\boldsymbol{\psi}_{t-1}, s_t = j, s_{t-1} = i)$$

for $i, j \in \{1, \ldots, M\}$. Associated to these forecasts are M^2 different mean squared error matrices

$$\mathbf{P}_{t|t-1}^{(i,j)} = E[(\mathbf{x}_t - \mathbf{x}_{t|t-1})(\mathbf{x}_t - \mathbf{x}_{t|t-1})^{'}|\boldsymbol{\psi}_{t-1}, s_t = j, s_{t-1} = i]$$

The computation algorithm is given by the following filter:

Theorem 8.1 *(Kim covariance filter) For $t \geq 0$, we have the following relations*

$$\mathbf{x}_{t|t-1}^{(i,j)} = \mathbf{A}_j \mathbf{x}_{t-1|t-1}^{i} + \boldsymbol{\gamma}_j \mathbf{z}_t \tag{43}$$

$$\mathbf{P}_{t|t-1}^{(i,j)} = \mathbf{A}_j \mathbf{P}_{t-1|t-1}^{i} \mathbf{A}_j^{'} + \mathbf{G}_j \mathbf{Q} \mathbf{G}_j^{'} \tag{44}$$

$$\boldsymbol{\eta}_{t|t-1}^{(i,j)} = \mathbf{y}_t - \mathbf{F}_j \mathbf{x}_{t|t-1}^{(i,j)} - \boldsymbol{\beta}_j \mathbf{z}_t \tag{45}$$

$$\mathbf{H}_t^{(i,j)} = \mathbf{F}_j \mathbf{P}_{t|t-1}^{(i,j)} \mathbf{F}_j^{'} + \mathbf{R} \tag{46}$$

$$\mathbf{K}_t^{(i,j)} = \mathbf{P}_{t|t-1}^{(i,j)} \mathbf{F}_j^{'} [\mathbf{H}_t^{(i,j)}]^{-1} \tag{47}$$

$$\mathbf{x}_{t|t}^{(i,j)} = \mathbf{x}_{t|t-1}^{(i,j)} + \mathbf{K}_t^{(i,j)} \boldsymbol{\eta}_{t|t-1}^{(i,j)} \tag{48}$$

$$\mathbf{P}_{t|t}^{(i,j)} = (\mathbf{I} - \mathbf{K}_t^{(i,j)}\mathbf{F}_j)\mathbf{P}_{t|t-1}^{(i,j)} \tag{49}$$

where $\mathbf{x}_{t-1|t-1}^{i}$ *is an inference about* $\mathbf{x}_{t-1}$ *based on information up to time* $t-1$, *given* $s_{t-1} = i$; $\boldsymbol{\eta}_{t|t-1}^{(i,j)}$ *is the conditional forecast error of* $\mathbf{y}_t$ *based on information up to time* $t-1$, *given* $s_t = j$ *and* $s_{t-1} = i$; $\mathbf{H}_t^{(i,j)}$ *is the conditional variance of* $\boldsymbol{\eta}_{t|t-1}^{(i,j)}$; *and* $\mathbf{K}_t^{(i,j)}$ *is the Kalman gain.*

Proof. Let us define the conditional forecast error of $\mathbf{y}_t$ at horizon 1 as

$$\begin{aligned}
\boldsymbol{\eta}_{t|t-1}^{(i,j)} &= \mathbf{y}_t - \mathbf{y}_{t|t-1}^{(i,j)} = \mathbf{y}_t - E(\mathbf{y}_t|\boldsymbol{\psi}_{t-1}, s_t = j, s_{t-1} = i) \\
&= \mathbf{y}_t - E(\mathbf{F}_{s_t}\mathbf{x}_t + \boldsymbol{\beta}_{s_t}\mathbf{z}_t + \mathbf{e}_t|\boldsymbol{\psi}_{t-1}, s_t = j, s_{t-1} = i) \\
&= \mathbf{y}_t - \mathbf{F}_j E(\mathbf{x}_t|\boldsymbol{\psi}_{t-1}, s_t = j, s_{t-1} = i) - \boldsymbol{\beta}_j E(\mathbf{z}_t|\boldsymbol{\psi}_{t-1}) \\
&= \mathbf{y}_t - \mathbf{F}_j\mathbf{x}_{t|t-1}^{(i,j)} - \boldsymbol{\beta}_j\mathbf{z}_t
\end{aligned}$$

which is formula (45). Using (45) and the geometric subspace decomposition as in [1] chp. 15, we get

$$\begin{aligned}
\mathbf{x}_{t|t}^{(i,j)} &= E(\mathbf{x}_t|\boldsymbol{\psi}_t, s_t = j, s_{t-1} = i) \\
&= E(\mathbf{x}_t|\boldsymbol{\psi}_{t-1}, \boldsymbol{\eta}_{t|t-1}^{(i,j)}, s_t = j, s_{t-1} = i) \\
&= E(\mathbf{x}_t|\boldsymbol{\psi}_{t-1}, s_t = j, s_{t-1} = i) + E(\mathbf{x}_t|\boldsymbol{\eta}_{t|t-1}^{(i,j)}) - E(\mathbf{x}_t) \\
&= \mathbf{x}_{t|t-1}^{(i,j)} + cov(\mathbf{x}_t, \boldsymbol{\eta}_{t|t-1}^{(i,j)})[var(\boldsymbol{\eta}_{t|t-1}^{(i,j)})]^{-1}\boldsymbol{\eta}_{t|t-1}^{(i,j)} \\
&= \mathbf{x}_{t|t-1}^{(i,j)} + cov(\mathbf{x}_t, \boldsymbol{\eta}_{t|t-1}^{(i,j)})[\mathbf{H}_t^{(i,j)}]^{-1}\boldsymbol{\eta}_{t|t-1}^{(i,j)}
\end{aligned}$$

Now we have

$$\begin{aligned}
cov(\mathbf{x}_t, \boldsymbol{\eta}_{t|t-1}^{(i,j)}) &= E((\mathbf{x}_t - \mathbf{x}_{t|t-1})\boldsymbol{\eta}_{t|t-1}^{(i,j)'}|\boldsymbol{\psi}_{t-1}, s_t = j, s_{t-1} = i) \\
&= E((\mathbf{x}_t - \mathbf{x}_{t|t-1})(\mathbf{F}_j(\mathbf{x}_t - \mathbf{x}_{t|t-1}) + \mathbf{e}_t)'|\boldsymbol{\psi}_{t-1}, s_t = j, s_{t-1} = i) \\
&= E((\mathbf{x}_t - \mathbf{x}_{t|t-1})(\mathbf{x}_t - \mathbf{x}_{t|t-1})'|\boldsymbol{\psi}_{t-1}, s_t = j, s_{t-1} = i)\mathbf{F}_j' \\
&= \mathbf{P}_{t|t-1}^{(i,j)}\mathbf{F}_j'
\end{aligned}$$

hence

$$\begin{aligned}
\mathbf{x}_{t|t}^{(i,j)} &= \mathbf{x}_{t|t-1}^{(i,j)} + \mathbf{P}_{t|t-1}^{(i,j)}\mathbf{F}_j'[\mathbf{H}_t^{(i,j)}]^{-1}\boldsymbol{\eta}_{t|t-1}^{(i,j)} \\
&= \mathbf{x}_{t|t-1}^{(i,j)} + \mathbf{K}_t^{(i,j)}\boldsymbol{\eta}_{t|t-1}^{(i,j)}
\end{aligned}$$

which proves formulae (47) and (48). We compute the filtering error covariance matrix $\mathbf{P}_{t|t}^{(i,j)} =$

$var(\mathbf{x}_t - \mathbf{x}_{t|t}^{(i,j)})$. Since the corresponding error $\mathbf{x}_t - \mathbf{x}_{t|t}^{(i,j)}$ is not correlated with $\mathbf{y}_1, \dots, \mathbf{y}_t$, it is not correlated with $\boldsymbol{\eta}_{t|t-1}^{(i,j)}$. Using (47) and (48), we get

$$\begin{aligned}
\mathbf{P}_{t|t}^{(i,j)} &= E((\mathbf{x}_t - \mathbf{x}_{t|t}^{(i,j)})(\mathbf{x}_t - \mathbf{x}_{t|t}^{(i,j)})') \\
&= E((\mathbf{x}_t - \mathbf{x}_{t|t-1}^{(i,j)} - \mathbf{K}_t^{(i,j)}\boldsymbol{\eta}_{t|t-1}^{(i,j)})(\mathbf{x}_t - \mathbf{x}_{t|t-1}^{(i,j)})') \\
&= E((\mathbf{x}_t - \mathbf{x}_{t|t-1}^{(i,j)})(\mathbf{x}_t - \mathbf{x}_{t|t-1}^{(i,j)})') - \mathbf{K}_t^{(i,j)}E(\boldsymbol{\eta}_{t|t-1}^{(i,j)}(\mathbf{x}_t - \mathbf{x}_{t|t-1}^{(i,j)})') \\
&= \mathbf{P}_{t|t-1}^{(i,j)} - \mathbf{K}_t^{(i,j)}E(\boldsymbol{\eta}_{t|t-1}^{(i,j)}(\mathbf{x}_t - \mathbf{x}_{t|t}^{(i,j)} + \mathbf{K}_t^{(i,j)}\boldsymbol{\eta}_{t|t-1}^{(i,j)})') \\
&= \mathbf{P}_{t|t-1}^{(i,j)} - \mathbf{K}_t^{(i,j)}E(\boldsymbol{\eta}_{t|t-1}^{(i,j)}\boldsymbol{\eta}_{t|t-1}^{(i,j)'})\mathbf{K}_t^{(i,j)'} \\
&= \mathbf{P}_{t|t-1}^{(i,j)} - \mathbf{K}_t^{(i,j)}var(\boldsymbol{\eta}_{t|t-1}^{(i,j)})\mathbf{K}_t^{(i,j)'} \\
&= \mathbf{P}_{t|t-1}^{(i,j)} - \mathbf{K}_t^{(i,j)}\mathbf{H}_t^{(i,j)}\mathbf{K}_t^{(i,j)'} \\
&= \mathbf{P}_{t|t-1}^{(i,j)} - \mathbf{K}_t^{(i,j)}\mathbf{H}_t^{(i,j)}[\mathbf{H}_t^{(i,j)}]^{-1}\mathbf{F}_j\mathbf{P}_{t|t-1}^{(i,j)} \\
&= \mathbf{P}_{t|t-1}^{(i,j)} - \mathbf{K}_t^{(i,j)}\mathbf{F}_j\mathbf{P}_{t|t-1}^{(i,j)} \\
&= (\mathbf{I} - \mathbf{K}_t^{(i,j)}\mathbf{F}_j)\mathbf{P}_{t|t-1}^{(i,j)}
\end{aligned}$$

which proves (49). To prove (43) we consider

$$\mathbf{x}_t = \mathbf{A}_{s_t}\mathbf{x}_{t-1} + \boldsymbol{\gamma}_{s_t}\mathbf{z}_t + \mathbf{G}_{s_t}\boldsymbol{\nu}_t$$

Computing the conditional expectation for each element with respect to $\boldsymbol{\Psi}_{t-1}, s_t = j, s_{t-1} = i$, and using the fact that the innovation $\boldsymbol{\nu}_t$ and the measurement errors are uncorrelated, we get

$$\begin{aligned}
\mathbf{x}_{t|t-1}^{(i,j)} &= E(\mathbf{x}_t|\boldsymbol{\Psi}_{t-1}, s_t = j, s_{t-1} = i) \\
&= E(\mathbf{A}_{s_t}\mathbf{x}_{t-1} + \boldsymbol{\gamma}_{s_t}\mathbf{z}_t + \mathbf{G}_{s_t}\boldsymbol{\nu}_t|\boldsymbol{\Psi}_{t-1}, s_t = j, s_{t-1} = i) \\
&= \mathbf{A}_jE(\mathbf{x}_{t-1}|\boldsymbol{\Psi}_{t-1}, s_{t-1} = i) + \boldsymbol{\gamma}_jE(\mathbf{z}_t|\boldsymbol{\Psi}_{t-1}) \\
&= \mathbf{A}_j\mathbf{x}_{t-1|t-1}^i + \boldsymbol{\gamma}_j\mathbf{z}_t
\end{aligned}$$

as $\mathbf{z}_t$ belongs to $\boldsymbol{\Psi}_{t-1}$. To prove (44), the corresponding error is (use (43))

$$\mathbf{x}_t - \mathbf{x}_{t|t-1}^{(i,j)} = \mathbf{A}_{s_t}\mathbf{x}_{t-1} + \boldsymbol{\gamma}_{s_t}\mathbf{z}_t + \mathbf{G}_{s_t}\boldsymbol{\nu}_t - \mathbf{A}_j\mathbf{x}_{t-1|t-1}^i - \boldsymbol{\gamma}_j\mathbf{z}_t$$

So we obtain

$$\begin{aligned}\mathbf{P}_{t|t-1}^{(i,j)} &= var(\mathbf{x}_t - \mathbf{x}_{t|t-1}^{(i,j)}|\mathbf{\Psi}_{t-1}, s_t = j, s_{t-1} = i) \\ &= var(\mathbf{A}_j(\mathbf{x}_{t-1} - \mathbf{x}_{t-1|t-1}^{i})) + var(\mathbf{G}_j\boldsymbol{\nu}_t) \\ &= \mathbf{A}_j var(\mathbf{x}_{t-1} - \mathbf{x}_{t-1|t-1}^{i})\mathbf{A}_j^{'} + \mathbf{G}_j var(\boldsymbol{\nu}_t)\mathbf{G}_j^{'} \\ &= \mathbf{A}_j\mathbf{P}_{t-1|t-1}^{i}\mathbf{A}_j^{'} + \mathbf{G}_j\mathbf{Q}\mathbf{G}_j^{'}.\end{aligned}$$

To prove (46), we consider

$$\mathbf{y}_t = \mathbf{F}_{s_t}\mathbf{x}_t + \boldsymbol{\beta}_{s_t}\mathbf{z}_t + \mathbf{e}_t$$

hence

$$\mathbf{y}_{t|t-1}^{(i,j)} = \mathbf{F}_j\mathbf{x}_{t|t-1}^{(i,j)} + \boldsymbol{\beta}_j\mathbf{z}_t$$

Thus we obtain

$$\mathbf{y}_t - \mathbf{y}_{t|t-1}^{(i,j)} = \mathbf{F}_j(\mathbf{x}_t - \mathbf{x}_{t|t-1}^{(i,j)}) + \mathbf{e}_t$$

and

$$\begin{aligned}\mathbf{H}_t^{(i,j)} &= var(\mathbf{y}_t - \mathbf{y}_{t|t-1}^{(i,j)}) \\ &= var(\mathbf{F}_j(\mathbf{x}_t - \mathbf{x}_{t|t-1}^{(i,j)})) + var(\mathbf{e}_t) \\ &= \mathbf{F}_j\mathbf{P}_{t|t-1}^{(i,j)}\mathbf{F}_j^{'} + \mathbf{R}\end{aligned}$$

which is (46). □

To make the above filtering operable, it is necessary to introduce the following approximations (see [5], p.6, for more details)

$$\mathbf{x}_{t|t}^{j} = \frac{\sum_{i=1}^{M} P(s_{t-1} = i, s_t = j|\mathbf{\Psi}_t)\mathbf{x}_{t|t}^{(i,j)}}{P(s_t = j|\mathbf{\Psi}_t)} \tag{50}$$

$$\mathbf{P}_{t|t}^{j} = \frac{\sum_{i=1}^{M} P(s_{t-1} = i, s_t = j|\mathbf{\Psi}_t)[P_{t|t}^{(i,j)} + (\mathbf{x}_{t|t}^{j} - \mathbf{x}_{t|t}^{(i,j)})(\mathbf{x}_{t|t}^{j} - \mathbf{x}_{t|t}^{(i,j)})^{'}]}{P(s_t = j|\mathbf{\Psi}_t)} \tag{51}$$

The last thing that remains to be considered to complete the filtering is to calculate $P(s_{t-1} = i, s_t = j|\mathbf{\Psi}_t)$ and other probability terms.

The *following procedure* explains how a complete basic filtering can be performed using the above formulae (43) through (51). Note that the basic filter accepts the three inputs $\mathbf{x}^i_{t-1|t-1}, \mathbf{P}^i_{t-1|t-1}$ and $P(s_{t-1} = i, s_{t-2} = i^{'}|\mathbf{\Psi}_{t-1})$, and has three outputs $\mathbf{x}^j_{t|t}, \mathbf{P}^j_{t|t}$ and $P(s_t = j|\mathbf{\Psi}_t)$.

Step 1. Calculate

$$
\begin{aligned}
P(s_{t-1} = i, s_t = j|\mathbf{\Psi}_{t-1}) &= \sum_{i^{'}=1}^{M} P(s_{t-2} = i^{'}, s_{t-1} = i|\mathbf{\Psi}_{t-1}) P(s_t = j|s_{t-1} = i) \\
&= P(s_t = j|s_{t-1} = i) \sum_{i^{'}=1}^{M} P(s_{t-2} = i^{'}, s_{t-1} = i|\mathbf{\Psi}_{t-1}) \\
&= p_{ij} \sum_{i^{'}=1}^{M} P(s_{t-2} = i^{'}, s_{t-1} = i|\mathbf{\Psi}_{t-1})
\end{aligned} \tag{52}
$$

Step 2. Calculate the joint conditional density function

$$
f(\mathbf{y}_t, s_{t-1} = i, s_t = j|\mathbf{\Psi}_{t-1}) = f(\mathbf{y}_t|\mathbf{\Psi}_{t-1}, s_{t-1} = i, s_t = j) P(s_{t-1} = i, s_t = j|\mathbf{\Psi}_{t-1}) \tag{53}
$$

where

$$
f(\mathbf{y}_t|\mathbf{\Psi}_{t-1}, s_{t-1} = i, s_t = j) = (2\pi)^{-\frac{n}{2}} |\mathbf{H}_t^{(i,j)}|^{-\frac{1}{2}} exp(-\frac{1}{2} \boldsymbol{\eta}_{t|t-1}^{(i,j)^{'}} [\mathbf{H}_t^{(i,j)}]^{-1} \boldsymbol{\eta}_{t|t-1}^{(i,j)})
$$

Step 3. Calculate

$$
P(s_{t-1} = i, s_t = j|\mathbf{\Psi}_t) = \frac{f(\mathbf{y}_t, s_{t-1} = i, s_t = j|\mathbf{\Psi}_{t-1})}{f(\mathbf{y}_t|\mathbf{\Psi}_{t-1})} \tag{54}
$$

where

$$
f(\mathbf{y}_t|\mathbf{\Psi}_{t-1}) = \sum_{i=1}^{M} \sum_{j=1}^{M} f(\mathbf{y}_t, s_{t-1} = i, s_t = j|\mathbf{\Psi}_{t-1})
$$

Step 4. From (50) and (51) and output from Step 3, we get $\mathbf{x}^j_{t|t}$ and $\mathbf{P}^j_{t|t}$. The remaining output $P(s_t = j|\mathbf{\Psi}_t)$ can be calculated by

$$
P(s_t = j|\mathbf{\Psi}_t) = \sum_{i=1}^{M} P(s_{t-1} = i, s_t = j|\mathbf{\Psi}_t)
$$

The sample conditional log-likelihood function can be obtained from Step 3

$$LL = logf(\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_1|\mathbf{\Psi}_0) = \sum_{t=1}^{T} logf(\mathbf{y}_t|\mathbf{\Psi}_{t-1})$$

The above filter is obtained under the assumption that parameters of the model are known. To estimate the parameters, we can maximize the log-likelihood function above with respect to the underlying unknown parameters by using the EM method (see Section 1.7).

1.8.3 Smoothing

Once parameters are estimated, we can get inference about s_t and $\mathbf{x}_t$ based on all the information in the sample, that is, $P(s_t = j|\mathbf{\Psi}_T)$ and $\mathbf{x}_{t|T}$, for $t = 1, \ldots, T$. Let us consider the following derivation of the joint probability that $s_t = j$ and $s_{t+1} = k$ based on full information set $\mathbf{\Psi}_T$ (here it is assumed that no lagged dependent variables appear in the model):

$$\begin{aligned} P(s_t = j, s_{t+1} = k|\mathbf{\Psi}_T) &= P(s_{t+1} = k|\mathbf{\Psi}_T)P(s_t = j|s_{t+1} = k, \mathbf{\Psi}_T) \\ &\sim P(s_{t+1} = k|\mathbf{\Psi}_T)P(s_t = j|s_{t+1} = k, \mathbf{\Psi}_t) \\ &= \frac{P(s_{t+1} = k|\mathbf{\Psi}_T)P(s_t = j|\mathbf{\Psi}_t)P(s_{t+1} = k|s_t = j)}{P(s_{t+1} = k|\mathbf{\Psi}_t)} \end{aligned}$$

and

$$P(s_t = j|\mathbf{\Psi}_T) = \sum_{k=1}^{M} P(s_t = j, s_{t+1} = k|\mathbf{\Psi}_T)$$

Note that the first formula involves an approximation. To investigate the nature of this approximation, define $\mathbf{h}_{t+1,T} = (\mathbf{y}'_{t+1}, \ldots, \mathbf{y}'_T, \mathbf{z}'_{t+1}, \ldots, \mathbf{z}'_T)'$ for $t < T$, that is, $\mathbf{h}_{t+1,T}$ is the vector of observations from date $t + 1$ to T. Then we have

$$\begin{aligned} P(s_t = j|s_{t+1} = k, \mathbf{\Psi}_T) &= P(s_t = j|s_{t+1} = k, \mathbf{h}_{t+1,T}, \mathbf{\Psi}_t) \\ &= \frac{f(s_t = j, \mathbf{h}_{t+1,T}|s_{t+1} = k, \mathbf{\Psi}_t)}{f(\mathbf{h}_{t+1,T}|s_{t+1} = k, \mathbf{\Psi}_t)} \\ &= \frac{P(s_t = j|s_{t+1} = k, \mathbf{\Psi}_t)f(\mathbf{h}_{t+1,T}|s_{t+1} = k, s_t = j, \mathbf{\Psi}_t)}{f(\mathbf{h}_{t+1,T}|s_{t+1} = k, \mathbf{\Psi}_t)} \end{aligned}$$

which equals $P(s_t = j|s_{t+1} = k, \mathbf{\Psi}_t)$ provided that

$$f(\mathbf{h}_{t+1,T}|s_{t+1} = k, s_t = j, \mathbf{\Psi}_t) = f(\mathbf{h}_{t+1,T}|s_{t+1} = k, \mathbf{\Psi}_t).$$

For the Hamiltonian model [2] with no lagged dependent variables, this equality holds. For the state-space model considered here, it does not hold exactly, and this why an approximation is involved above.

Like the basic filtering in (1.8.2), the smoothing algorithm for the vector $\mathbf{x}_t$ can be written as follows.

Theorem 8.2 *(Smoothing)*

Given $s_t = j$ and $s_{t+1} = k$, we have for $t = 1, \ldots, T-1$

(55) $$\mathbf{x}_{t|T}^{(j,k)} = \mathbf{x}_{t|t}^{j} + \widetilde{\mathbf{P}}_t^{(j,k)}(\mathbf{x}_{t+1|T}^{k} - \mathbf{x}_{t+1|t}^{(j,k)})$$

(56) $$\mathbf{P}_{t|T}^{(j,k)} = \mathbf{P}_{t|t}^{j} + \widetilde{\mathbf{P}}_t^{(j,k)}(\mathbf{P}_{t+1|T}^{k} - \mathbf{P}_{t+1|t}^{(j,k)})\widetilde{\mathbf{P}}_t^{(j,k)'}$$

where $\widetilde{\mathbf{P}}_t^{(j,k)} = \mathbf{P}_{t|t}^{j}\mathbf{A}_k'[\mathbf{P}_{t+1|t}^{(j,k)}]^{-1}$, $\mathbf{x}_{t|T}^{(j,k)}$ is the inference of $\mathbf{x}_t$ based on the full sample, and $\mathbf{P}_{t|T}^{(j,k)}$ is the variance-covariance matrix of $\mathbf{x}_{t|T}^{(j,k)}$ (note that $\mathbf{x}_{t|t}^{j}$ and $\mathbf{P}_{t|t}^{j}$ are given by formulae (50) and (51)).

Proof. We have to compute

$$\mathbf{x}_{t|T}^{(j,k)} = E(\mathbf{x}_t|\mathbf{\Psi}_T, s_t = j, s_{t+1} = k)$$

Reasoning as in [1], chp.15, from iterated projections, we get

$$\mathbf{x}_{t|T}^{(j,k)} = E(E(\mathbf{x}_t|I_t)|\mathbf{\Psi}_T)$$

where I_t is the set of variables $\mathbf{y}_1, \ldots, \mathbf{y}_t, \mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)}, \mathbf{z}_t, \ldots, \mathbf{z}_T, \mathbf{w}_t, \ldots, \mathbf{w}_T$, where $\mathbf{w}_t = (\mathbf{e}_t' \; \boldsymbol{\nu}_t')'$, given $s_t = j$ and $s_{t+1} = k$. We can decompose $E(\mathbf{x}_t|I_t)$ into two orthogonal projections

$\mathbf{x}_{t|t}^{j}$ and $E(\mathbf{x}_t|\mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)}) - E(\mathbf{x}_t)$. The second component is given by

$$\begin{aligned}E(\mathbf{x}_t|\mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)}) - E(\mathbf{x}_t) &= cov(\mathbf{x}_t, \mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)})[var(\mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)})]^{-1}(\mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)})\\ &= cov(\mathbf{x}_t, \mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)})[P_{t+1|t}^{(j,k)}]^{-1}(\mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)})\end{aligned}$$

By formula (43) we have

$$\begin{aligned}\mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)} &= \mathbf{x}_{t+1}^{k} - \mathbf{A}_k\mathbf{x}_{t|t}^{j} - \boldsymbol{\gamma}_k\mathbf{z}_t\\ &= \mathbf{A}_k\mathbf{x}_t + \boldsymbol{\gamma}_k\mathbf{z}_t + \mathbf{G}_k\boldsymbol{\nu}_t - \mathbf{A}_k\mathbf{x}_{t|t}^{j} - \boldsymbol{\gamma}_k\mathbf{z}_t\\ &= \mathbf{A}_k(\mathbf{x}_t - \mathbf{x}_{t|t}^{j}) + \mathbf{G}_k\boldsymbol{\nu}_t.\end{aligned}$$

Thus we get

$$\begin{aligned}cov(\mathbf{x}_t, \mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)}) &= cov(\mathbf{x}_t, \mathbf{A}_k(\mathbf{x}_t - \mathbf{x}_{t|t}^{j}) + \mathbf{G}_k\boldsymbol{\nu}_t)\\ &= cov(\mathbf{x}_t, \mathbf{A}_k(\mathbf{x}_t - \mathbf{x}_{t|t}^{j})) = E((\mathbf{x}_t - \mathbf{x}_{t|t}^{j})(\mathbf{x}_t - \mathbf{x}_{t|t}^{j})'\mathbf{A}_k')\\ &= var(\mathbf{x}_t - \mathbf{x}_{t|t}^{j})\mathbf{A}_k' = \mathbf{P}_{t|t}^{j}\mathbf{A}_k'\end{aligned}$$

Finally we get

$$E(\mathbf{x}_t|I_t) = \mathbf{x}_{t|t}^{j} + \mathbf{P}_{t|t}^{j}\mathbf{A}_k'[\mathbf{P}_{t+1|t}^{(j,k)}]^{-1}(\mathbf{x}_{t+1}^{k} - \mathbf{x}_{t+1|t}^{(j,k)}).$$

Taking the conditional expectation with respect to $\boldsymbol{\Psi}_T$ gives

$$\begin{aligned}\mathbf{x}_{t|T}^{(j,k)} &= \mathbf{x}_{t|t}^{j} + \mathbf{P}_{t|t}^{j}\mathbf{A}_k'[\mathbf{P}_{t+1|t}^{(j,k)}]^{-1}(\mathbf{x}_{t+1|T}^{k} - \mathbf{x}_{t+1|t}^{(j,k)})\\ &= \mathbf{x}_{t|t}^{j} + \widetilde{\mathbf{P}}_t^{(j,k)}(\mathbf{x}_{t+1|T}^{k} - \mathbf{x}_{t+1|t}^{(j,k)})\end{aligned}$$

which proves (55). Formula (55) implies

$$\mathbf{x}_{t|T}^{(j,k)} - \widetilde{\mathbf{P}_t}^{(j,k)}\mathbf{x}_{t+1|T}^{k} = \mathbf{x}_{t|t}^{j} - \widetilde{\mathbf{P}}_t^{(j,k)}\mathbf{x}_{t+1|t}^{(j,k)}$$

hence

$$\mathbf{x}_t - \mathbf{x}_{t|T}^{(j,k)} + \widetilde{\mathbf{P}}_t^{(j,k)}\mathbf{x}_{t+1|T}^{k} = \mathbf{x}_t - \mathbf{x}_{t|t}^{j} + \widetilde{\mathbf{P}}_t^{(j,k)}\mathbf{x}_{t+1|t}^{(j,k)}.$$

Now $\mathbf{x}_t - \mathbf{x}_{t|T}^{(j,k)}$ is uncorrelated with $\mathbf{y}_1, \ldots, \mathbf{y}_T, \mathbf{z}_1, \ldots, \mathbf{z}_T$ and therefore with $\mathbf{x}_{t+1|T}^{k}$; for the same reason, $\mathbf{x}_t - \mathbf{x}^j_{t|t}$ and $\mathbf{x}^{(j,k)}_{t+1|t}$ are uncorrelated, so that

$$\mathbf{P}^{(j,k)}_{t|T} + \widetilde{\mathbf{P}}^{(j,k)}_t var(\mathbf{x}^k_{t+1|T})\widetilde{\mathbf{P}}^{(j,k)'}_t = \mathbf{P}^j_{t|t} + \widetilde{\mathbf{P}}^{(j,k)}_t var(\mathbf{x}^{(j,k)}_{t+1|t})\widetilde{\mathbf{P}}^{(j,k)'}_t$$

hence

$$\mathbf{P}^{(j,k)}_{t|T} = \mathbf{P}^j_{t|t} + \widetilde{\mathbf{P}}^{(j,k)}_t [var(\mathbf{x}^{(j,k)}_{t+1|t}) - var(\mathbf{x}^k_{t+1|T})]\widetilde{\mathbf{P}}^{(j,k)'}_t$$

Moreover, taking the noncorrelation into acccount, the identity

$$\mathbf{x}_{t+1} - \mathbf{x}^{(j,k)}_{t+1|t} + \mathbf{x}^{(j,k)}_{t+1|t} = \mathbf{x}_{t+1} - \mathbf{x}^k_{t+1|T} + \mathbf{x}^k_{t+1|T}$$

implies that

$$\mathbf{P}^{(j,k)}_{t+1|t} + var(\mathbf{x}^{(j,k)}_{t+1|t}) = \mathbf{P}^k_{t+1|T} + var(\mathbf{x}^k_{t+1|T})$$

hence

$$var(\mathbf{x}^{(j,k)}_{t+1|t}) - var(\mathbf{x}^k_{t+1|T}) = \mathbf{P}^k_{t+1|T} - \mathbf{P}^{(j,k)}_{t+1|t}.$$

Finally, we have

$$\mathbf{P}^{(j,k)}_{t|T} = \mathbf{P}^j_{t|t} + \widetilde{\mathbf{P}}^{(j,k)}_t (\mathbf{P}^k_{t+1|T} - \mathbf{P}^{(j,k)}_{t+1|t})\widetilde{\mathbf{P}}^{(j,k)'}_t$$

which is formula (56). □

As $P(s_t = j|\mathbf{\Psi}_T)$ is not dependent upon $\mathbf{x}_{t|T}$, we can first calculate smoothed probabilities, and then they can be used to get smoothed values of $\mathbf{x}_t$ and $\mathbf{x}_{t|T}$. Given the above smoothing algorithms, actual smoothing can be performed by applying approximations similar to those used in the basic filtering.

Step 1. Run through the basic filter in (1.8.2) for $t = 1, \dots, T$, and store the resulting sequences

$$\mathbf{x}^{(i,j)}_{t|t-1}, \mathbf{P}^{(i,j)}_{t|t-1}, \mathbf{x}^j_{t|t}, \mathbf{P}^j_{t|t},$$

$$P(s_t = j|\mathbf{\Psi}_{t-1}) = \sum_{i=1}^{M} P(s_{t-1} = i, s_t = j|\mathbf{\Psi}_{t-1})$$

and $P(s_t = j|\mathbf{\Psi}_t)$ from formulae (43), (44), (50), (51), (52), and from Step 4 in (1.8.2), respectively, for $t = 1, \dots, T$.

Step 2. For $t = T-1, T-2, \ldots, 1$, get the smoothed joint probability $P(s_t = j, s_{t+1} = k|\mathbf{\Psi}_T)$ and $P(s_t = j|\mathbf{\Psi}_T)$ according to the initial relations in (1.8.2), and save them. Here $P(s_T = j|\mathbf{\Psi}_T)$ (the starting value for smoothing) is given by the last iteration of the basic filter.

Step 3. We use the smoothed probabilities from Step 2 to collapse the M^2 elements of $\mathbf{x}_{t|T}^{(j,k)}$ and $\mathbf{P}_{t|T}^{(j,k)}$ into M elements by taking weighted averages. At each iteration of (55) and (56) for $t = T-1, T-2, \ldots, 1$, collapse the M^2 elements into M in the following way by taking weighted averages over state s_{t+1}:

$$\mathbf{x}_{t|T}^{j} = \frac{\sum_{k=1}^{M} P(s_t = j, s_{t+1} = k|\mathbf{\Psi}_T)\mathbf{x}_{t|T}^{(j,k)}}{P(s_t = j|\mathbf{\Psi}_T)}$$

$$\mathbf{P}_{t|T}^{j} = \frac{\sum_{k=1}^{M} P(s_t = j, s_{t+1} = k|\mathbf{\Psi}_T)[P_{t|T}^{(j,k)} + (\mathbf{x}_{t|T}^{j} - \mathbf{x}_{t|T}^{(j,k)})(\mathbf{x}_{t|T}^{j} - \mathbf{x}_{t|T}^{(j,k)})']}{P(s_t = j|\mathbf{\Psi}_T)}$$

Step 4. From Step 3, the smoothed value of $\mathbf{x}_{t|T}^{j}$ is dependent upon states at time t. By taking a weighted average over the states at time t, we can get $\mathbf{x}_{t|T}$ from

$$\mathbf{x}_{t|T} = \sum_{j=1}^{M} P(s_t = j|\mathbf{\Psi}_T)\mathbf{x}_{t|T}^{j}.$$

As remarked in [6, p.119], the procedure proposed in Kim seems to work in practice, but theoretical results concerning the effects of the above approximations to optimal filtering are missing. Futher results on switching state-space models like above can be found in Billio and Montford (1995) [*Switching State Space Models. Likelihood function, Filtering and Smoothing*, CREST Working Paper].

Finally, we observe that the approach of this section permits to treat any Markov-switching ARMA(p, q) model as

$$\mathbf{y}_t = -\sum_{i=1}^{p} \boldsymbol{\phi}_i(s_t)\mathbf{y}_{t-i} + \sum_{j=1}^{q} \mathbf{\Theta}_j(s_t)\boldsymbol{\epsilon}_{t-j} + \boldsymbol{\epsilon}_t.$$

As shown in Section 1.4, it possesses a state-space representation as above:

$$\begin{cases} \mathbf{y}_t = \mathbf{C}(s_t)\mathbf{z}_t + \boldsymbol{\epsilon}_t \\ \mathbf{z}_{t+1} = \mathbf{A}(s_t)\mathbf{z}_t + \mathbf{B}\boldsymbol{\epsilon}_t \end{cases}$$

where $\mathbf{z}_t = (\mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p}, \boldsymbol{\epsilon}'_{t-1}, \ldots, \boldsymbol{\epsilon}'_{t-q})'$ is the lag vector, and the matrices $\mathbf{C}(s_t), \mathbf{A}(s_t)$ and $\mathbf{B}$ are described in Section 1.4. Now the elements $\boldsymbol{\phi}_i(s_t)$ and $\boldsymbol{\Theta}_j(s_t)$ in $\mathbf{C}(s_t)$ and $\mathbf{A}(s_t)$ are regime dependent.

1.9 Determination of the Number of Regimes

In this section we consider vector autoregressive processes with Markovian regime shifts in the intercept term (in short, MSI-AR). The case with regime changes in the mean can be treated in a similar manner. Following [6, chps.3 and 7] we illustrate finite order ARMA representations for such processes. Then we use these results to develop a strategy for selecting simultaneously the number of regimes and the order of the autoregression. Let us consider the following $n \times 1$ vector MSI-AR(p), $p \geq 0$, model

$$\mathbf{A}(L)\mathbf{y}_t = \boldsymbol{\nu}(s_t) + \mathbf{u}_t \tag{57}$$

where $\mathbf{A}(L) = \sum_{i=0}^{p} \mathbf{A}_i L^i$ with $\mathbf{A}_0 = \mathbf{I}_n$ is an $n \times n$ matrix polynomial in the lag operator (here the variance-covariance matrix of $\mathbf{u}_t$ and the autoregressive parameters $\mathbf{A}_i$ are assumed to be regime invariant). As usual, the M-state (irreducible and ergodic) Markov chain (s_t) can be represented by the AR(1) equation $\boldsymbol{\xi}_t = \mathbf{P}'\boldsymbol{\xi}_{t-1} + \mathbf{v}_t$. For $s_t = m$, $\nu_m = \nu(s_t = m)$ is $n \times 1$, and $\boldsymbol{\Lambda} = (\nu_1 \cdots \nu_M)$ is $n \times M$. The above MSI-AR(p) model has the state-space representation

$$\begin{cases} \mathbf{A}(L)\mathbf{y}_t = \boldsymbol{\Lambda}\boldsymbol{\xi}_t + \mathbf{u}_t \\ \boldsymbol{\xi}_t = \mathbf{P}'\boldsymbol{\xi}_{t-1} + \mathbf{v}_t \end{cases} \tag{58}$$

since $\nu(s_t) = \sum_{m=1}^{M} \nu_m I(s_t = m) = \mathbf{\Lambda}\boldsymbol{\xi}_t$. The transition equation in (58) differs from a stable linear AR(1) process by the fact that one eigenvalue of $\mathbf{P}$ is equal to one and the covariance matrix of $\mathbf{v}_t$ is singular, due to the adding-up restriction $\mathbf{i}_M^{'}\boldsymbol{\xi}_t = 1$. For analytical purposes, a slightly different formulation of the transition equation is more useful, where the above restriction is eliminated. This procedure alters representation (58), and we consider a new state vector $\boldsymbol{\delta}_t$ which is $(M-1)$ dimensional:

$$\boldsymbol{\delta}_t = \begin{pmatrix} \xi_{1,t} - \pi_1 \\ \vdots \\ \xi_{M-1,t} - \pi_{M-1} \end{pmatrix}$$

where $\boldsymbol{\pi} = (\pi_1 \cdots \pi_M)^{'}$ is the $M \times 1$ vector of ergodic probabilities of the Markov chain. The transition matrix, $\mathbf{F}$ say, associated with the state vector $\boldsymbol{\delta}_t$ is given by

$$\mathbf{F} = \begin{pmatrix} p_{1,1} - p_{M,1} & \cdots & p_{M-1,1} - p_{M,1} \\ \vdots & & \vdots \\ p_{1,M-1} - p_{M,M-1} & \cdots & p_{M-1,M-1} - p_{M,M-1} \end{pmatrix}$$

Then we have

$$\boldsymbol{\delta}_t = \mathbf{F}\boldsymbol{\delta}_{t-1} + \mathbf{w}_t$$

where $\mathbf{w}_t = [\mathbf{I}_{M-1} \quad -\mathbf{i}_{M-1}]\mathbf{v}_t$. The measurement equation in (58) can be reformulated as

$$\mathbf{A}(L)\mathbf{y}_t = \Lambda\boldsymbol{\pi} + \Lambda(\boldsymbol{\xi}_t - \boldsymbol{\pi}) + \mathbf{u}_t$$

Taking expectation gives

$$\mathbf{A}(L)E(\mathbf{y}_t) = \Lambda\boldsymbol{\pi} + \Lambda(E(\boldsymbol{\xi}_t) - \boldsymbol{\pi}) = \Lambda\boldsymbol{\pi}$$

since $E(\boldsymbol{\xi}_t) = \boldsymbol{\pi}$. Setting $E(\mathbf{y}_t) = \boldsymbol{\mu}_y$, we obtain

$$\mathbf{A}(L)\mathbf{y}_t = \mathbf{A}(L)\boldsymbol{\mu}_y + \Lambda(\boldsymbol{\xi}_t - \boldsymbol{\pi}) + \mathbf{u}_t$$

hence

$$\mathbf{A}(L)(\mathbf{y}_t - \boldsymbol{\mu}_y) = \mathbf{M}\boldsymbol{\delta}_t + \mathbf{u}_t$$

where $\mathbf{M} = (\nu_1 - \nu_M \cdots \nu_{M-1} - \nu_M)$ is $n \times (M-1)$. Thus we get the *unrestricted* state-space representation

$$\begin{cases} \mathbf{A}(L)(\mathbf{y}_t - \boldsymbol{\mu}_y) = \mathbf{M}\boldsymbol{\delta}_t + \mathbf{u}_t \\ \boldsymbol{\delta}_t = \mathbf{F}\boldsymbol{\delta}_{t-1} + \mathbf{w}_t \end{cases} \tag{59}$$

where $(\mathbf{w}_t)$ is a martingale difference sequence with a nonsingular covariance matrix and the innovation sequence in the measurement equation is unaltered. Setting $\mathbf{z}_t = (\boldsymbol{\delta}_t^{'} \ \mathbf{u}_t^{'})^{'}$, we get the state-space representation

$$\begin{cases} \mathbf{A}(L)(\mathbf{y}_t - \boldsymbol{\mu}_y) = [\mathbf{M} \quad \mathbf{I}_n]\mathbf{z}_t \\ \mathbf{z}_t = \begin{bmatrix} \mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{z}_{t-1} + \begin{bmatrix} \mathbf{w}_t \\ \mathbf{u}_t \end{bmatrix} \end{cases} \tag{60}$$

By [6, chp.3], any MSI-AR(p) process, represented as in (60), admits a finite order ARMA representation. For this, we need the following lemma, due to Lütkepohl (see, for example, Lemma 1 of [6],p.55)

Lemma 9.1

Suppose that $\mathbf{x}_t$ is a R-dimensional AR(p) process with $\mathbf{A}(L)\mathbf{x}_t = \mathbf{u}_t$. Let $\mathbf{G}$ be a $K \times R$ matrix of rank K. Then

$$\mathbf{y}_t = \mathbf{G}\mathbf{x}_t = \mathbf{G}\mathbf{A}(L)^{-1}\mathbf{u}_t = \mathbf{G}|\mathbf{A}(L)|^{-1}\mathbf{A}(L)^{*}\mathbf{u}_t$$

has an ARMA(p^*, q^*) representation with

$$p^* \leq deg|\mathbf{A}(L)|$$

$$q^* \leq \max_{i,j} deg\, \mathbf{A}_{ij}(L) - deg\,|\mathbf{A}(L)| + p^*$$

where deg(·) denotes the degree of a polynomial, $|\mathbf{A}(L)|$ is the determinant of $\mathbf{A}(L)$, $\mathbf{A}(L)^*$ is the adjoint of $\mathbf{A}(L)$, and $\mathbf{A}_{ij}(L)$ is the (i,j)th cofactor of $\mathbf{A}(L)$.

Proposition 9.2 *[Krolzig, Prop.2, chp.3]*

Let $\mathbf{y}_t$ denote an $n \times 1$ vector MSI-AR(p) process, $p = 0$, satisfying (60). Then there exists a final equation form ARMA(p^*, q^*) representation with $p^* = q^* \leq M - 1$:

$$\gamma(L)(\mathbf{y}_t - \boldsymbol{\mu}_y) = \mathbf{B}(L)\boldsymbol{\epsilon}_t$$

where $\boldsymbol{\epsilon}_t$ is a zero mean vector white noise process,

$$\gamma(L) = 1 - \gamma_1 L - \cdots - \gamma_{M-1} L^{M-1}$$

is the scalar AR operator, and $\mathbf{B}(L) = I_n - \mathbf{B}_1 L - \cdots - \mathbf{B}_{M-1} L^{M-1}$ is a $n \times (n + M - 1)$-dimensional matrix polynomial in the lag operator (of order $M - 1$).

Proof. Solving the transition equation in (60) gives

$$\begin{aligned}\mathbf{z}_t &= \begin{bmatrix} \sum_{i=0}^{\infty} \mathbf{F}^i \mathbf{w}_{t-i} \\ \mathbf{u}_t \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{\infty} (\mathbf{F}L)^i \mathbf{w}_t \\ \mathbf{u}_t \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{I}_{M-1} - \mathbf{F}L)^{-1} \mathbf{w}_t \\ \mathbf{u}_t \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{F}(L) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{u}_t \end{bmatrix}\end{aligned}$$

where $\mathbf{F}(L) = \mathbf{I}_{M-1} - \mathbf{F}L$ (here we have used the fact that all the eigenvalues of $\mathbf{F}$ are less than 1 in modulus). Inserting the above vector MA(∞) representation for $\mathbf{z}_t$ in the measurement equation gives

$$\mathbf{y}_t - \boldsymbol{\mu}_y = \begin{bmatrix} \mathbf{M} & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{F}(L) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{u}_t \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{M} & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{F}(L)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{u}_t \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{MF}(L)^{-1} & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{w}_t \\ \mathbf{u}_t \end{bmatrix}$$

$$= \mathbf{MF}(L)^{-1}\mathbf{w}_t + \mathbf{u}_t$$

Applying Lemma 9.1 we get the final equation form of ARMA$(M-1, M-1)$ model

$$|F(L)|(\mathbf{y}_t - \boldsymbol{\mu}_y) = \mathbf{M}F(L)^*\mathbf{w}_t + |F(L)|\mathbf{u}_t$$

Note that $p^* = q^* = M - 1$ is satisfied if the scalar lag polynomial $\gamma(L) = |F(L)|$ and $B(L)$ are coprime. □

Proposition 9.3 *[Krolzig, Prop.3, chp.3]*

Suppose that $\mathbf{y}_t$ *is an* $n \times 1$ *vector MSI-AR(p) process,* $p > 0$*, satisfying (60). Under quite general regularity conditions,* $\mathbf{y}_t$ *possess the ARMA(*$M+np-1, M+(n-1)p-1$*) representation*

$$\mathbf{C}(L)(\mathbf{y}_t - \boldsymbol{\mu}_y) = \mathbf{B}(L)\boldsymbol{\epsilon}_t$$

where $\boldsymbol{\epsilon}_t$ *is a zero mean vector white noise.*

Proof. It is a simple extension of Proposition 9.2. Consider the process $\mathbf{y}_t^* = A(L)(\mathbf{y}_t - \boldsymbol{\mu}_y)$. Since the relation $A(L)(\mathbf{y}_t - \boldsymbol{\mu}_y) = \mathbf{M}\boldsymbol{\delta}_t + \mathbf{u}_t$ holds from (59), the transformed process $\mathbf{y}_t^*$ satisfies the conditions of Lemma 9.1. This MSI-AR(0) process has the ARMA$(M-1, M-1)$ representation

$$|F(L)|\mathbf{y}_t^* = \mathbf{M}F(L)^*\mathbf{w}_t + |F(L)|\mathbf{u}_t$$

which induces the ARMA$(M+p-1, M-1)$ representation

$$|F(L)|A(L)(\mathbf{y}_t - \boldsymbol{\mu}_y) = \mathbf{M}F(L)^*\mathbf{w}_t + |F(L)|\mathbf{u}_t$$

Multiplying with the adjoint $A(L)^*$ gives the final equation form

$$|F(L)||A(L)|(\mathbf{y}_t - \boldsymbol{\mu}_y) = A(L)^*\mathbf{M}F(L)^*\mathbf{w}_t + |F(L)|A(L)^*\mathbf{u}_t$$

which is an ARMA$(M + np - 1, M + (n - 1)p - 1)$. $\square$

By Proposition 9.3 any univariate MSI-AR(p), $p > 0$, has an ARMA$(M + p - 1, M - 1)$ representation. Thus, an ARMA structure in the autocovariance function may reveal the characteristics of a data generating MSI-AR process. More precisely, the determination of the number of regimes, as well as the number of autoregressive parameters, can be based on currently available procedures to estimate the order of ARMA models. In principle, any of the existing model selection criteria may be applied for identifying M and p. For example, for specifying univariate ARMA models, one can use the classical *Box-Jenkins strategy* (see, for example, [1], chapter 6).

References

[1] GOURIEROUX C.- MONFORT A:, "Time Series and Dynamic Models", *Themes in Modern Econometrics*, Cambridge University Press, Cambridge U.K., 1997.

[2] HAMILTON J.D., "Analysis of Time Series Subject to Changes in Regime", *Journal of Econometrics* 45 (1990), 39-70.

[3] HAMILTON J.D., "Estimation, Inference and Forecasting of Time Series Subject to Changes in Regime",in *Handbook of Statistics* (Maddala G.S., Rao C.R. and Vinod H.D. eds), Vol.11, North-Holland, New York, 1993.

[4] HAMILTON J.D., "Time Series Analysis", *Princeton University Press*, Princeton, N.J., 1994.

[5] KIM C.J., "Dynamic Linear Models with Markov-Switching", *Journal of Econometrics* 60 (1994), 1-22.

[6] KROLZIG H.M., "Markov-Switching Vector Autoregressions: modelling, statistical inference and application to business cycle analysis", *Springer Verlag*, Berlin, Heidelberg, N.Y.,1997.

Chapter 2

Markov-Switching VARMA Models

2.1 Determining the Number of Regimes in Markov-Switching VAR and VMA Models

Abstract. *We give stable finite order VARMA(p^*, q^*) representations for M-state Markov switching second-order stationary time series whose autocovariances satisfy a certain matrix relation. The upper bounds for p^* and q^* are elementary functions of the dimension K of the process, the number M of regimes, the autoregressive and moving average orders of the initial model. If there is no cancellation, the bounds become equalities, and this solves the identification problem. Our classes of time series include every M-state Markov switching multivariate moving average models and autoregressive models in which the regime variable is uncorrelated with the observable. Our results include, as particular cases, those obtained by Krolzig (1997), and improve the bounds given by Zhang and Stine (2001) and Francq and Zakoïan (2001) for our classes of dynamic models. Data simulations and an application on foreign exchange rates complete the paper. [JEL Classification: C01, C32, C50, C52]*

Keywords: Second-order stationary time series, VMA models, VAR models, State-Space models, Markov chains, changes in regime, regime number.

2.1.1 Introduction

In this paper we consider dynamic models whose parameters can change as a result of a regime-shift variable, described as the outcome of an unobserved Markov chain. Such models have attracted much interest in the literature for their applications in areas as economics, statistics, and finance. A key problem arising in applications is to determine the number of Markov regimes for which a switching model gives an adequate representation of the observed data. In practice, the state dimension of the Markov chain is sometimes dictated by the actual application or it is determined in an informal manner by visual inspection of plots. However, there exists in the literature likelihood ratio test developed under non-standard conditions which help testing Markov switching models (see Hansen (1992)). The current methods for determining the state dimension are mainly based either on complexity-penalized likelihood criteria (see, for example, Psaradakis and Spagnolo (2003), Olteanu and Rynkiewicz (2007), and Ríos and Rodríguez (2008)) or on finite order VARMA representations of the initial switching models (see, for example, Krolzig (1997), Zhang and Stine (2001) and Francq and Zakoïan (2001)). The parameters of the VARMA representations can be determined by evaluating the autocovariance function of the Markov-switching models. It turns out that the above parameters are elementary functions of the dimension of the dynamic process, the number of regimes and the orders of the switching autoregressive moving-average model. As the sample autocovariances are more easily calculated than maximum (penalized) likelihood estimates of the model parameters, the bounds arising from the above-mentioned elementary functions are very useful for selecting the number of regimes and the orders of the switching moving-average autoregression. Some bounds are previously determined by Krolzig (1997), Zhang and Stine (2001) and Francq and Zakoïan (2001) for some Markov regime switching models of different type. Surprisingly, we show that the bounds given by Krolzig maintain their validity for Markov switching time series whose autocovariances satisfy a matrix relation specified in the statement of Theorem 2.2. This allows us to improve the bounds obtained by Zhang and Stine (2001) in Theorem 4 and Francq and Zakoïan (2001) in Section 4.3 for a large class of dynamic models. This class includes every multivariate regime switching Moving Average (MA) process and multivariate regime switching Autoregressive (AR) processes, in which the regime variable is uncorrelated with the observable.

2.1 Determining the Number of Regimes in Markov-Switching VAR and VMA Models

The main results of the paper are Theorems 2.2, 3.5, 4.2 and we are going to illustrate them (the specifics of the models and the regularity assumptions will be given in the next sections). The first result states that a second-order stationary dynamic process, whose autocovariances satisfy a certain matrix relation, has a stable VARMA representation whose AR and MA orders are well-specified elementary functions. The second result relates to M-state switching multivariate Moving Average model MA(q) of the type $\mathbf{y}_t = \boldsymbol{\nu}_{s_t} + \boldsymbol{\Theta}_{s_t}(L)\mathbf{u}_t$. We will show that it admits a stable VARMA(p^*, q^*) representation, whose autoregressive lag polynomial is scalar, and where the orders of the stable VARMA satisfy $p^* \leq M - 1$ and $q^* \leq M + q - 1$. Finally, the last result regards to a M-state switching K-dimensional Autoregressive model AR(p) of the type $\boldsymbol{\phi}_{s_t}(L)\mathbf{y}_t = \boldsymbol{\nu}_{s_t} + \boldsymbol{\Sigma}_{s_t}(L)\mathbf{u}_t$. Assuming that the regime variable is uncorrelated with the observable, we prove that it admits a stable VARMA(p^*, q^*) representation, whose autoregressive lag polynomial is scalar, and where the orders of the stable VARMA satisfy $p^* \leq M + Kp - 1$ and $q^* \leq M + (K-1)p - 1$. The assumption of uncorrelated regimes with the observable is, of course, satisfied when the autoregressive part is regime-invariant (as done in Krolzig (1997)). However, it is also a reasonable assumption when coinceving the change in regimes as an outside event from the economic system as, for instance, abrupt natural events or unexpected wars. This leads us to the fact that if the lag polynomials of the autoregressive and moving-average parts of the stable VARMA(p^*, q^*) are coprime, then equalities hold in the previous relations. In other words, if there is no cancellation, then the identification problem is completely solved. In any case, the above result allows us to determine a lower bound for the number of states. This means that a VARMA representation, and hence a VARMA structure in the autocovariance function, may reveal the characteristics of a data generating MS(M)-VMA and MS(M)-VAR processes. More precisely, the determination of the number of regimes, as well as the number of autoregressive (or moving-average) parameters, can be based on currently available procedures to estimate the orders of VARMA models. In principle, any of the existing model selection criteria may be applied for identifying M and p or q. In the case of MS(M)-VMA(q) model, evaluating estimates $(\widehat{p^*}, \widehat{q^*})$ of the orders of the stable VARMA(p^*, q^*) representation, we get the estimates $\widehat{M} = \widehat{p^*} + 1$ and $\widehat{q} = \widehat{q^*} - \widehat{p^*}$, for M and q, respectively, when there is no cancellation (see Theorem 3.5). Furthermore, for MS(M)-VAR(p) model, in which the regime variable is uncorrelated with the observable, having estimates $(\widehat{p^*}, \widehat{q^*})$ of the orders of the stable

VARMA(p^*, q^*) representation, gives us $\widehat{M} = K(\widehat{q^*}+1) - (K-1)(\widehat{p^*}+1)$ and $\widehat{p} = \widehat{p^*} - \widehat{q^*}$, for M and p, respectively, when there is no cancellation and the autoregressive lag polynomial of the stable representation is scalar (see Theorem 4.2). The rest of the paper is organized as follows. In Section 2 we recall a characterization of VARMA process in terms of autocovariances, as given by Zhang and Stine (2001). Then we extend Lemma 1 of the quoted paper, and give a different proofs of Krolzig's results on VARMA representations of certain MS processes (see Krolzig (1997), Propositions 2, 3, and 4, Chp. 3) and of some results of Zhang and Stine (2001) and Francq and Zakoïan (2001). In Section 3 (resp. 4) we show that an MS(M)-MA(q), $q \geq 0$, (resp. MS(M)-VAR(p), $p \geq 0$, for which the regime variable is uncorrelated with the observable) has a VARMA(p^*, q^*) representation, where the upper bounds for p^* and q^* are induced by specific elementary functions. In Section 5 we discuss the implications of our results for model selection, and illustrate some simulation experiments and numerical applications. In Section 6 we include our results on the exchange rata data from Engle and Hamilton (1990). Section 7 concludes. The proofs of theorems in Section 3 and 4 are reported in the Appendix which completes the paper.

2.1.2 VARMA Representations

In this section, we introduce the model and the basic notation concerning with it. In particular, we prove new algebraic results (Theorem 2.2 and Corollary 2.3) which will be used in the next Sections 3 and 4 for the determination of the number of regimes in Markov-switching VMA and VAR models. As a consequence, we give different and simple proofs of some results, previously obtained by several authors, for classes of Markov switching time series which are included in our models.

Let $\mathbf{y} = (\mathbf{y}_t)$ be a second-order stationary K-dimensional process. Then $\mathbf{y}$ is said to have a stable and invertible VARMA(p, q) *representation* if it satisfies a finite difference equation $\boldsymbol{\phi}(L)\mathbf{y}_t = \boldsymbol{\Theta}(L)\mathbf{u}_t$, where $\boldsymbol{\phi}(L) = \sum_{i=0}^{p} \boldsymbol{\phi}_i L^i$ and $\boldsymbol{\Theta}(L) = \sum_{j=0}^{q} \boldsymbol{\Theta}_j L^j$ are $K \times K$ matrix polynomials in the lag operator L, $\boldsymbol{\phi}_0 = \boldsymbol{\Theta}_0 = \mathbf{I}_K$, $\boldsymbol{\phi}_p \neq \mathbf{0}, \boldsymbol{\Theta}_q \neq \mathbf{0}$. Here the variables $\mathbf{y}_{-1}, \ldots, \mathbf{y}_{-p}$, $\mathbf{u}_{-1}, \ldots, \mathbf{u}_{-q}$ are assumed to be uncorrelated with $\mathbf{u}_t$ for every $t \geq 0$. The process $\mathbf{u} = (\mathbf{u}_t)$ is a zero mean white noise, i.e., $E(\mathbf{u}_t \mathbf{u}_\tau^{\prime}) = \delta_\tau^t G$ with $|G| \neq 0$ (through the paper, the symbol $|A|$ denotes the determinant of a square matrix A, and δ_τ^t denotes the Kronecker symbol, i.e., $\delta_\tau^t = 1$ if $t = \tau$ and zero otherwise). To avoid redundancy, $\boldsymbol{\phi}(L)$ and $\boldsymbol{\Theta}(L)$ are coprime, and to guarantee invertibility, we assume that the polynomials $|\boldsymbol{\phi}(z)|$ and $|\boldsymbol{\Theta}(z)|$, $z \in \mathbb{C}$, have all their roots strictly outside the unit circle. This definition implies that the orders p and q are minimal in the usual sense. Finally, the process $\mathbf{y} = (\mathbf{y}_t)$ is second-order stationary if the mean $E(\mathbf{y}_t)$ and the autocovariances $\boldsymbol{\Gamma}_{\mathbf{y}}(t,h)$ are independent of t. So we can write $\boldsymbol{\mu}_{\mathbf{y}} = E(\mathbf{y}_t)$ and $\boldsymbol{\Gamma}_{\mathbf{y}}(h) = \boldsymbol{\Gamma}_{\mathbf{y}}(t,h)$. We start with the following well-knowm result which characterizes the minimal VARMA(p,q) model in terms of its autocovariance function (through the paper we always assume that the process is not deterministic). For the proof see Zhang and Stine (2001), Theorem 1. Let L be the backward shift operator, $L\Gamma_{\mathbf{y}}(h) = \Gamma_{\mathbf{y}}(h-1)$, where $\Gamma_{\mathbf{y}}$ is the autocovariance function of the observed process $\mathbf{y} = (\mathbf{y}_t)$.

Theorem 2.1. *Suppose that the K-dimensional process $\mathbf{y} = (\mathbf{y}_t)$ is second-order stationary (or equivalently, weakly stationary) and the covariances $\Gamma_{\mathbf{y}}(h)$, $h \in \mathbb{Z}$, satisfy a finite difference equation of order p and rank $q+1$, that is, there exist $K \times K$ matrices $\mathbf{A}_i$, $i = 0, \ldots, p$, with $\mathbf{A}_0 = \mathbf{I}_K$ and $\mathbf{A}_p \neq \mathbf{0}$, such that $B(L)\Gamma_{\mathbf{y}}(h) \neq 0$ for $h = q$ and vanishes for every $h \geq q+1$, where $B(L) = \sum_{i=0}^{p} \mathbf{A}_i L^i$. Then $\mathbf{y} = (\mathbf{y}_t)$ has a VARMA(p^*, q^*) representation, where $p^* \leq p$ and $q^* \leq q$. If the pair (p,q) is minimal, then we have equalities $p^* = p$ and $q^* = q$.*

Note that the dimension K is absent from p^* and q^*. This depends on the fact that the autoregressive part of VARMA(p^*, q^*) consists of matrices, in general. In what follows, we also seek special VARMA(p^*, q^*) representation whose autoregressive part consists of scalars. This corresponds to the *final equation form* usually considered in the statements of Krolzig's book (1997); see, for example, Formula 3.10, p.58.

The following result generalizes Lemma 1 of Zhang and Stine (2001), proved for the case $p = q = 0$.

Theorem 2.2. *Suppose that the K-dimensional process $\mathbf{y} = (\mathbf{y}_t)$ is second-order stationary and the autocovariances of $\mathbf{y}$ satisfy*

$$B(L)\Gamma_{\mathbf{y}}(h) = A' Q^h B$$

for every $h \geq q \geq 0$, where all the matrices on the right hand-side are nonzero matrices, Q is $M \times M$, A and B are $M \times K$ matrices, $B(L) = \sum_{i=0}^{p} \mathbf{B}_i L^i$ is a $K \times K$ matrix polynomial of degree $p \geq 0$, with $\mathbf{B}_0 = I_K$ and $\mathbf{B}_p \neq \mathbf{0}$. Then $\mathbf{y} = (\mathbf{y}_t)$ has a stable VARMA(p^*, q^*) representation, where $p^* \leq M + p$ and $q^* \leq M + q - 1$. If we require that the autoregressive part of such a representation consists of scalars (not matrices) and assume the usual regularity conditions on the roots of the polynomial $|B(z)|$, $z \in \mathbb{C}$, to guarantee the invertibility of $B(L)$, the bounds become $p^* \leq M + Kp$ and $q^* \leq M + (K-1)p + q - 1$.

Proof. The Cayley-Hamilton theorem implies that there exist real numbers $f_1 \ldots f_M \in \mathbb{R}$ such that

$$Q^M - f_1 Q^{M-1} - \cdots - f_M I_M = \varphi_Q(Q) = \mathbf{0} \tag{2.1}$$

where $\varphi_Q(\lambda) = \lambda^M - f_1\lambda^{M-1} - \cdots - f_M$ is the characteristic polynomial of Q, that is, $\varphi_Q(\lambda) = |\lambda I_M - Q|$. The hypothesis of the statement implies the following relations

$$\begin{aligned}
\Gamma_{\mathbf{y}}(q+h+M) + \mathbf{B}_1\Gamma_{\mathbf{y}}(q+h+M-1) + \cdots + \mathbf{B}_p\Gamma_{\mathbf{y}}(q+h+M-p) &= A^{'}Q^{q+h+M}B \\
\Gamma_{\mathbf{y}}(q+h+M-1) + \mathbf{B}_1\Gamma_{\mathbf{y}}(q+h+M-2) + \cdots + \mathbf{B}_p\Gamma_{\mathbf{y}}(q+h+M-p-1) &= A^{'}Q^{q+h+M-1}B \\
&\vdots \\
\Gamma_{\mathbf{y}}(q+h) + \mathbf{B}_1\Gamma_{\mathbf{y}}(q+h-1) + \cdots + \mathbf{B}_p\Gamma_{\mathbf{y}}(q+h-p) &= A^{'}Q^{q+h}B
\end{aligned}$$

for every $h \geq 0$. Multiplying the last M lines with $-f_1, \ldots, -f_M$ and adding all equations, we get

$$\Gamma_{\mathbf{y}}(q+h+M) + \sum_{j=1}^{M+p} \mathbf{C}_j \Gamma_{\mathbf{y}}(q+h+M-j) = A^{'}Q^{q+h}\varphi_Q(Q)B = \mathbf{0} \tag{2.2}$$

for some matrices $\{\mathbf{C}_j\}$ and for every $h \geq 0$. The right hand-side of (2.2) is zero by (2.1). Formula (2.2) can be written in the equivalent form

$$\Gamma_{\mathbf{y}}(h) + \sum_{j=1}^{M+p} \mathbf{C}_j \Gamma_{\mathbf{y}}(h-j) = \mathbf{0} \tag{2.3}$$

for every $h \geq M+q$. From (2.3) and Theorem 2.1 we can conclude that

$$p^* \leq M+p \qquad \text{and} \qquad q^* \leq M+q-1.$$

To get a stable VARMA(p^*, q^*) representation whose autoregressive part consists of scalars (not matrices), we multiply (2.1) by Q^{q-p}

$$Q^{M+q-p} - f_1 Q^{M+q-p-1} - \cdots - f_M Q^{q-p} = \mathbf{0}. \tag{2.4}$$

Premultiplying (2.4) by $A^{'}$ and postmultiplying (2.4) by B yield

$$A^{'}Q^{M+q-p}B - f_1 A^{'}Q^{M+q-p-1}B - \cdots - f_M A^{'}Q^{q-p}B = \mathbf{0}$$

hence

$$B(L)\Gamma_{\mathbf{y}}(M+q-p) - f_1 B(L)\Gamma_{\mathbf{y}}(M+q-p-1) - \cdots - f_M B(L)\Gamma_{\mathbf{y}}(q-p) = \mathbf{0}$$

by using the matrix relation of the statement. Premultiplying the last equation by the adjoint $B(L)^*$ of $B(L)$ defined as $B(L)^* = |B(L)|B(L)^{-1}$ recalling that $B(L)$ is invertible by hypothesis (or equivalently, $B(L)^* B(L) = |B(L)|\mathbf{I}_K$), we get

$$|B(L)|(\Gamma_{\mathbf{y}}(M+q-p) - f_1\Gamma_{\mathbf{y}}(M+q-p-1) - \cdots - f_M\Gamma_{\mathbf{y}}(q-p)) = \mathbf{0}$$

where the polynomial $|B(L)|$ has degree Kp. Doing the matrix products term-by-term, taking in mind the definition of the operator L and collecting similar terms, we get a finite scalar difference equation of the form

$$\Gamma_{\mathbf{y}}(M+q+Kp-p) + \eta_1\Gamma_{\mathbf{y}}(M+q+Kp-p-1) + \cdots + \eta_{M+Kp}\Gamma_{\mathbf{y}}(q-p) = \mathbf{0} \tag{2.5}$$

where the coefficients $\{\eta_j\}$ are scalars. Now the last result in the statement follows from (2.5) and Theorem 2.1, that is, we get $p^* \leq M+Kp$, and $q^* \leq M+q+(K-1)p-1$. □

Corollary 2.3. *Under the hypothesis of Theorem 2.2, if the autocovariances of* $\mathbf{y} = (\mathbf{y}_t)$ *satisfy*

$$B(L)\Gamma_{\mathbf{y}}(h) = \sum_{i=1}^{r} A_i^{'} Q_i^h B_i$$

for every $h \geq q \geq 0$*, where* A_i *and* B_i *are* $M_i \times K$ *nonzero matrices,* Q_i *is an* $M_i \times M_i$ *nonzero matrix, for* $i = 1, \ldots, r$*, and* $B(L)$ *is a* $K \times K$ *matrix polynomial in* L *of degree* $p \geq 0$*, with* $\mathbf{B}_0 = \mathbf{I}_K$*. Then* $\mathbf{y} = (\mathbf{y}_t)$ *has a stable VARMA(*p^*, q^**) representation, where* $p^* \leq \sum_{i=1}^{r} M_i + p$ *and* $q^* \leq \sum_{i=1}^{r} M_i + q - 1$*. If* $B(L)$ *is invertible and we require that the autoregressive part of such a representation consists of scalars (not matrices), the bounds become* $p^* \leq \sum_{i=1}^{r} M_i + Kp$ *and* $q^* \leq \sum_{i=1}^{r} M_i + (K-1)p + q - 1$*.*

Proof. Setting $A^{'} = [A_1^{'} \, \ldots \, A_r^{'}]$, $B = [B_1^{'} \, \ldots \, B_r^{'}]^{'}$ and $Q = diag(Q_1 \, \ldots Q_r)$, we get $B(L)\Gamma_{\mathbf{y}}(h) = A^{'}Q^h B$. The result now follows from Theorem 2.2. □

To complete the section we give different proofs of Krolzig's results on VARMA representations of certain MS processes (see Krolzig (1997), Chp. 3).

Proposition 2.4. (Krolzig(1997), Proposition 2, p.56) *Let* $\mathbf{y} = (\mathbf{y}_t)$ *be a* K*-dimensional Hidden Markov-chain process (called MSI(M)-VAR(0) process in Krolzig (1997),p.50)*

$$\mathbf{y}_t = \boldsymbol{\nu}_{s_t} + \mathbf{u}_t \qquad \mathbf{u}_t \sim IID(\mathbf{0}, \boldsymbol{\Sigma}_u).$$

Then $\mathbf{y}$ *admits a stable VARMA(*p^*, q^**) representation with* $p^* = q^* \leq M - 1$*.*

Proof. The autocovariances of $\mathbf{y}$ satisfy $\Gamma_{\mathbf{y}}(h) = A^{'}F^h B$ for every $h \geq 1$, where A and B are nonzero $(M-1) \times K$ matrices, and F is $(M-1) \times (M-1)$ (see Krolzig (1997), Section 3.3.2, Formula (3.21)). Now apply Theorem 2.2 for $p = 0$, $q = 1$ and $M - 1$ instead of M. □

Proposition 2.5. (Krolzig(1997), Proposition 3, p.57) *Let* $\mathbf{y} = (\mathbf{y}_t)$ *be a* K*-dimensional* M*-state Markov switching process (called MSI(M)-VAR(p) process,* $p > 0$*, in Krolzig (1997),p.50)*

$$A(L)\mathbf{y}_t = \boldsymbol{\nu}_{s_t} + \mathbf{u}_t \qquad \mathbf{u}_t \sim IID(\mathbf{0}, \boldsymbol{\Sigma}_u).$$

where $A(L) = I_K - A_1 L - \cdots - A_p L^p$*,* $A_p \neq 0$*, is invertible and regime invariant. Then* $\mathbf{y} = (\mathbf{y}_t)$

admits a stable VARMA(p^*, q^*) representation

$$C(L)(\mathbf{y}_t - \mu_{\mathbf{y}}) = B(L)\boldsymbol{\epsilon}_t$$

where $p^* \leq M + p - 1$ and $q^* \leq M - 1$, $C(L)$ is a $K \times K$ dimensional lag polynomial of order $M + p - 1$, $B(L)$ is a $K \times K$ dimensional lag polynomial of order $M - 1$, and $\boldsymbol{\epsilon}_t$ is a zero mean vector white noise process. If we require that the autoregressive part of such a stable representation is scalar, the bounds become $p^* \leq M + Kp - 1$ and $q^* \leq M + (K-1)p - 1$, as pointed out in Krolzig (1997), p.58, Formula (3.10).

Proof. The autocovariances of $\mathbf{y}$ satisfy $\Gamma_{\mathbf{y}}(h) - \sum_{j=1}^{p} A_j \Gamma_{\mathbf{y}}(h-j) = A' F^h B$ for every $h \geq 1$, where A and B are nonzero $(M-1) \times K$ matrices, and F is $(M-1) \times (M-1)$ (see Krolzig (1997), Section 3.3.4, Formula (3.27)). The result now follows from Theorem 2.2 for $q = 1$ and $M - 1$ instead of M. □

Proposition 2.6. (Krolzig(1997), Proposition 4, p.58) *Let $\mathbf{y} = (\mathbf{y}_t)$ be a K-dimensional M-state Markov switching process (called MSM(M)-VAR(p) process, $p > 0$, in Krolzig (1997),p.50)*

$$A(L)(\mathbf{y}_t - \boldsymbol{\mu}_{s_t}) = \mathbf{u}_t \qquad \mathbf{u}_t \sim IID(\mathbf{0}, \boldsymbol{\Sigma}_u).$$

where $A(L)$ is as above. Then there exists a final equation form VARMA(p^*, q^*) representation

$$\gamma(L)(\mathbf{y}_t - \mu_{\mathbf{y}}) = B(L)\boldsymbol{\epsilon}_t$$

where $p^* \leq M + Kp - 1$ and $q^* \leq M + Kp - 2$, $\gamma(L)$ is a scalar lag polyonomial of order $M + Kp - 1$, $B(L)$ is a $(K \times K)$ dimensional lag polynomial of order $M + Kp - 2$, and $\boldsymbol{\epsilon}_t$ is a zero mean vector white noise process.

Proof. The autocovariances of $\mathbf{y}$ satisfy $\Gamma_{\mathbf{y}}(h) = V' F^h W + X' A^h Z$, for every $h \geq 0$, where V and W are nonzero $(M-1) \times K$ matrices, X and Z are nonzero $(Kp) \times K$ nonzero matrices, F is $(M-1) \times (M-1)$ and A is $(Kp) \times (Kp)$ (see Krolzig (1997), Section 3.3.3, Formula (3.24)). Now apply Corollary 2.3 by setting $r = 2$ and $p = q = 0$ in the last statement of it. □

Note that there is a typographic error in the statement of Proposition 4, Krolzig (1997), p.58, that is, one reads $q^* \leq M - Kp - 2$. However, the first minus sign in this inequality is a typo as remarked at line -4, p.58, in the same reference.

For completeness, we also recall the bounds for the regime number obtained by Francq and Zakoïan (2001) in Section 4.3. Here we give a different proof of their result by using Theorem 2.2 above.

Proposition 2.7. (Francq and Zakoïan (2001), Sec. 4.3) *Let $\mathbf{y} = (\mathbf{y}_t)$ be a K-dimensional second-order stationary centered dynamic process which satisfies the following MS(M) VARMA(p, q) model (in the notation of the quoted paper)*

$$\mathbf{y}_t = \sum_{i=1}^{p} a_i(s_t)\mathbf{y}_{t-i} + \boldsymbol{\epsilon}_t + \sum_{j=1}^{q} b_j(s_t)\boldsymbol{\epsilon}_{t-j}$$

where $a_i(s_t)$ and $b_j(s_t)$ are $K \times K$ random matrices, $\boldsymbol{\epsilon}_t = \sigma(s_t)\boldsymbol{\eta}_t$, where $\sigma(s_t)$ is a $K \times K$ random matrix and $\boldsymbol{\eta}_t$ is a centered white noise with $E(\boldsymbol{\eta}_t\boldsymbol{\eta}_\tau^{'}) = \delta_\tau^t \Omega$ (Ω non singular). Then $\mathbf{y} = (\mathbf{y}_t)$ admits a stable VARMA(p^*, q^*) representation, where $p^*, q^* \leq MK(p+q)$.

Proof. It was shown in Francq and Zakoïan, Section 4.3, that the autocovariance of $\mathbf{y} = (\mathbf{y}_t)$ is given by $\Gamma_{\mathbf{y}}(h) = E(\mathbf{y}_t\mathbf{y}_{t-h}^{'}) = (\mathbf{e}^{'} \otimes \mathbf{f}^{'})(P^*)^h W(0)\mathbf{f}$, for every $h > 0$, where $(\mathbf{e}^{'} \otimes \mathbf{f}^{'})$ and $W(0)\mathbf{f}$ are nontrivial $K \times [MK(p+q)]$ and $[MK(p+q)] \times K$ matrices, respectively, and P^* is $[MK(p+q)] \times [MK(p+q)]$. Now we apply Theorem 2.2 where on the right side $Q = P^*$ is a square matrix of order $MK(p+q)$ and on the left side $B(L) = \mathbf{I}_K$, i.e., we must set $p = 0$ and $q = 1$ according to the notation in the statement of Theorem 2.2. □

Finally, we obtain the main result in Zhang and Stine (2001) by using Theorem 2.2 above.

Proposition 2.8. (Zhang and Stine (2001), Theorem 4) *Let $\mathbf{y} = (\mathbf{y}_t)$ be a K-dimensional second-order stationary Markov regime switching VAR(p)*

$$\mathbf{y}_t = A_{s_t}^{(1)}\mathbf{y}_{t-1} + \cdots + A_{s_t}^{(p)}\mathbf{y}_{t-p} + \boldsymbol{\Sigma}_{s_t}\mathbf{v}_t$$

where $\mathbf{v}_t \sim IID(\mathbf{0}, \mathbf{I}_K)$, $\boldsymbol{\Sigma}_{s_t}$ is a $K \times K$ positive definitive matrix, $A_{s_t}^{(i)}$ is $K \times K$ and $\{s_t\}$ is independent of $\{\mathbf{v}_t\}$. Then $\mathbf{y} = (\mathbf{y}_t)$ admits a stable VARMA(p^*, q^*) representation, where $p^* \leq M(Kp)^2$ and $q^* \leq M(Kp)^2 - 1$.

Proof. It was shown in Zhang and Stine (2001), Formula (29), that the autocovariance of $\mathbf{y} = (\mathbf{y}_t)$ is given by $\operatorname{vec}\Gamma_{\mathbf{y}}(h) = \mathbf{Q}F_1^h\mathbf{R}$, for every $h \geq 0$, where $\mathbf{Q}$ and $\mathbf{R}$ are nontrivial $K^2 \times [M(Kp)^2]$ and $[M(Kp)^2] \times 1$ matrices, respectively, and F_1 is $[M(Kp)^2] \times [M(Kp)^2]$. Now we apply Theorem 2.2 where on the right side $Q = F_1$ is a square matrix of order $M(Kp)^2$ and on the left side $B(L) = \mathbf{I}_K$, i.e., we must set $p = 0$ and $q = 0$ according to the notation in the statement of Theorem 2.2. □

In the next sections, we improve the bounds given by the previous authors, and show that the bounds given by Krolzig (1997) maintain their validity in the general case of MS(M)- VMA(q) models and for the class of MS(M)-VAR(p) models in which the regime variable is uncorrelated with the observable.

2.1.3 Markov Switching Moving Average Models

In this section we consider Markov-switching models with the following moving-average form (in short, MS(M)-MA(q)):

$$\mathbf{y}_t = \boldsymbol{\nu}_{s_t} + \mathbf{\Theta}_{s_t}(L)\mathbf{u}_t \tag{3.1}$$

Here we allows Markovian shifts in the intercept term; the case with regime changes in the mean can be treated in a similar manner. As usual, $\mathbf{y} = (\mathbf{y}_t)$ is a K-dimensional random process, $\mathbf{\Theta}_{s_t}(L) = \sum_{j=0}^{q} \mathbf{\Theta}_{s_t,j}L^j$, with $\mathbf{\Theta}_{s_t,0} = \mathbf{\Sigma}_{s_t}$ (nonsingular symmetric $K \times K$ matrix) and $\mathbf{\Theta}_{s_t,q} \neq \mathbf{0}$. The process $\mathbf{u} = (\mathbf{u}_t)$ is a zero mean white noise with $E(\mathbf{u}_t\mathbf{u}_\tau^{'}) = \delta_\tau^t\mathbf{I}_K$. The M-state Markov chain $s = (s_t)$ is irreducible, stationary and ergodic with transition matrix $\mathbf{P} = (p_{ij})$, where $p_{ij} = P(s_{t+1} = j|s_t = i)$, and stationary distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)^{'}$. Irreducibility implies that $\pi_m > 0$, for $m = 1, \ldots, M$, meaning that all unobservable states are possible. As remarked in Francq and Zakoïan (2001), Example 2, p.351, a Markov-switching moving-average process is always second-order stationary. It is sufficient to observe that the terms $\boldsymbol{\nu}_{s_t}$ and $\mathbf{\Theta}_{s_t,j}\mathbf{u}_{t-j}$, $j = 0, \ldots, q$ in (3.1) belong to the space of square summable vector function L^2. The Markov chain follows an AR(1) model

$$\boldsymbol{\xi}_t = \mathbf{P}^{'}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t \tag{3.2}$$

where $\boldsymbol{\xi}_t$ is the random $M \times 1$ vector whose mth element is equal to 1 if $s_t = m$ and zero otherwise. The innovation $\mathbf{v} = (\mathbf{v}_t)$ is a zero mean martingale difference sequence with respect to an increasing σ-field (for more details, see Krolzig (1997), p.34). By direct computations, we have

$$\begin{aligned} &E(\boldsymbol{\xi}_t) = \boldsymbol{\pi} \\ &E(\boldsymbol{\xi}_t\boldsymbol{\xi}^{'}_{t+h}) = \mathbf{D}\mathbf{P}^h \\ &E(\mathbf{v}_t\mathbf{v}^{'}_{\tau}) = \delta^t_{\tau}(\mathbf{D} - \mathbf{P}^{'}\mathbf{D}\mathbf{P}) \end{aligned} \tag{3.3}$$

where $\mathbf{D} = diag(\pi_1, \ldots, \pi_M)$ and $h \geq 0$ (here, and in the sequel, we use the convention that $A^h = \mathbf{I}$, identity matrix, if $h = 0$ for every square matrix A). We also assume that $(s_t, \mathbf{u}_t)$ is a strictly stationary process defined on some probability space, and that (s_t) is independent of $(\mathbf{u}_t)$. Our formulation includes the Hidden Markov chain processes of Krolzig (1997), Chp.3, and the Markov mean-variance switching models of Zhang and Stine (2001), Section 3.1, which is the case $q = 0$. Setting $\boldsymbol{\Lambda} = (\boldsymbol{\nu}_1 \ldots \boldsymbol{\nu}_M)$ and $\boldsymbol{\Theta}_j = (\boldsymbol{\Theta}_{1,j} \ldots \boldsymbol{\Theta}_{M,j})$ for $j = 0, \ldots, q$, where $\boldsymbol{\Theta}_0 = \boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1 \ldots \boldsymbol{\Sigma}_M)$, the process $\mathbf{y} = (\mathbf{y}_t)$ in (3.1) admits the following state-space representation

$$\begin{aligned} &\mathbf{y}_t = \boldsymbol{\Lambda}\boldsymbol{\xi}_t + \sum_{j=0}^{q} \boldsymbol{\Theta}_j(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)L^j\mathbf{u}_t \\ &\boldsymbol{\xi}_t = \mathbf{P}^{'}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t. \end{aligned} \tag{3.4}$$

Taking expectation gives $\boldsymbol{\mu}_{\mathbf{y}} = E(\mathbf{y}_t) = \boldsymbol{\Lambda}E(\boldsymbol{\xi}_t) = \boldsymbol{\Lambda}\boldsymbol{\pi}$ as $E(\boldsymbol{\xi}_t) = \boldsymbol{\pi}$. In the next theorem we compute the autocovariance function of the process $\mathbf{y} = (\mathbf{y}_t)$. This extends Theorem 3 from Zhang and Stine (2001) proved for the case $q = 0$.

Theorem 3.1. *The autocovariance function of the process* $\mathbf{y} = (\mathbf{y}_t)$ *in (3.1) is given by*

$$i) \qquad \Gamma_{\mathbf{y}}(h) = \mathbf{\Lambda}(\mathbf{Q}^{'})^h \mathbf{D}(\mathbf{I}_M - \delta_h^0 \mathbf{P}_\infty)\mathbf{\Lambda}^{'} + \sum_{j=h}^{q} \mathbf{\Theta}_j((\mathbf{P}^{'})^h \mathbf{D} \otimes \mathbf{I}_K)\mathbf{\Theta}_{j-h}^{'}$$

for $h = 0, \ldots, q$; *and*

$$ii) \qquad \Gamma_{\mathbf{y}}(h) = \mathbf{\Lambda}(\mathbf{Q}^{'})^h \mathbf{D}\mathbf{\Lambda}^{'} \quad \text{for every} \quad h \geq q+1$$

where $\mathbf{Q} = \mathbf{P} - \mathbf{P}_\infty$, $\mathbf{P}_\infty = \lim_n \mathbf{P}^n = \mathbf{i}_M \boldsymbol{\pi}^{'}$ *and* $\mathbf{i}_M = (1, 1, \ldots, 1)^{'}$.

Theorem 3.2. *Assume* $\mathbf{\Lambda} \neq \mathbf{0}$. *Then the process* $\mathbf{y} = (\mathbf{y}_t)$ *in (3.1) has a VARMA*(p^*, q^*) *representation, where* $p^* \leq M - 1$ *and* $q^* \leq M + q - 1$.

Note that the remaining case $\mathbf{\Lambda} = \mathbf{0}$ (and hence $\widetilde{\mathbf{\Lambda}} = \mathbf{0}$) will be included in the next Theorem 3.5. Now we use an argument discussed in Krolzig (1997), Section 2.3. The transition equation in (3.4) differs from a stable linear AR(1) process by the fact that one eigenvalue of $\mathbf{P}^{'}$ is equal to one and the covariance matrix of $\mathbf{v}_t$ is singular, due to the adding-up restriction $\mathbf{i}_M^{'}\boldsymbol{\xi}_t = 1$. For analytical purposes, a slightly different formulation of the transition equation is more useful, where the above restriction is eliminated. This procedure alters representation (3.4), and we consider a new state $(M-1)$-dimensional vector defined by $\boldsymbol{\delta}_t = (\xi_{1,t} - \pi_1 \ \ldots \ \xi_{M-1,t} - \pi_{M-1})^{'}$. The transition matrix, $\mathbf{F}$ say, associated with the state vector $\boldsymbol{\delta}_t$ is given by

$$\mathbf{F} = \begin{pmatrix} p_{11} - p_{M1} & \cdots & p_{M-1,1} - p_{M1} \\ \vdots & & \vdots \\ p_{1,M-1} - p_{M,M-1} & \cdots & p_{M-1,M-1} - p_{M,M-1} \end{pmatrix}$$

which is an $(M-1) \times (M-1)$ nonsingular matrix with all eigenvalues inside the unit circle. Then we have

$$\boldsymbol{\delta}_t = \mathbf{F}\boldsymbol{\delta}_{t-1} + \mathbf{w}_t \qquad (3.5)$$

where $\mathbf{w}_t = [\mathbf{I}_{M-1} \quad -\mathbf{i}_{M-1}]\mathbf{v}_t$. By direct computations, we have

$$
\begin{aligned}
& E(\boldsymbol{\delta}_t) = \mathbf{0} \\
& E(\boldsymbol{\delta}_t \boldsymbol{\delta}'_{t+h}) = \widetilde{\mathbf{D}}(\mathbf{F}')^h \\
& E(\mathbf{w}_t \mathbf{w}'_\tau) = \delta_\tau^t(\widetilde{\mathbf{D}} - \mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}')
\end{aligned}
\tag{3.6}
$$

where $\widetilde{\mathbf{D}} = \mathbf{A}\mathbf{D}(\mathbf{I}-\mathbf{P}_\infty)\mathbf{A}'$ and $\mathbf{A} = [\mathbf{I}_{M-1} \quad \mathbf{o}_{M-1}]$ is $(M-1)\times M$ (here $\mathbf{o}_{M-1}$ is the $(M-1)\times 1$ vector of zeros). More explicitly, we get

$$
\widetilde{\mathbf{D}} = \begin{pmatrix}
\pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_{M-1} \\
-\pi_1\pi_2 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_{M-1} \\
\vdots & \vdots & & \vdots \\
-\pi_{M-1}\pi_1 & -\pi_{M-1}\pi_2 & \cdots & \pi_{M-1}(1-\pi_{M-1})
\end{pmatrix}.
$$

We can see that $|\widetilde{\mathbf{D}}| = |\mathbf{D}| = \pi_1\pi_2\cdots\pi_M \neq 0$ as the Markov chain is irreducible. Now the measurement equation in (3.4) can be reformulated as

$$
\mathbf{y}_t = \boldsymbol{\Lambda}\boldsymbol{\pi} + \boldsymbol{\Lambda}(\boldsymbol{\xi}_t - \boldsymbol{\pi}) + \sum_{j=0}^{q} \boldsymbol{\Theta}_j[(\boldsymbol{\xi}_t - \boldsymbol{\pi}) \otimes \mathbf{I}_K]\mathbf{u}_{t-j} + \sum_{j=0}^{q} \boldsymbol{\Theta}_j(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_{t-j}.
$$

Then the process $\mathbf{y} = (\mathbf{y}_t)$ in (3.1) admits a second state-space representation

$$
\begin{aligned}
\mathbf{y}_t &= \boldsymbol{\Lambda}\boldsymbol{\pi} + \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\delta}_t + \sum_{j=0}^{q} \widetilde{\boldsymbol{\Theta}_j}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_{t-j} + \sum_{j=0}^{q} \boldsymbol{\Theta}_j(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_{t-j} \\
\boldsymbol{\delta}_t &= \mathbf{F}\boldsymbol{\delta}_{t-1} + \mathbf{w}_t
\end{aligned}
\tag{3.7}
$$

where $\widetilde{\boldsymbol{\Lambda}} = (\boldsymbol{\nu}_1 - \boldsymbol{\nu}_M \ \cdots \ \boldsymbol{\nu}_{M-1} - \boldsymbol{\nu}_M)$ and $\widetilde{\boldsymbol{\Theta}_j} = (\boldsymbol{\Theta}_{1,j} - \boldsymbol{\Theta}_{M,j} \ \cdots \ \boldsymbol{\Theta}_{M-1,j} - \boldsymbol{\Theta}_{M,j})$ for every $j = 0, \ldots, q$. Equation (3.7) is also called the *unrestricted* state-space representation of $\mathbf{y}$, where $\mathbf{w} = (\mathbf{w}_t)$ is a martingale difference sequence with a nonsingular covariance matrix and the innovation sequence in the measurement equation is unaltered. Note that (3.7) can be written in short as

$$
\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\delta}_t + \sum_{j=0}^{q} \widetilde{\boldsymbol{\Theta}_j}[(\boldsymbol{\delta}_t + \widetilde{\boldsymbol{\pi}}) \otimes \mathbf{I}_K]L^j\mathbf{u}_t
$$

where $\boldsymbol{\mu}_{\mathbf{y}} = E(\mathbf{y}_t) = \boldsymbol{\Lambda}\boldsymbol{\pi}$ and $\widetilde{\boldsymbol{\pi}} = (\pi_1 - \pi_M \ \cdots \ \pi_{M-1} - \pi_M)^{'}$. Using representation (3.7) and doing computations similar to those in the proof of Theorem 3.1, we get

Theorem 3.3. *The autocovariance function of the process* $\mathbf{y} = (\mathbf{y}_t)$ *in (3.1) is given by*

$$i) \quad \Gamma_{\mathbf{y}}(h) = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^h\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Lambda}}^{'} + \sum_{j=h}^{q}\widetilde{\boldsymbol{\Theta}}_j[(\mathbf{F}^h\widetilde{\mathbf{D}})\otimes\mathbf{I}_K]\widetilde{\boldsymbol{\Theta}}_{j-h}^{'} + \sum_{j=h}^{q}\boldsymbol{\Theta}_j[(\mathbf{D}\mathbf{P}_\infty)\otimes\mathbf{I}_K]\boldsymbol{\Theta}_{j-h}^{'}$$

for $h = 0,\ldots,q$; *and*

$$ii) \quad \Gamma_{\mathbf{y}}(h) = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^h\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Lambda}}^{'} \qquad \text{for every} \qquad h \geq q+1.$$

Now $\Gamma_{\mathbf{y}}(h) = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^h\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Lambda}}^{'}$ for $h \geq q+1$ is in the form of Theorem 2.2, with $p = 0$ and $q+1$ instead of q. Since $\mathbf{F}$ is $(M-1)\times(M-1)$, we can apply directly Theorem 2.2 to get an alternative proof of Theorem 3.2, assuming that $\widetilde{\boldsymbol{\Lambda}} \neq \mathbf{0}$. However, the next two results are valid in the general case.

Theorem 3.4. *The process* $\mathbf{y} = (\mathbf{y}_t)$ *in (3.1) admits the MA(∞) representation*

$$\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}} = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}(L)^{-1}\mathbf{w}_t + \sum_{j=0}^{q}\widetilde{\boldsymbol{\Theta}_j}[(\mathbf{F}(L)^{-1}\mathbf{w}_t)\otimes\mathbf{I}_K]L^j\mathbf{u}_t + \sum_{j=0}^{q}\boldsymbol{\Theta}_j(\boldsymbol{\pi}\otimes\mathbf{I}_K)L^j\mathbf{u}_t$$

where $\mathbf{F}(L) = \mathbf{I}_{M-1} - \mathbf{F}L$.

Now we compute explicitly a VARMA representation for the process $\mathbf{y} = (\mathbf{y}_t)$ in (3.1). The final equation form of the stable representation could be very useful also when dealing with inference problems. Moreover, this gives a new proof of Theorem 3.2 and generalizes Proposition 2 from Krolzig (1997), Section 3.2.3.

Theorem 3.5. *The process* $\mathbf{y} = (\mathbf{y}_t)$ *in (3.1) admits a final equation form VARMA(p^*, q^*) representation with* $p^* \leq M-1$ *and* $q^* \leq M+q-1$

$$\gamma(L)(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = \mathbf{C}(L)\boldsymbol{\epsilon}_t$$

where $\gamma(L) = |F(L)|$ *is the scalar AR operator of degree* $M-1$, $\mathbf{C}(L)$ *is the matrix lag polynomial* *of degree $M+q-1$ given by*

$$\mathbf{C}(L) = [\widetilde{\mathbf{\Lambda}}\mathbf{F}^*(L) \quad \widetilde{\mathbf{\Theta}}_0(F^*(L)\otimes\mathbf{I}_K) \quad \cdots \quad \widetilde{\mathbf{\Theta}}_q(F^*(L)\otimes\mathbf{I}_K)L^q \quad |\mathbf{F}(L)|\sum_{j=0}^{q}\mathbf{\Theta}_j(\boldsymbol{\pi}\otimes\mathbf{I}_K)L^j]$$

and $\boldsymbol{\epsilon}_t = (\mathbf{w}_t^{'} \quad \mathbf{u}_t^{'}(\mathbf{w}_t^{'}\otimes\mathbf{I}_K)\cdots\mathbf{u}_t^{'}(\mathbf{w}_{t+q}^{'}\otimes\mathbf{I}_K) \quad \mathbf{u}_t^{'})^{'}$ *is a zero mean vector white noise with* $var(\boldsymbol{\epsilon}_t) = diag(\widetilde{\mathbf{D}}-\mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}^{'} \quad (\widetilde{\mathbf{D}}-\mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}^{'})\otimes\mathbf{I}_K \quad \cdots \quad (\widetilde{\mathbf{D}}-\mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}^{'})\otimes\mathbf{I}_K \quad \mathbf{I}_K)$. *Here* $\mathbf{F}^*(L)$ *is the adjoint of* $\mathbf{F}(L) = \mathbf{I}_{M-1}-\mathbf{F}L$. *Note that* $p^* = M-1$ *and* $q^* = M+q-1$ *are satisfied if* $\gamma(L)$ *and* $\mathbf{C}(L)$ *are coprime, so the identification problem is completely solved, that is,* $M = p^*+1$ *and* $q = q^*-p^*$ *(hence* $q^* \geq p^*$ *in this case).*

To end the section we treat the forecasting for the above Markov switching moving average model. Predictions of Markov switching VARMA models can be based on the state–space representations obtained above. By ignoring the parameter estimation problem, i.e., the fact that the parameters of the multivariate Markov switching process are unknown and must therefore be estimated, the mean squared prediction error optimal forecast can be generated by the conditional expectation $\widehat{\mathbf{y}}_{t+h|t} = E(\mathbf{y}_{t+h}|\mathbf{Y}_t)$, for $h\geq 1$, where $\mathbf{Y}_t = (\mathbf{y}_t^{'}, \mathbf{y}_{t-1}^{'}, \dots)^{'}$. From (3.2) we get the forecast of the hidden Markov chain, that is, $\widehat{\boldsymbol{\xi}}_{t+h|t} = (\mathbf{P}^{'})^h\widehat{\boldsymbol{\xi}}_{t|t}$, for $h\geq 1$, where $\widehat{\boldsymbol{\xi}}_{t|t} = E(\boldsymbol{\xi}_t|\mathbf{Y}_t)$. Equivalently, from (3.5) we obtain $\widehat{\boldsymbol{\delta}}_{t+h|t} = \mathbf{F}^h\widehat{\boldsymbol{\delta}}_{t|t}$, for $h\geq 1$, where $\widehat{\boldsymbol{\delta}}_{t|t}$ is the $(M-1)$ vector formed by the columns, but the last one, of $\widehat{\boldsymbol{\xi}}_{t|t}-\boldsymbol{\pi}$. Inserting this formula into Equation (3.7) gives the h–step predictor in the case of MS(M) MA(q) models. More precisely, we have

$$\begin{aligned}\widehat{\mathbf{y}}_{t+h|t} &= E(\mathbf{\Lambda}\boldsymbol{\pi}+\widetilde{\mathbf{\Lambda}}\boldsymbol{\delta}_{t+h}+\sum_{j=0}^{q}\widetilde{\mathbf{\Theta}}_j(\boldsymbol{\delta}_{t+h}\otimes\mathbf{I}_K)\mathbf{u}_{t+h-j}+\sum_{j=0}^{q}\mathbf{\Theta}_j(\boldsymbol{\pi}\otimes\mathbf{I}_K)\mathbf{u}_{t+h-j}|\mathbf{Y}_t)\\ &= \mathbf{\Lambda}\boldsymbol{\pi}+\widetilde{\mathbf{\Lambda}}\widehat{\boldsymbol{\delta}}_{t+h|t} = \mathbf{\Lambda}\boldsymbol{\pi}+\widetilde{\mathbf{\Lambda}}\mathbf{F}^h\widehat{\boldsymbol{\delta}}_{t|t}.\end{aligned}$$

Since the eigenvalues of $\mathbf{F}$ are all inside the unit circle, the forecasts of $\mathbf{y}_{t+h}$ converge to the unconditional mean of the process as h goes to infinity, that is, we have

$$\lim_{h\to\infty}\widehat{\mathbf{y}}_{t+h|t} = \mathbf{\Lambda}\boldsymbol{\pi} = \boldsymbol{\mu}_{\mathbf{y}}.$$

Our prediction formula $\widehat{\mathbf{y}}_{t+h|t}$ is based on the quantity $\widehat{\boldsymbol{\delta}}_{t|t}$, or equivalently, on $\widehat{\boldsymbol{\xi}}_{t|t}$. It is known that $\widehat{\boldsymbol{\xi}}_{t|t}$ can be computed by iterating on the following pair of recursive formulae

$$\widehat{\boldsymbol{\xi}}_{t|t} = \frac{\boldsymbol{\eta}_t \odot \widehat{\boldsymbol{\xi}}_{t|t-1}}{\boldsymbol{\eta}_t' \widehat{\boldsymbol{\xi}}_{t|t-1}} \qquad \widehat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{P}' \widehat{\boldsymbol{\xi}}_{t|t}$$

where the symbol $\odot$ denotes the element-by-element multiplication and $\boldsymbol{\eta}_t$ is the $(M \times 1)$ vector whose jth component is the conditional density of $\mathbf{y}_t$ given $s_t = j$ and $\mathbf{Y}_{t-1}$. See, for example, Krolzig (1997), Chp.5, Formulae (5.4) and (5.5). The iteration is started by assuming that the initial state vector is drawn from the stationary unconditional probability distribution of the Markov chain, that is, $\widehat{\boldsymbol{\xi}}_{1|0} = \boldsymbol{\pi}$.

2.1.4 Markov Switching Autoregressive Models

Let $\mathbf{y} = (\mathbf{y}_t)$ be a K-dimensional second-order stationary dynamic process satisfying the following Markov switching autoregressive model (in short, MS(M)-VAR(p)):

$$\boldsymbol{\phi}_{s_t}(L)\mathbf{y}_t = \boldsymbol{\nu}_{s_t} + \boldsymbol{\Sigma}_{s_t}\mathbf{u}_t \tag{4.1}$$

where $\mathbf{u}_t \sim IID(\mathbf{0}, \mathbf{I}_K)$ and $\boldsymbol{\phi}_{s_t}(L) = \sum_{i=0}^{p} \boldsymbol{\phi}_{s_t,i} L^i$ with $\boldsymbol{\phi}_{s_t,0} = \mathbf{I}_k$ and $|\boldsymbol{\phi}_{s_t,p}| \neq 0$. As usual, we assume that the polynomials $|\boldsymbol{\phi}_{s_t}(z)|$ have all their roots strictly outside the unit circle. Sufficient conditions ensuring second-order stationarity for Markov-switching VAR models and Markov-switching VARMA models can be found, for example, in Karlsen (1990a and 1990b) and Francq and Zakoïan (2001), respectively. Moreover, in Francq and Zakoïan (2002) it was shown that, under appropriate moment conditions, the powers of the stationary solutions admit weak ARMA representations, which are potentially useful for statistical applications. Define

$$\boldsymbol{\Lambda} = (\boldsymbol{\nu}_1 \cdots \boldsymbol{\nu}_M) \qquad \boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1 \cdots \boldsymbol{\Sigma}_M)$$

and

$$\boldsymbol{\phi}(L) = (\sum_{i=0}^{p} \boldsymbol{\phi}_{1,i} L^i \cdots \sum_{i=0}^{p} \boldsymbol{\phi}_{M,i} L^i).$$

Then the process $\mathbf{y} = (\mathbf{y}_t)$ in (4.1) admits the following state-space representation

$$\begin{aligned} &\boldsymbol{\phi}(L)(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{y}_t = \boldsymbol{\Lambda}\boldsymbol{\xi}_t + \boldsymbol{\Sigma}(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{u}_t \\ &\boldsymbol{\xi}_t = \mathbf{P}^{'}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t. \end{aligned} \tag{4.2}$$

Taking expectation gives $\boldsymbol{\phi}(1)(\boldsymbol{\pi} \otimes \mathbf{I}_K)\boldsymbol{\mu}_{\mathbf{y}} = \boldsymbol{\Lambda}\boldsymbol{\pi}$. Assuming the invertibility of the $K \times K$ matrix $R = \boldsymbol{\phi}(1)(\boldsymbol{\pi} \otimes \mathbf{I}_K)$, we can write $\boldsymbol{\mu}_{\mathbf{y}} = R^{-1}\boldsymbol{\Lambda}\boldsymbol{\pi}$. Set $\mathbf{x}_t = \boldsymbol{\Lambda}\boldsymbol{\xi}_t + \boldsymbol{\Sigma}(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{u}_t$. For every $h \geq 0$ and assuming that the regime variable $\boldsymbol{\xi}_{t+h}$ is uncorrelated with $\mathbf{y}_t$, we have

$$\begin{aligned} cov(\mathbf{x}_{t+h}, \mathbf{y}_t) &= cov(\boldsymbol{\phi}(L)(\boldsymbol{\xi}_{t+h} \otimes \mathbf{I}_K)\mathbf{y}_{t+h}, \mathbf{y}_t) \\ &= \boldsymbol{\phi}(L)[E(\boldsymbol{\xi}_{t+h}) \otimes cov(\mathbf{y}_{t+h}, \mathbf{y}_t)] \\ &= \boldsymbol{\phi}(L)(\boldsymbol{\pi} \otimes \mathbf{I}_K)[1 \otimes cov(\mathbf{y}_{t+h}, \mathbf{y}_t)] \\ &= B(L)\Gamma_{\mathbf{y}}(h) \end{aligned}$$

where $B(L) = \boldsymbol{\phi}(L)(\boldsymbol{\pi} \otimes \mathbf{I}_K)$ is a $K \times K$ matrix lag polynomial of degree p. By explicit computations, we can see that $B(L) = \sum_{i=0}^{p} B_i L^i$, with $B_0 = \mathbf{I}_K$, where $B_i = \boldsymbol{\phi}_i(\boldsymbol{\pi} \otimes \mathbf{I}_K)$ is $K \times K$ and $\boldsymbol{\phi}_i = (\boldsymbol{\phi}_{1,i} \cdots \boldsymbol{\phi}_{M,i})$ is $K \times (KM)$ for every $i = 1, \ldots, p$. As done in Section 3, we can substitute $\boldsymbol{\xi}_t$ with the state $(M-1) \times 1$ vector $\boldsymbol{\delta}_t$ in order to obtain the unrestricted state-space representation

$$\begin{aligned} \widetilde{\boldsymbol{\phi}(L)}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{y}_t + \boldsymbol{\phi}(L)(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{y}_t &= \boldsymbol{\Lambda}\boldsymbol{\pi} + \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\delta}_t + \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t \\ \boldsymbol{\delta}_t &= \mathbf{F}\boldsymbol{\delta}_{t-1} + \mathbf{w}_t \end{aligned} \tag{4.4}$$

where $\widetilde{\boldsymbol{\Lambda}} = (\boldsymbol{\nu}_1 - \boldsymbol{\nu}_M \cdots \boldsymbol{\nu}_{M-1} - \boldsymbol{\nu}_M)$, $\widetilde{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_M \cdots \boldsymbol{\Sigma}_{M-1} - \boldsymbol{\Sigma}_M)$ and $\widetilde{\boldsymbol{\phi}(L)} = (\sum_{i=1}^{p}(\boldsymbol{\phi}_{1,i} - \boldsymbol{\phi}_{M,i})L^i \cdots \sum_{i=1}^{p}(\boldsymbol{\phi}_{M-1,i} - \boldsymbol{\phi}_{M,i})L^i)$. From the transition equation in (4.4) we obtain $\boldsymbol{\delta}_{t+h} = \mathbf{F}^h\boldsymbol{\delta}_t + \sum_{j=0}^{h-1}\mathbf{F}^j\mathbf{w}_{t+h-j}$. Using this relation, $\mathbf{x}_{t+h}$ can be expressed as

$$\begin{aligned} \mathbf{x}_{t+h} =& \boldsymbol{\Lambda}\boldsymbol{\pi} + \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^h\boldsymbol{\delta}_t + \sum_{j=0}^{h-1}\widetilde{\boldsymbol{\Lambda}}\mathbf{F}^j\mathbf{w}_{t+h-j} + \widetilde{\boldsymbol{\Sigma}}[(\mathbf{F}^h\boldsymbol{\delta}_t) \otimes \mathbf{I}_K]\mathbf{u}_{t+h} \\ &+ \sum_{j=0}^{h-1}\widetilde{\boldsymbol{\Sigma}}[(\mathbf{F}^j\mathbf{w}_{t+h-j}) \otimes \mathbf{I}_K]\mathbf{u}_{t+h} + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_{t+h}. \end{aligned} \tag{4.5}$$

By (4.5), we obtain

$$cov(\mathbf{x}_{t+h}, \mathbf{y}_t) = cov(\boldsymbol{\Lambda}\boldsymbol{\pi} + \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^h\boldsymbol{\delta}_t, \mathbf{y}_t) = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^h cov(\boldsymbol{\delta}_t, \mathbf{y}_t) = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^h E(\boldsymbol{\delta}_t\mathbf{y}_t^{'}) \tag{4.6}$$

for every $h > 0$. For $h = 0$, we have

$$\begin{aligned} cov(\mathbf{x}_t, \mathbf{y}_t) &= cov(\boldsymbol{\Lambda}\boldsymbol{\pi} + \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\delta}_t + \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t, \mathbf{y}_t) \\ &= \widetilde{\boldsymbol{\Lambda}}E(\boldsymbol{\delta}_t\mathbf{y}_t^{'}) + \widetilde{\boldsymbol{\Sigma}}E[(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t\mathbf{y}_t^{'}] + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)E(\mathbf{u}_t\mathbf{y}_t^{'}). \end{aligned} \tag{4.7}$$

Now we are going to compute $E(\boldsymbol{\delta}_t\mathbf{y}_t^{'})$, $E(\mathbf{u}_t\mathbf{y}_t^{'})$ and $E[(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t\mathbf{y}_t^{'}]$. Postmultiplying the measurement equation in (4.4) by $\boldsymbol{\delta}_t^{'}$ and taking expectation give

$$\boldsymbol{\phi}(1)(\boldsymbol{\pi} \otimes \mathbf{I}_K)E(\mathbf{y}_t\boldsymbol{\delta}_t^{'}) = \widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{D}}$$

hence

$$E(\boldsymbol{\delta}_t\mathbf{y}_t^{'}) = \widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Lambda}}^{'}[R^{'}]^{-1}. \tag{4.8}$$

Postmultiplying the measurement equation in (4.2) and (4.4) by $\mathbf{u}_t^{'}$ and equating them, we get

$$\boldsymbol{\phi}(L)(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{y}_t\mathbf{u}_t^{'} = \boldsymbol{\Lambda}\boldsymbol{\pi}\mathbf{u}_t^{'} + \widetilde{\boldsymbol{\Lambda}}\delta_t\mathbf{u}_t^{'} + \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t\mathbf{u}_t^{'} + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t\mathbf{u}_t^{'}.$$

Taking expectation gives

$$\boldsymbol{\phi}(1)(\boldsymbol{\pi} \otimes \mathbf{I}_K)E(\mathbf{y}_t\mathbf{u}_t^{'}) = \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)$$

hence

$$E(\mathbf{u}_t\mathbf{y}_t^{'}) = (\boldsymbol{\pi}^{'} \otimes \mathbf{I}_K)\boldsymbol{\Sigma}^{'}[R^{'}]^{-1}. \tag{4.9}$$

Reasoning as above by using $\mathbf{u}_t^{'}(\boldsymbol{\delta}_t^{'} \otimes \mathbf{I}_K)$ instead of $\mathbf{u}_t^{'}$, we get

$$\boldsymbol{\phi}(1)(\boldsymbol{\pi} \otimes \mathbf{I}_K)E[\mathbf{y}_t\mathbf{u}_t^{'}(\boldsymbol{\delta}_t^{'} \otimes \mathbf{I}_K)] = \widetilde{\boldsymbol{\Sigma}}(\widetilde{\mathbf{D}} \otimes \mathbf{I}_K)$$

hence

$$E[(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t\mathbf{y}_t^{'}] = (\widetilde{\mathbf{D}} \otimes \mathbf{I}_K)\widetilde{\boldsymbol{\Sigma}}^{'}[R^{'}]^{-1}. \tag{4.10}$$

Substituting Formulae (4.8), (4.9) and (4.10) into (4.6) and (4.7), we get

$$cov(\mathbf{x}_{t+h}, \mathbf{y}_t) = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^h\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Lambda}}^{'}[R^{'}]^{-1} \tag{4.11}$$

for every $h > 0$, and

$$cov(\mathbf{x}_t, \mathbf{y}_t) = [\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Lambda}}^{'} + \boldsymbol{\Sigma}(\mathbf{D}\mathbf{P}_\infty \otimes \mathbf{I}_K)\boldsymbol{\Sigma}^{'} + \widetilde{\boldsymbol{\Sigma}}(\widetilde{\mathbf{D}} \otimes \mathbf{I}_K)\widetilde{\boldsymbol{\Sigma}}^{'}][R^{'}]^{-1}. \tag{4.12}$$

Collecting Formulae (4.3), (4.11) and (4.12) gives the following result:

Theorem 4.1. *Under the hypothesis that the regime variable is uncorrelated with the observable, the autocovariance function of the second-order stationary process $\mathbf{y} = (\mathbf{y}_t)$ in (4.1) is given by*

$$i)\quad \mathbf{B}(L)\Gamma_{\mathbf{y}}(0) = [\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Lambda}}^{'} + \boldsymbol{\Sigma}(\mathbf{D}\mathbf{P}_\infty \otimes \mathbf{I}_K)\boldsymbol{\Sigma}^{'} + \widetilde{\boldsymbol{\Sigma}}(\widetilde{\mathbf{D}} \otimes \mathbf{I}_K)\widetilde{\boldsymbol{\Sigma}}^{'}][R^{'}]^{-1}; \quad \textit{and}$$
$$ii)\quad \mathbf{B}(L)\Gamma_{\mathbf{y}}(h) = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^h\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Lambda}}^{'}[R^{'}]^{-1} \quad \textit{for every} \quad h > 0,$$

where $R = \boldsymbol{\phi}(1)(\boldsymbol{\pi} \otimes \mathbf{I}_K)$ has been assumed to be nonsingular.

Applying Theorem 2.2 for $q = 1$ and taking in mind that $\mathbf{F}$ is $(M-1) \times (M-1)$, we get

Theorem 4.2. *Suppose that the regime variable $\boldsymbol{\xi}_{t+h}$ is uncorrelated with $\mathbf{y}_t$ for every $h \geq 0$ and $\widetilde{\boldsymbol{\Lambda}} \neq \mathbf{0}$. Then the MS(M)-VAR(p) process $\mathbf{y} = (\mathbf{y}_t)$ in (4.1) admits a stable VARMA(p^*, q^*) representation with $p^* \leq M + p - 1$ and $q^* \leq M - 1$. If we require that the autoregressive lag polynomial of such a stable representation is scalar, then the bounds become $p^* \leq M + Kp - 1$ and $q^* \leq M + (K-1)p - 1$. If there is no cancellation, then the identification problem is completely solved and the above relations become equalities. In particular, the last formulae imply $M = K(q^* + 1) - (K-1)(p^* + 1)$ and $p = p^* - q^*$.*

Theorem 4.2 can be reformulated by using the hypothesis $\boldsymbol{\Lambda} \neq \mathbf{0}$ as done in Theorem 3.2. This arises from the autocovariances expressed by using the matrix $\mathbf{Q}$ instead of $\mathbf{F}$, similarly as in Theorem 4.1. Moreover, to justify the hypothesis of having $\widetilde{\boldsymbol{\Lambda}} \neq \mathbf{0}$ (respectively, $\boldsymbol{\Lambda} \neq \mathbf{0}$), we refer to Krolzig (1997), p.53, line 15, where it was assumed the identifiability of the regimes, $\boldsymbol{\nu}_i \neq \boldsymbol{\nu}_j$ for $i \neq j$, in order to render the results unique. To end the section we compute explicitly a VARMA representation for the process $\mathbf{y} = (\mathbf{y}_t)$ in (4.1). This gives a new proof of Theorem 4.2 and extends Proposition 3 from Krolzig (1997), Section 3.2.4. We start with the more simple case in which the autoregressive lag polynomial of the initial process is state independent.

Theorem 4.3. *Under quite general regularity conditions, the process $\mathbf{y} = (\mathbf{y}_t)$ in (4.1), with $\mathbf{\Phi}_{s_t}(L) = A(L)$ is state independent, has a VARMA(p^*, q^*) representation with $p^* \leq M+Kp-1$ and $q^* \leq M + (K-1)p - 1$*

$$\gamma(L)(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = \mathbf{C}(L)\boldsymbol{\epsilon}_t$$

where $\gamma(L) = |F(L)||A(L)|$ is the scalar AR operator of degree $M+Kp-1$, $\mathbf{C}(L) = [A(L)^*\widetilde{\Lambda}F(L)^* \quad A(L)^*\widetilde{\Sigma}(F(L)^* \otimes \mathbf{I}_K) \qquad |F(L)|A(L)^*\Sigma(\pi \otimes \mathbf{I}_K)]$ is a matrix lag polynomial of degree $M + (K-1)p - 1$, and $\boldsymbol{\epsilon}_t = (\boldsymbol{w}_t' \quad \boldsymbol{u}_t'(\boldsymbol{w}_t' \otimes \mathbf{I}_K) \quad \boldsymbol{u}_t')'$ is a zero mean vector white noise process with $\text{var}(\boldsymbol{\epsilon}_t) = \text{diag}(\widetilde{\mathbf{D}} - \mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}', (\widetilde{\mathbf{D}} - \mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}') \otimes \mathbf{I}_K, \mathbf{I}_K)$.

In the general case in which the autoregression part of the initial process is state dependent but the regime variable is uncorrelated with the observable, we can proceed as follows. By Theorem 4.1 the autocovariances of the process satisfy a finite difference equation of order $p^* = M + Kp - 1$ and rank $q^* + 1 = M + (K-1)p$. Then the process can be represented by a stable VARMA(p^*, q^*), whose autoregression lag polynomial is assumed to be scalar. Given the process $(\mathbf{y}_t)$, we can estimate the coefficients of the stable VARMA(p^*, q^*) via OLS. If there is no cancellation between the AR and MA part of the estimated VARMA(p^*, q^*), then we get the representation of Theorem 4.2 with equalities.

To complete the section we also discuss the forecasting for our Markov switching autoregressive model. So let us consider the MS(M)-VAR(p) model in (4.4). Then we can write

$$\mathbf{y}_t + \sum_{i=1}^{p} \widetilde{\mathbf{\Phi}}_i(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{y}_{t-i} + \sum_{i=1}^{p} \mathbf{\Phi}_i(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{y}_{t-i} = \mathbf{\Lambda}\boldsymbol{\pi} + \widetilde{\mathbf{\Lambda}}\boldsymbol{\delta}_t + \widetilde{\mathbf{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t + \mathbf{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t$$

where $\mathbf{\Phi}_i = (\mathbf{\Phi}_{1,i} \cdots \mathbf{\Phi}_{M,i})$ and $\widetilde{\mathbf{\Phi}}_i = (\mathbf{\Phi}_{1,i} - \mathbf{\Phi}_{M,i} \cdots \mathbf{\Phi}_{M-1,i} - \mathbf{\Phi}_{M,i})$, for every $i = 1, \ldots, p$

(see Section 4). The one-step predictor $\widehat{\mathbf{y}}_{t+1|t}$ can be calculated as above, so we get

$$\widehat{\mathbf{y}}_{t+1|t} = \mathbf{\Lambda}\boldsymbol{\pi} + \widetilde{\mathbf{\Lambda}}\mathbf{F}\widehat{\boldsymbol{\delta}}_{t|t} - \sum_{i=1}^{p} \widetilde{\mathbf{\Phi}}_i(\mathbf{F}\otimes\mathbf{I}_K)(\widehat{\boldsymbol{\delta}}_{t|t}\otimes\mathbf{I}_K)\mathbf{y}_{t+1-i} - \sum_{i=1}^{p}\mathbf{\Phi}_i(\boldsymbol{\pi}\otimes\mathbf{I}_K)\mathbf{y}_{t+1-i}.$$

For h–step predictions, $h > 1$, the task is much more complicated, and the last formula generalizes as follows

$$\widehat{\mathbf{y}}_{t+h|t} = \mathbf{\Lambda}\boldsymbol{\pi} + \widetilde{\mathbf{\Lambda}}\mathbf{F}^h\widehat{\boldsymbol{\delta}}_{t|t} - \sum_{i=1}^{p} \widetilde{\mathbf{\Phi}}_i(\mathbf{F}\otimes\mathbf{I}_K)^h(\widehat{\boldsymbol{\delta}}_{t|t}\otimes\mathbf{I}_K)\widehat{\mathbf{y}}_{t+h-i|t} - \sum_{i=1}^{p}\mathbf{\Phi}_i(\boldsymbol{\pi}\otimes\mathbf{I}_K)\widehat{\mathbf{y}}_{t+h-i|t}$$

which in practice gives a recursive formula. Also in this case, taking the limits for $h \to \infty$ on both sides yields

$$\lim_{h\to\infty}\widehat{\mathbf{y}}_{t+h|t} = \mathbf{\Lambda}\boldsymbol{\pi} - \sum_{i=1}^{p}\mathbf{\Phi}_i(\boldsymbol{\pi}\otimes\mathbf{I}_K)\lim_{h\to\infty}\widehat{\mathbf{y}}_{t+h-i|t}$$

hence

$$\lim_{h\to\infty}\widehat{\mathbf{y}}_{t+h|t} = R^{-1}\mathbf{\Lambda}\boldsymbol{\pi} = \boldsymbol{\mu}_{\mathbf{y}}.$$

where $R = \mathbf{\Phi}(1)(\boldsymbol{\pi}\otimes\mathbf{I}_K)$. However, in applied work it is customary to follow a suggestion of Doan, Litterman and Sims (1984) for which the sequence of predicted values $\{\widehat{\mathbf{y}}_{t+1|t}, \dots, \widehat{\mathbf{y}}_{t+h|t}, \dots\}$ is substituted by the sequence $\{\widehat{\mathbf{y}}_{t+1|t}, \dots, \widehat{\mathbf{y}}_{t+h|t+h-1}, \dots\}$. See, for example, Krolzig (1997), Section 4.4, for more details on this construction. Of course, the calculation of the filtered regime probabilities $\widehat{\boldsymbol{\xi}}_{t|t}$ (and hence $\widehat{\boldsymbol{\delta}}_{t|t}$) can be performed by the recursive formulae listed at the end of Section 3.

2.1.5 Data Simulation

In this section, we perform some MonteCarlo experiment for the estimation of the number of states given by the estimated lower bound obtained in the previous sections and penalized likelihood criteria such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan-Quinn Criterion (HQC). For the computation of the orders of the stable VARMA we use the 3-pattern method (TPM) proposed by Choi (1992). We perform three experiments.

Parameters	1.A		1.B		1.C	
P	1	0	0.8	0.2	0.8	0.2
	1	0	0.2	0.8	0.2	0.8
λ	4	4	4	12	4	6

Table 2.1: First experiment: Poisson Markov regime switching models with two states. In table we report the transition matrices Ps and Poisson Means λ for cases A–C.

The first is taken from Zhang and Stine (2001) in which we generate three different Poisson Markov regime switching models with two states. The second is a Markov switching process with two states and autoregressive dynamic with one lag (in short, MS(2)-AR(1) process) and the third is a two-states Markov switching with two lags in the autoregressive part (in short, MS(2)-AR(2)). The data-generating processes have gaussian i.i.d. errors and the parameters are reported below in Table 1, 2 and 3.

With regards to the first experiment, we consider three different Markov switching Poisson models (Table 1). The first model, denoted as case A, is a one-state model; that is, y_t is an i.i.d. sequence of Poisson random variables with λ= 4. Cases B and C are two-state models and share the same transition matrix but different means. With λ_1= 4 and λ_2 = 12, the two states of case B are more distinct than those of case C with λ_1 = 4 and λ_2 = 6.

Concerning with the second experiment, the parameters are set in accordance to Table 2. It is a Markov switching process with two states and autoregressive dynamic with one lag and it can be written as $y_t = \mu_{s_t} + \phi_{s_t} y_{t-1} + \sigma_{s_t} u_t$ with $s_t \in \{1, 2\}$. The three models in A, B and C share the same transition probability matrix and the values for the intercept and the variance but there are diffences in the autoregressive parameters (this follow some experiments as in Psaradakis and Spagnolo, 2003). The last case D, instead, has a different transition matrix which gives a more persistent chain.

Finally, in the third experiment, we consider a Markov switching process with two states and autoregressive dynamic with two lags, written as $y_t = \mu_{s_t} + \phi_{1s_t} y_{t-1} + \phi_{2s_t} y_{t-2} + \sigma_{s_t} u_t$ with $s_t \in \{1, 2\}$. Here we want to compare the performance of the bounds proposed in the present paper (we will denote it by CAV) with those proposed by Zhang and Stine (2001) (in short, ZS) and Francq and Zakoïan (2001) (in short, FZ) for those autoregressive markov switching models. Case B considers a more persistent process compared to the baseline case A and case

Parameters	2.A		2.B		2.C		2.D	
P	0.6	0.4	0.6	0.4	0.6	0.4	0.8	0.2
	0.4	0.6	0.4	0.6	0.4	0.6	0.2	0.8
μ	0	3	0	3	0	3	0	3
σ	1	$\sqrt{1.5}$	1	$\sqrt{1.5}$	1	$\sqrt{1.5}$	1	$\sqrt{1.5}$
ϕ	0.3	0.3	0.3	0.6	0.3	0.9	0.3	0.6

Table 2.2: Second experiment: Markov switching process with two states and autoregressive dynamic with one lag. In table we report the transition matrices Ps and the parameters of the process (means μs, standard deviation σs and autoregressive coefficients ϕs) for cases A–D.

Parameters	3.A		3.B		3.C	
P	0.6	0.4	0.9	0.1	0.6	0.4
	0.4	0.6	0.1	0.9	0.4	0.6
μ	0	3	0	3	0	3
σ	1	$\sqrt{1.5}$	1	$\sqrt{1.5}$	1	$\sqrt{1.5}$
ϕ_1	0.3	0.6	0.3	0.6	0.3	0.3
ϕ_2	0.4	0.8	0.4	0.8	0.1	0.1

Table 2.3: Third experiment: Markov switching process with two states and autoregressive dynamic with two lags. In table we report the transition matrices Ps and the parameters of the process (means μs, standard deviation σs and autoregressive coefficients ϕ_1s and ϕ_2s) for cases A–C.

C instead considers same autoregressive coefficients in different states.

The experiments simulate artificial time series of length $T + 50$ with $T \in \{100, 500, 1000\}$; the first 50 initial data points are discarded to minimize the effect of initial conditions. 100 MonteCarlo replications are carried out for each trial. When complexity-penalized likelihood criteria are computed, we use the recursive procedure discussed by Hamilton and the penalization constants are the usual proposed in that literature (1 for AIC, $\frac{1}{2}lnN$ for BIC and $lnlnN$ for HQC).

The simulation results from the first experiment are reported in Table 4. With respect to the one-state case (A) only the TPM applied on our bounds seems to correctly predict the exact number of states most of the times, while the likelihood criteria seem to overestimate that. The same happens for case B and in case C for larger samples. Overall, the TPM does better than any other likelihood methods.

With regards to the second experiment, the results are shown in Table 5. Here close

		1.A				1.B				1.C			
N	$\hat{M}$	AIC	BIC	HQC	TPM	AIC	BIC	HQC	TPM	AIC	BIC	HQC	TPM
100	1	0	0	0	45	0	0	0	5	0	0	0	39
	2	53	74	66	15	22	41	26	64	59	77	67	26
	3	47	26	34	40	78	59	74	31	41	23	33	35
500	1	0	0	0	44	0	0	0	0	0	0	0	26
	2	28	58	44	25	4	6	5	62	11	51	27	36
	3	72	42	56	31	96	94	95	38	89	49	73	38
1000	1	0	0	0	40	0	0	0	0	0	0	0	8
	2	19	58	31	23	0	2	2	60	8	27	11	62
	3	81	42	69	37	100	98	98	40	92	73	89	28

Table 2.4: Simulation results of the first experiment: Poisson Markov regime switching models with two states for cases A–C. We report the number of regimes chosen by the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and Hannan-Quinn Criterion (HQC). TPM denotes the number of regimes chosen in accordance to the bounds presented in the present paper.

conclusions can be drawn. In fact, the likelihood methods overestimate the number of states, with the exception of the Bayesian Criterion (BIC) in small sample, while the TPM detects it most of the times. These conclusions are robust to the change in the transition probability matrix, when choosing a more persistent chain (case D).

The results of the third experiment are in Table 6. As expected from theoretical aspects, using the bounds of Zhang and Stine (2001) (ZS in short) or Francq and Zakoïan (2001) (FZ in short) we tend to underestimate the number of states since these bounds are larger and then less informative. Whereas when using the bounds obtained in this work (denote by CAV), we are able to detect the exact number of regimes most of the time; and this choice is robust to the change of the transition probability matrix (case B) or of the values of the autoregressive coefficients (case C).

2.1.6 Application on foreign exchange rates

As an application on real data, we want to consider and complete the example of Zhang and Stine (2001) on foreign exchange rates. The data are the same used in Engle and Hamilton (1990), who consider quarterly data for French franc, British pound and German mark for the period from 1973:Q3 to 1988:Q4. Engle and Hamilton (1990) proposed to model the logarithm of exchange rates as a two states Markov-switching autoregressive of order one. In line with

		2.A				2.B			
N	$\hat{M}$	AIC	BIC	HQC	TPM	AIC	BIC	HQC	TPM
100	1	0	0	0	0	0	0	0	0
	2	38	87	65	75	24	63	40	76
	3	62	13	35	25	76	37	60	24
500	1	0	0	0	0	0	0	0	0
	2	0	21	1	74	1	3	1	70
	3	100	79	99	26	99	97	99	30
1000	1	0	0	0	0	0	0	0	0
	2	0	0	0	68	0	0	0	72
	3	100	100	100	32	100	100	100	28
		2.C				2.D			
N	$\hat{M}$	AIC	BIC	HQC	TPM	AIC	BIC	HQC	TPM
100	1	0	0	0	0	0	0	0	0
	2	13	48	21	74	10	58	24	75
	3	87	52	79	26	90	42	76	25
500	1	0	0	0	0	0	0	0	0
	2	9	16	13	75	0	2	1	82
	3	91	84	87	25	100	98	99	18
1000	1	0	0	0	0	0	0	0	0
	2	10	17	14	76	0	0	0	74
	3	90	83	86	24	100	100	100	26

Table 2.5: Simulation results of the second experiment: Markov switching process with two states and autoregressive dynamic with one lag (in short, MS(2)-AR(1) process) for cases A–D. We report the number of regimes chosen by the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and Hannan-Quinn Criterion (HQC). TPM denotes the number of regimes chosen in accordance to the bounds presented in the present paper.

		3.A			3.B			3.C		
N	$\hat{M}$	CAV	ZS	FZ	CAV	ZS	FZ	CAV	ZS	FZ
100	1	0	100	100	0	100	100	1	100	100
	2	73	0	0	96	0	0	75	0	0
	3	27	0	0	4	0	0	24	0	0
500	1	0	100	100	0	100	100	0	100	100
	2	75	0	0	20	0	0	70	0	0
	3	25	0	0	80	0	0	30	0	0
1000	1	0	100	100	0	100	100	0	100	100
	2	75	0	0	50	0	0	60	0	0
	3	25	0	0	50	0	0	40	0	0

Table 2.6: Simulation results of the third experiment: Markov switching process with two states and autoregressive dynamic with two lags (in short, MS(2)-AR(2) process) for cases A–C. We report the number of regimes chosen by using the three-pattern method and applying different bounds either those obtained in Zhang and Stine (2001) (ZS), Francq and Zakoïan (2001) (FZ) or in the present paper (CAV).

Criteria	Franc	Pound	Mark
AIC	2	2	1
BIC	1	1	1
HQC	1	1	1
CAV	2	1	2
ZS	0	0	0
FZ	1	0	1

Table 2.7: Estimates of the number of regimes for quarterly data on French franc, British pound and German mark for the period from 1973:Q3 to 1988:Q4 based on penalized likelihood criteria (Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan-Quinn Criterion (HQC)) and based on three-pattern method computing the lower bounds as in Zhang and Stine (2001) (ZS), Francq and Zakoïan (2001) (FZ) or in the present paper (CAV).

Zhang and Stine (2001), when we fit Gaussian regime switching models, penalized likelihood criteria as Bayesian Information Criterion (BIC) and Hannan-Quinn Criterion (HQC) choose $M = 1$ for all three currencies while Akaike Information Criterion (AIC) chooses 2 regimes for franc and pound and only one regime for the mark. What is interesting is that, when we compute the orders of the stable VARMA and evaluate the lower bounds as proposed in the papers of Zhang and Stine (2001) (ZS), Francq and Zakoïan (2001) (FZ) and in the present one (CAV), we find that our bounds are more precise and then more informative. In particular, our bounds propose the existence of two regimes for franc and mark and one regime for the pound (as reported in Table 7), while using the bounds of ZS and FZ we are not able to infer any information on regime switching from the data. Finally, note that the last methodology compared to penalized methods is less demanding and computationally faster since it does not request likelihood calculations.

2.1.7 Conclusion

In this paper, for M-state Markov switching multivariate moving average models and autoregressive models in which the regime variable is uncorrelated with the observable, we give finite order VARMA(p^*, q^*) representations. The parameters of the VARMA can be determined by evaluating the autocovariance function of the Markov-switching models. It turns out that upper bounds for p^* and q^* are elementary functions of the dimension K of the process, the number M of regimes, and the orders p and q. In particular, the order of the stable VARMA admits a simple form: $p^* \leq M-1$, $q^* \leq M+q-1$ for M-state switching VMA(q) models and $p^* \leq M+Kp-1$, $q^* \leq M+p(K-1)+q-1$ for M-state switching VAR(p) models. This result yields an easily computed method for setting a lower bound on the number of regimes from an estimated autocovariance function. Our results include, as particular cases, those obtained by Krolzig (1997), and improve the bounds found in the literature in the works of Zhang and Stine (2001) and Francq and Zakoïan (2001) for our classes of dynamic models. Our simulation results indicate the procedure is more precise than penalized likelihood criteria such as AIC, BIC and HQC which require more elaborate procedures and assumption of a specific probability model and the associated likelihood calculations. Moreover, having bounds for the number of states which are small than those of Zhang and Stine (2001) or Francq and Zakoïan (2001) give estimates for the number of states which are more precise and then more informative. This is shown both with simulated experiments and with real data application on exchange rates.

References

[1] Choi, B. S. (1992) ARMA Model Identification, New York, Springer-Verlag.

[2] Doan, T., Litterman, R. and Sims, C. (1984) *Forecasting and conditional projection using realistic prior distributions*, Econometric Reviews **3**, 1–100.

[3] Engle, C. and Hamilton, J.D. (1990) *Long swings in the dollar: are they in the data and do markets know it?*, American Economic Review **80**, 689–713.

[4] Francq, C. and Zakoïan, J.M. (2001) *Stationarity of Multivariate Markov-Switching ARMA Models*, Journal of Econometrics **102**, 339–364.

[5] Francq, C. and Zakoïan, J.M. (2002) *Autocovariance Structure of Powers of Switching-Regime ARMA Processes*, ESAIM: Probability and Statistics **6**, 259–270.

[6] Hansen B.E., (1992) *The likelihood Ratio Test under Nonstandard Conditions: testing the Markov Switching Model of GNP*, Journal of Applied Econometrics **7**, S61–S82.

[7] Karlsen, H.A (1990a) *Doubly stochastic vector AR(1) processes*, Technical Report, Department of Mathematics, University of Bergen, Norway.

[8] Karlsen, H.A (1990b) *A class of non-linear time series models*, Ph.D. Thesis, University of Bergen, Norway.

[9] Krolzig, H.M. (1997) *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference and Application to Business Cycle Analysis*, Springer Verlag, Berlin-Heidelberg-New York.

[10] Olteanu, M. and Rynkiewicz, J. (2007) *Estimating the number of regimes in a switching autoregressive model*, HAL (Hyper Articles en Ligne), CCSD (Centre pour la Communication Scientifique Directe).

[11] Psaradakis, Z. and Spagnolo, N. (2003) *On the determination of the number of regimes in Markov-switching autoregressive models*, Journal of Time Series Analysis **24** (2), 237–252.

[12] Psaradakis, Z. and Spagnolo, N. (2006) *Joint determination of the state dimension and autoregressive order for models with in Markov regime switching*, Journal of Time Series Analysis **27** (5), 753–766.

[13] Ríos, R. and Rodríguez, L.A. (2008) *Penalized estimate of the number of states in Gaussian linear AR with Markov regime*, Electronic Journal of Statistics, **2**, 1111–1128.

[14] Zhang, J. and Stine, R.A. (2001) *Autocovariance structure of Markov regime switching models and model selection.* Journal of Time Series Analysis **22**(1), 107–124.

2.1.8 Appendix

In this section, we give proofs of Theorems in Sections 3 and 4.

Proof of Theorem 3.1. The following are well-known facts (see, for example, Zhang and Stine (2001), Section 3.1): $\mathbf{D}\mathbf{P}_\infty = \boldsymbol{\pi}\boldsymbol{\pi}'$, $\mathbf{P}_\infty^n = \mathbf{P}^n\mathbf{P}_\infty = \mathbf{P}_\infty\mathbf{P}^n = \mathbf{P}_\infty$ and $\mathbf{Q}^n = \mathbf{P}^n - \mathbf{P}_\infty$

for every $n \geq 1$. First we treat the case $h = 0$. Then we have

$$\begin{aligned}
\Gamma_{\mathbf{y}}(0) &= E(\mathbf{y}_t\mathbf{y}_t^{'}) - E(\mathbf{y}_t)E(\mathbf{y}_t^{'}) \\
&= E(\mathbf{y}_t\mathbf{y}_t^{'}) - \boldsymbol{\Lambda}\boldsymbol{\pi}\boldsymbol{\pi}^{'}\boldsymbol{\Lambda}^{'} \\
&= E(\mathbf{y}_t\mathbf{y}_t^{'}) - \boldsymbol{\Lambda}\mathbf{D}\mathbf{P}_\infty\boldsymbol{\Lambda}^{'}
\end{aligned}$$

and

$$\begin{aligned}
E(\mathbf{y}_t\mathbf{y}_t^{'}) &= E[(\boldsymbol{\Lambda}\boldsymbol{\xi}_t + \sum_{j=0}^{q}\boldsymbol{\Theta}_j(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{u}_{t-j})(\sum_{j=0}^{q}\mathbf{u}_{t-j}^{'}(\boldsymbol{\xi}_t^{'} \otimes \mathbf{I}_K)\boldsymbol{\Theta}_j^{'} + \boldsymbol{\xi}_t^{'}\boldsymbol{\Lambda}^{'})] \\
&= \boldsymbol{\Lambda}E(\boldsymbol{\xi}_t\boldsymbol{\xi}_t^{'})\boldsymbol{\Lambda}^{'} + \sum_{j=0}^{q}\boldsymbol{\Theta}_j[E(\boldsymbol{\xi}_t\boldsymbol{\xi}_t^{'}) \otimes \mathbf{I}_K]\boldsymbol{\Theta}_j^{'} \\
&= \boldsymbol{\Lambda}\mathbf{D}\boldsymbol{\Lambda}^{'} + \sum_{j=0}^{q}\boldsymbol{\Theta}_j(\mathbf{D} \otimes \mathbf{I}_K)\boldsymbol{\Theta}_j^{'}
\end{aligned}$$

hence

$$\Gamma_{\mathbf{y}}(0) = \boldsymbol{\Lambda}\mathbf{D}(\mathbf{I}_M - \mathbf{P}_\infty)\boldsymbol{\Lambda}^{'} + \sum_{j=0}^{q}\boldsymbol{\Theta}_j(\mathbf{D} \otimes \mathbf{I}_K)\boldsymbol{\Theta}_j^{'}$$

which proves i) for $h = 0$. For $h = 1, \ldots, q$, we have

$$\Gamma_{\mathbf{y}}(-h) = cov(\mathbf{y}_t, \mathbf{y}_{t+h}) = E(\mathbf{y}_t\mathbf{y}_{t+h}^{'}) - E(\mathbf{y}_t)E(\mathbf{y}_{t+h}^{'}) = E(\mathbf{y}_t\mathbf{y}_{t+h}^{'}) - \boldsymbol{\Lambda}\mathbf{D}\mathbf{P}_\infty\boldsymbol{\Lambda}^{'}$$

and

$$\begin{aligned}
E(\mathbf{y}_t\mathbf{y}_{t+h}^{'}) &= E[(\boldsymbol{\Lambda}\boldsymbol{\xi}_t + \sum_{j=0}^{q}\boldsymbol{\Theta}_j(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{u}_{t-j})(\sum_{i=0}^{q}\mathbf{u}_{t+h-i}^{'}(\boldsymbol{\xi}_{t+h}^{'} \otimes \mathbf{I}_K)\boldsymbol{\Theta}_i^{'} + \boldsymbol{\xi}_{t+h}^{'}\boldsymbol{\Lambda}^{'})] \\
&= \boldsymbol{\Lambda}E(\boldsymbol{\xi}_t\boldsymbol{\xi}_{t+h}^{'})\boldsymbol{\Lambda}^{'} + \sum_{j=0}^{q}\sum_{i=0}^{q}\boldsymbol{\Theta}_j[E(\boldsymbol{\xi}_t\boldsymbol{\xi}_{t+h}^{'}) \otimes \delta_{t+h-i}^{t-j}\mathbf{I}_K]\boldsymbol{\Theta}_i^{'} \\
&= \boldsymbol{\Lambda}\mathbf{D}\mathbf{P}^h\boldsymbol{\Lambda}^{'} + \sum_{i=0}^{q-h}\boldsymbol{\Theta}_i[(\mathbf{D}\mathbf{P}^h) \otimes \mathbf{I}_K]\boldsymbol{\Theta}_{i+h}^{'}
\end{aligned}$$

hence

$$\begin{aligned}
\Gamma_{\mathbf{y}}(-h) &= \boldsymbol{\Lambda}\mathbf{D}(\mathbf{P}^h - \mathbf{P}_\infty)\boldsymbol{\Lambda}^{'} + \sum_{i=0}^{q-h}\boldsymbol{\Theta}_i[(\mathbf{D}\mathbf{P}^h) \otimes \mathbf{I}_K]\boldsymbol{\Theta}_{i+h}^{'} \\
&= \boldsymbol{\Lambda}\mathbf{D}\mathbf{Q}^h\boldsymbol{\Lambda}^{'} + \sum_{i=0}^{q-h}\boldsymbol{\Theta}_i[(\mathbf{D}\mathbf{P}^h) \otimes \mathbf{I}_K]\boldsymbol{\Theta}_{i+h}^{'}.
\end{aligned}$$

Now taking transposition and setting $j = i + h$, we get

$$\Gamma_{\mathbf{y}}(h) = \mathbf{\Lambda}(\mathbf{Q}^{'})^h \mathbf{D}\mathbf{\Lambda}^{'} + \sum_{j=h}^{q} \mathbf{\Theta}_j[((\mathbf{P}^{'})^h \mathbf{D}) \otimes \mathbf{I}_K]\mathbf{\Theta}_{j-h}^{'}$$

which proves i) for $h = 1, \ldots, q$. For every $h \geq q + 1$, we have

$$E(\mathbf{y}_t \mathbf{y}_{t+h}^{'}) = \mathbf{\Lambda} E(\boldsymbol{\xi}_t \boldsymbol{\xi}_{t+h}^{'})\mathbf{\Lambda}^{'} = \mathbf{\Lambda}\mathbf{D}\mathbf{P}^h\mathbf{\Lambda}^{'}$$

and

$$\Gamma_{\mathbf{y}}(-h) = \mathbf{\Lambda}\mathbf{D}(\mathbf{P}^h - \mathbf{P}_\infty)\mathbf{\Lambda}^{'} = \mathbf{\Lambda}\mathbf{D}\mathbf{Q}^h\mathbf{\Lambda}^{'}$$

hence

$$\Gamma_{\mathbf{y}}(h) = \mathbf{\Lambda}(\mathbf{Q}^{'})^h \mathbf{D}\mathbf{\Lambda}^{'}$$

which proves ii). □

Proof of Theorem 3.2. For $h \geq q+1$, the autocovariance function $\Gamma_{\mathbf{y}}(h) = \mathbf{\Lambda}(\mathbf{Q}^{'})^h\mathbf{D}\mathbf{\Lambda}^{'}$ is in the form specified in Theorem 2.2 with $p = 0$ and $q + 1$ instead of q. As remarked in Zhang and Stine (2001) (see the proof of Theorem 3), the minimal polynomial of Q has a zero constant term as Q is singular. So the proof is now slightly different from that of Theorem 2.2. Let λ_i, $i = 1, \ldots, M$, be the eigenvalues of $\mathbf{P}^{'}$, where we set $\lambda_1 = 1$ as $\mathbf{P}^{'}\boldsymbol{\pi} = \boldsymbol{\pi}$. Since the Markov chain is ergodic, all other eigenvalues of $\mathbf{P}^{'}$ are inside the unit circle. It follows that the eigenvalues of $\mathbf{Q}^{'}$ are $\mu_1 = 0$ and $\mu_i = \lambda_i - \lim_n \lambda_i^n = \lambda_i$ for $i = 2, \ldots, M$. Since $\mathbf{Q}^{'}$ is an $M \times M$ singular matrix, its minimal polynomial can be written as $\varphi(x) = x^M - f_1 x^{M-1} - \cdots - f_{M-1}x$, where the coefficient f_{M-1} may be zero. An argument similar to that used in the proof of Theorem 2.2 gives

$$\Gamma_{\mathbf{y}}(M + q) - f_1\Gamma_{\mathbf{y}}(M + q - 1) - \cdots - f_{M-1}\Gamma_{\mathbf{y}}(q + 1) = 0.$$

The result now follows from Theorem 2.1 with $h = M + q$, $p^* \leq M - 1$ and $q^* \leq h - 1$. □

Proof of Theorem 3.4 From (3.5) we get $\boldsymbol{\delta}_t = (\mathbf{I}_{M-1} - \mathbf{F}L)^{-1}\mathbf{w}_t = \mathbf{F}(L)^{-1}\mathbf{w}_t$ (here we have used the fact that all the eigenvalues of $\mathbf{F}$ are less than 1 in modulus). Inserting the above relation in (3.7) gives the result of the statement. □

Proof of Theorem 3.5 Premultiplying by $|\mathbf{F}(L)|$ the MA(∞) representation of Theorem 3.4 yields

$$|\mathbf{F}(L)|(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^*(L)\mathbf{w}_t + \sum_{j=0}^{q} \widetilde{\boldsymbol{\Theta}}_j[(\mathbf{F}^*(L)\mathbf{w}_t) \otimes \mathbf{I}_K]L^j\mathbf{u}_t + |\mathbf{F}(L)|\sum_{j=0}^{q} \boldsymbol{\Theta}_j(\boldsymbol{\pi} \otimes \mathbf{I}_K)L^j\mathbf{u}_t$$
$$= \widetilde{\boldsymbol{\Lambda}}\mathbf{F}^*(L)\mathbf{w}_t + \sum_{j=0}^{q} \widetilde{\boldsymbol{\Theta}}_j[\mathbf{F}^*(L) \otimes \mathbf{I}_K]L^j(\mathbf{w}_{t+j} \otimes \mathbf{I}_K)\mathbf{u}_t + |\mathbf{F}(L)|\sum_{j=0}^{q} \boldsymbol{\Theta}_j(\boldsymbol{\pi} \otimes \mathbf{I}_K)L^j\mathbf{u}_t$$

which is a VARMA(p^*, q^*) representation as claimed in the statement. □

Proof of Theorem 4.3 From (4.4) we get $\boldsymbol{\delta}_t = F(L)^{-1}\mathbf{w}_t$ as usual. Equating (4.1) and (4.4) and substituting the last formula, we get

$$A(L)\mathbf{y}_t = \boldsymbol{\Lambda\pi} + \widetilde{\boldsymbol{\Lambda}}\boldsymbol{\delta}_t + \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t$$
$$= A(1)\boldsymbol{\mu}_{\mathbf{y}} + \widetilde{\boldsymbol{\Lambda}}F(L)^{-1}\mathbf{w}_t + \widetilde{\boldsymbol{\Sigma}}(F(L)^{-1} \otimes \mathbf{I}_K)(\mathbf{w}_t \otimes \mathbf{I}_K)\mathbf{u}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t$$

hence

$$A(L)(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = \widetilde{\boldsymbol{\Lambda}}F(L)^{-1}\mathbf{w}_t + \widetilde{\boldsymbol{\Sigma}}(F(L)^{-1} \otimes \mathbf{I}_K)(\mathbf{w}_t \otimes \mathbf{I}_K)\mathbf{u}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t. \tag{4.13}$$

Premultiplying (4.13) by $|F(L)|$ yields

$$|F(L)|A(L)(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = \widetilde{\boldsymbol{\Lambda}}F(L)^*\mathbf{w}_t + \widetilde{\boldsymbol{\Sigma}}(F(L)^* \otimes \mathbf{I}_K)(\mathbf{w}_t \otimes \mathbf{I}_K)\mathbf{u}_t + |F(L)|\boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t. \tag{4.14}$$

Now the regularity conditions of the statement mean that $A(L)$ is invertible, that is, $A(L)^*A(L) = |A(L)|\mathbf{I}_K$. Premultiplying (4.14) by $A(L)^*$, we get the VARMA(p^*, q^*) representation, with $p^* \leq M + Kp - 1$ and $q^* \leq M + (K-1)p - 1$ (use the fact that the degree of $|A(L)|$ is Kp):

$$|F(L)||A(L)|(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = A(L)^*\widetilde{\boldsymbol{\Lambda}}F(L)^*\mathbf{w}_t + A(L)^*\widetilde{\boldsymbol{\Sigma}}(F(L)^* \otimes \mathbf{I}_K)(\mathbf{w}_t \otimes \mathbf{I}_K)\mathbf{u}_t$$
$$+ |F(L)|A(L)^*\boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t$$

which is a model as required in the statement. □

2.2 Business Cycle and Markov Switching Models with Distributed Lags: a Comparison between US and Euro Area

Abstract. Abstract. *Business cycle models are often investigated by using reduced form time series models, other than (or in alternative to) structural highly grounded in economic theory models. Reduced form VARMA with fixed parameters play a key role in business cycle analysis, but it is often found that by their very nature they do not typically capture the changing phases and regimes which characterize the economy. ln this paper we show that well-known state space systems used to analyse business cycle in several empirical works can be comprised into a broad class of non linear models, the MSI-VARMA. These processes are M-state Markov switching VARMA models for which the intercept term depends not only on the actual regime but also on the last r regimes. We give stable finite order VARMA representations for these processes, where upper bounds for the stable VARMA orders are elementary functions of the parameters of the initial switching model. If there is no cancellation, the bounds become equalities, and this solves the identification problem. This result allows us to study US and European business cycles and to determine the number of regimes most appropriate for the description of the economic systems. Two regimes are confirmed for the US economy; the European business cycle exhibits, instead, strong non-linearities and more regimes are necessary. This is taken into account when performing estimation and regime identification. [JEL Classification: C01, C50, C32, E32]*

Keywords: Time series, Varma models, Markov chains, changes in regime, regime number, business cycle models.

2.2.1 Introduction

In last years, we have witnessed a renewed interest in estimating and forecasting economic growth rates and turning points in economic activity. To address those issues, business cycle models have been proposed in the literature which are highly grounded in economic theory. Since Hamilton (1989) those models, which are usually in state space form, have been accessorized with devices being able to capture asymmetries and turning points in business cycle dynamics: Markov switching features. However, estimation tasks of Markov switching state space are not easily performed due to complexity of estimation algorithms. Therefore, researchers typically deal with easier equipments such as reduced form Vector Autoregressive Moving Average (VARMA) models, but it is often found that VARMA models with fixed parametes do not typically capture the changing phases characterizing the economy. We note that well-known state space models of business cycle can be rewritten into a special case of Markov switching VARMA models (in short, MS-VARMA), which we call Markov switching VARMA with distributed lags (in short, MSI-VARMA). These models for time series allow the parameters to change as a result of the outcome of an unobserved Markov chain, which is a M-state discrete variable. Having such a convenient reduced form permits simpler inference and a more tractable framework. In particular, differently from the baseline of a MS-VARMA (see Cavicchioli (2013)), we study MSI-VARMA models where the intercept term depends not only on the actual regime, but also on the last r regimes. A key problem when using non-linear models is the determination of the number of states which best describes the observed data. In fact, in empirical applications where such non-linear specifications are employed, the number of regimes is sometimes dictated by the particular application or is determined in an informal manner by visual inspection of plots. In this work we propose a method for the determination of the number of regimes which relies on the computation of the autocovariance function and on finite order (or stable) VARMA(p^*, q^*) representation of the initial switching model (see Krolzig (1997) and Cavicchioli (2013)). More precisely, the parameters of the VARMA representation can be determined by evaluating the autocovariance function of the Markov switching model. It turns out that the orders p^* and q^* of the stable VARMA representations are elementary functions of the dimension of the process, the number M of regimes, the number of lags r on the intercept and the AR and MA orders of the initial swiching model. Moreover, if there is no cancellation among the roots of the autoregressive and moving average polynomials, the bounds become equalities, and this solves the identification problem. With these results in hand, we are thus able to propose an easier toolkit for the study of economic phases in applied works. Particularly, business cycle analysis takes its first steps from the main empirical facts establishing that during a postwar period, contraction has typically been followed by a high-growth recovery that quickly boosts output to its prerecession level. This two phase pattern was initially proposed by Hamilton (1989) and Lam (1990) and we call it the Lam-Hamilton-Kim model. An alternative description was initiated by Friedman (1964; 1993) who observes that postwar fluctuations in real output should be thought of having three phases rather than two - contractions, high-growth recoveries and moderate-growth subsequent recoveries (see also DeLong and Summers (1988) and Kim and Nelson (1999)). In the sequel we will refer to this alternative model as Friedman-Kim-Nelson. More recent works on turning points detection and economic growth forecast can be found in Krolzig (2001,2004), Billio and Casarin (2010) or Billio, Ferrara, Guégan and Mazzi (2013)). Since Hamilton (1989) and his application for the study of US cycles, two regimes have been considered in many studies. On the contrary, in some recent papers which analyze the Euro area dynamics, more regimes have been suggested. For example, Billio, Casarin, Ravazzolo and Van Dijk (2012) considered Markov Switching models and in their application to US and EU industrial production data, for a period of time including the last recession, they find that four regimes (strong-recession, contraction, normal-growth, and high-growth) are necessary to identify some important features of the cycle.

The main contributions of our paper are twofold. Firstly, we show that the most used models for business cycle analysis can be comprised into the broad class of non linear model, the MSI-VARMA, and we obtain new results related to it. Specifically, we give its stable VARMA representation and the orders can be determined by evaluating the autocovariance function of the initial switching model. Secondly, we propose a new and more rigorous way for the determination of the number of regimes and apply it to analyse the US and Euro business cycles. In particular, we are able to assess that two regimes are sufficient when modeling the US business cycle but more regimes are necessary when we consider the Euro area. This is the preliminary stage to obtain correct estimation and to identify regimes.

The rest of the paper is organized as follows. In Section 2 we review some facts linking Markov switching models to business cycle analysis and thus we introduce the MS-VAR model. In Section 3 we study MSI-VARMA models starting from the baseline case of an hidden Markov chain process with distributed lags in the regime; here we give upper bounds for the stable VARMA orders both via autocovariance function and via explicit determination of the stable VARMA. Then we extend these results in Section 4 where theorems are stated for every MSI-VARMA model in which the regime variable is uncorrelated with the observable. Section 5 introduces the Lam-Hamilton-Kim and the Friedman-Kim-Nelson models of business cycle fluctuations and shows that these state space models can be expressed as MSI-VARMA. Finally, an application on the determination of the number of regimes for the US and Euro real GDP is conducted in Section 6, followed by estimation and regime identification. Section 7 concludes. Proofs are given in the Appendix.

2.2.2 Markov Switching Models and Business Cycle

Many economic time series occasionally exhibit dramatic breaks in their behavior, associated with events such as financial crises, abrupt changes in government policy or in the price of production factors. Of particular interest for economists is the statistical measurement and forecasting of business cycles. Since the early work of Bruns and Mitchell (1946), many attempts have been made to identify cycles and to provide a turning-point chronology that dates the cycle for a given country or economic area. The modern tools to deal with business cycle analysis refer to nonlinear parametric modelling, which are flexible enough to take into account certain stylized facts, such as asymmetries in the phase of the cycle. In fact, there is a large literature that uses Markov switching (MS) models to recognize business cycle phases. The starting point of this strand of literature is the recognition that there is a relationship between the concepts of changes in cyclical phases and changes in regime. The most representative works are the univariate regime-switching model proposed by Hamilton (1989) and its multivariate extension allowing both for co-movement of macroeconomic variables and switching regime as in Kim and Nelson (1999). The relationship between turning point and change in regime has been confirmed by a number of empirical studies, as Krolzig (2001, 2004) and Billio, Ferrara, Guégan and Mazzi

(2013) among the others. When using parametric models to analyze the cycle, the most used models are certainly MS-VAR. This approach do not assume any a priori definition of the business cycle: by means of the switching approach, different regimes are identified. Indeed, these regimes differ in terms of average growth rates and/or growth volatilities. Let us now introduce the MS-VAR model. Consider the K-dimensional second-order stationary dynamic process $\mathbf{y} = (\mathbf{y}_t)$ satisfying the following Markov switching autoregressive model

$$\boldsymbol{\phi}_{s_t}(L)\mathbf{y}_t = \boldsymbol{\nu}_{s_t} + \boldsymbol{\Sigma}_{s_t}\mathbf{u}_t \tag{2.1}$$

where $\mathbf{u}_t \sim IID(\mathbf{0}, \mathbf{I}_K)$ and $\boldsymbol{\phi}_{s_t}(L) = \sum_{i=0}^{p} \boldsymbol{\phi}_{s_t,i} L^i$ with $\boldsymbol{\phi}_{s_t,0} = \mathbf{I}_k$ and $\boldsymbol{\phi}_{s_t,p} \neq \mathbf{0}$. As usual, we assume that the polynomials $|\boldsymbol{\phi}_{s_t}(z)|$ have all their roots strictly outside the unit circle. Sufficient conditions ensuring second-order stationarity for Markov-switching VAR models and Markov-switching VARMA models can be found, for example, in Francq and Zakoïan (2001). The regime (s_t) follows an M-state ergodic irreducible Markov chain with $\mathbf{P} = (p_{ij})$ being the $(M \times M)$ matrix of transition probabilities $p_{ij} = Pr(s_t = j|s_{t-1} = i)$, for $i, j = 1, \ldots, M$. In general, from a statistical point of view, the order p, the number of states M, the parameters and the transition matrix $\mathbf{P}$ are unknown. However, it is established in the literature that such models admit finite-order VARMA(p^*, q^*) representations. Several authors (Krolzig (1997), Zhang and Stine (2001), Francq and Zakoïan (2001,2002), Cavicchioli (2013) have looked at the problem of finding upper bounds for p^* and q^*, expressed as functions of various parameters of the initial switching model. One possible application of such bounds are corresponding lower bounds for M which in principle could be useful in real data situation. Another method for the determination of regimes' number refers to complexity-penalized likelihood criteria, such as AIC, BIC, HQC (see Psaradakis and Spagnolo 2003, Olteanu and Rynkiewicz 2007, Rios and Rodriguez 2008). However, these criteria are not widely used in empirical literature, possibly for the computational burder required. The sample autocovariances are instead more easily calculated than maximum (penalized) likelihood estimates of the model parameters and the bounds arising from the above-mentioned elementary functions are very useful for selecting the number of regimes and the orders of the switching autoregression. In the sequel we will consider a generalization of these models where the intercept depends not only on the actual regime but also on the some previous regimes and we propose a method for the determination of regimes number. This is interesting since well-known models used for business cycle analysis can be recomprised in this framework, thus having a rigorous way for detection of regimes and estimation.

2.2.3 The MSI(M, r) - VAR(0) Model

Let us consider a generalization of the MS(M)-VAR(p) model (see Cavicchioli (2013)), for which we assume that the intercept term depends not only on the actual regime but also on the last r ($r \geq 0$) regimes

$$\boldsymbol{\nu}_t = \boldsymbol{\nu}_{s_t, s_{t-1}, \ldots, s_{t-r}} = \sum_{j=0}^{r} \boldsymbol{\nu}_{j, s_{t-j}} = \sum_{j=0}^{r} \sum_{m=1}^{M} \boldsymbol{\nu}_{jm} I(s_{t-j} = m)$$

where the indicator function $I(s_t = \ \cdot)$ takes on the value 1 if $s_t = m$ and zero otherwise. This specification is called MSI(M, r) - VAR(p) model. Here we treat the case $r > 0$ (for $r = 0$ see Cavicchioli (2013)). (The basic reference for our arguments and techniques is the Krolzig book (1997)). First we consider the K-dimensional MSI(M, r)-VAR(0) process:

$$\mathbf{y}_t = \sum_{j=0}^{r} \boldsymbol{\nu}_{j, s_{t-j}} + \boldsymbol{\Sigma}_{s_t} \mathbf{u}_t \tag{3.1}$$

where $\mathbf{u}_t \sim IID(\mathbf{0}, \mathbf{I}_K)$ and (s_t) follows an M-state ergodic irreducible Markov chain. The Markov chain follows an AR(1) process

$$\boldsymbol{\xi}_t = \mathbf{P}' \boldsymbol{\xi}_{t-1} + \mathbf{v}_t.$$

where $\mathbf{P} = (p_{ij})$ is the $(M \times M)$ matrix of transition probabilities $p_{ij} = Pr(s_t = j | s_{t-1} = i)$, for $i, j = 1, \ldots, M$, and $\boldsymbol{\xi}_t$ denotes the random $(M \times 1)$ vector whose mth element is equal to 1 if $s_t = m$ and zero otherwise. Here the innovation process $(\mathbf{v}_t)$ is a martingale difference sequence defined by $\mathbf{v}_t = \boldsymbol{\xi}_t - E(\boldsymbol{\xi}_t | \boldsymbol{\xi}_{t-1})$; it is uncorrelated with $\mathbf{u}_t$ and past values of $\mathbf{u}$, $\boldsymbol{\xi}$ or $\mathbf{y}$. The $(M \times 1)$ vector of the ergodic probabilities is denoted by $\boldsymbol{\pi} = E(\boldsymbol{\xi}_t) = (\pi_1, \ldots, \pi_M)'$. It turns out to be the eigenvector of $\mathbf{P}$ associated with the unit eigenvalue, that is, the vector $\boldsymbol{\pi}$ satisfies

$\mathbf{P}'\boldsymbol{\pi} = \boldsymbol{\pi}$. The eigenvector $\boldsymbol{\pi}$ is normalized so that its elements sum to unity. Irreducibility implies that $\pi_m > 0$, for $m = 1, \ldots, M$, meaning that all unobservable states are possible. The process in (3.1) has a first state space representation as follows

$$\begin{cases} \mathbf{y}_t = \sum_{j=0}^{r} \boldsymbol{\Lambda}_j \boldsymbol{\xi}_{t-j} + \boldsymbol{\Sigma}(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{u}_t \\ \boldsymbol{\xi}_t = \mathbf{P}'\boldsymbol{\xi}_{t-1} + \mathbf{v}_t \end{cases} \tag{3.2}$$

where $\boldsymbol{\Lambda}_j = (\boldsymbol{\nu}_{j1} \ldots \boldsymbol{\nu}_{jM})$, for $j = 0, \ldots, r$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1 \ldots \boldsymbol{\Sigma}_M)$.

Theorem 3.1 *Assume* $\mathbf{P}$ *non singular. For every* $h \geq r > 0$*, the autocovariance function of the process* $(\mathbf{y}_t)$ *in (3.1) satisfies*

$$\boldsymbol{\Gamma}_{\mathbf{y}}(h) = \mathbf{A}'(\mathbf{Q}')^h \mathbf{B}$$

where

$$\mathbf{A}' = \sum_{i=0}^{r} \boldsymbol{\Lambda}_i (\mathbf{P}')^{-i} \qquad \mathbf{B} = \sum_{j=0}^{r} (\mathbf{P}')^j \mathbf{D} \boldsymbol{\Lambda}_j' \qquad \mathbf{Q} = \mathbf{P} - \mathbf{P}_\infty \qquad \mathbf{P}_\infty = \lim_n \mathbf{P}^n.$$

We can always obtain a second state space representation in the following way:

$$\mathbf{y}_t = \sum_{j=0}^{r} \boldsymbol{\Lambda}_j (\boldsymbol{\xi}_{t-j} - \boldsymbol{\pi}) + (\sum_{j=0}^{r} \boldsymbol{\Lambda}_j)\boldsymbol{\pi} + \boldsymbol{\Sigma}((\boldsymbol{\xi}_t - \boldsymbol{\pi}) \otimes \mathbf{I}_K)\mathbf{u}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t$$

or equivalently

$$\mathbf{y}_t = (\sum_{j=0}^{r} \boldsymbol{\Lambda}_j)\boldsymbol{\pi} + \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j \boldsymbol{\delta}_{t-j} + \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t$$

where

$$\widetilde{\boldsymbol{\Lambda}}_j = (\boldsymbol{\nu}_{j1} - \boldsymbol{\nu}_{jM} \ldots \boldsymbol{\nu}_{jM-1} - \boldsymbol{\nu}_{jM}) \qquad \widetilde{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_M \ldots \boldsymbol{\Sigma}_{M-1} - \boldsymbol{\Sigma}_M).$$

Here $\boldsymbol{\delta}_t$ is the $(M-1) \times 1$ vector formed by the columns, but the last one, of $\boldsymbol{\xi}_t - \boldsymbol{\pi}$. Of course, the last formula is obtained by using the restrictions $\mathbf{i}_M'\boldsymbol{\xi}_t = 1$ and $\mathbf{i}_M'\boldsymbol{\pi} = 1$, where $\mathbf{i}_M'$ denotes the $(M \times 1)$ vector of ones. So we get

$$
\begin{cases}
\mathbf{y}_t = (\sum_{j=0}^{r} \mathbf{\Lambda}_j)\boldsymbol{\pi} + \sum_{j=0}^{r} \widetilde{\mathbf{\Lambda}}_j \boldsymbol{\delta}_{t-j} + \widetilde{\mathbf{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t + \mathbf{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t \\
\boldsymbol{\delta}_t = \mathbf{F}\boldsymbol{\delta}_{t-1} + \mathbf{w}_t
\end{cases}
\tag{3.3}
$$

where $E(\mathbf{w}_t\mathbf{w}_t^{'}) = \widetilde{\mathbf{D}} - \mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}^{'}$, $\quad E(\mathbf{w}_t\mathbf{w}_\tau^{'}) = \mathbf{0}$ for $t \neq \tau$,

$$
\widetilde{\mathbf{D}} = \begin{pmatrix}
\pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_{M-1} \\
-\pi_1\pi_2 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_{M-1} \\
\vdots & \vdots & & \vdots \\
-\pi_{M-1}\pi_1 & -\pi_{M-1}\pi_2 & \cdots & \pi_{M-1}(1-\pi_{M-1})
\end{pmatrix}
$$

and

$$
\mathbf{F} = \begin{pmatrix}
p_{11} - p_{M1} & \cdots & p_{M-1,1} - p_{M1} \\
\vdots & & \vdots \\
p_{1,M-1} - p_{M,M-1} & \cdots & p_{M-1,M-1} - p_{M,M-1}
\end{pmatrix}
$$

which is an $(M-1) \times (M-1)$ matrix with all eigenvalues inside the unit circle. Since $\boldsymbol{\delta}_t$ has zero mean, the unconditional expectation of the initial process is given by $E(\mathbf{y}_t) = \boldsymbol{\mu}_{\mathbf{y}} = (\sum_{j=0}^{r} \mathbf{\Lambda}_j)\boldsymbol{\pi}$, as before. By iteration of the transition equation in (3.3), we also obtain $E(\boldsymbol{\delta}_t\boldsymbol{\delta}_{t+h}^{'}) = \widetilde{\mathbf{D}}(\mathbf{F}^{'})^h$ for every $h \geq 0$.

For Model (3.1), we obtain the following main result:

Theorem 3.2. *The second-order stationary dynamic process defined in (3.1), with $r > 0$, has a stable VARMA(p^*, q^*) representation, where $p^* \leq M-1$ and $q^* \leq M+r-2$. If the lag polynomials of the AR and MA parts of the VARMA(p^*, q^*) have no roots in common, equalities hold in the previous relations, and the identification problem is completely solved, that is, $M = p^* + 1$ and $r = q^* - p^* + 1$ (in this case, we have $q^* \geq p^*$).*

Now we also determine explicitly a stable VARMA(p^*, q^*) representation for the process $(\mathbf{y}_t)$ in (3.1), with $r > 0$. In our computation, the autoregressive lag polynomial of such a stable VARMA is shown to be scalar.

Theorem 3.3. *The second-order stationary dynamic process defined in (3.1), with $r > 0$, has* *a stable VARMA(p^*, q^*) representation, with $p^* \leq M-1$ and $q^* \leq M+r-2$,*

$$\gamma(L)(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = \mathbf{C}(L)\boldsymbol{\epsilon}_t$$

where $\gamma(L) = |F(L)|$ is a scalar polynomial of degree $M-1$ in L (that is, the determinant of the matrix $F(L) = \mathbf{I}_{M-1} - \mathbf{F}L$, where L is the lag operator) and $\mathbf{C}(L)$ is the $[K \times (KM+M-1)]$-dimensional lag polynomial matrix of degree $M+r-2$ in L given by

$$\mathbf{C}(L) = [\sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j F(L)^* L^j \qquad \widetilde{\boldsymbol{\Sigma}}(F(L)^* \otimes \mathbf{I}_K) \qquad |F(L)|\boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)]$$

where $F(L)^*$ is the adjoint matrix of $F(L)$. Furthermore, $\boldsymbol{\epsilon}_t = (\mathbf{w}_t^{'} \quad \mathbf{u}_t^{'}(\mathbf{w}_t^{'} \otimes \mathbf{I}_K) \quad \mathbf{u}_t^{'})^{'}$ is a zero mean white noise process with $\mathrm{var}(\boldsymbol{\epsilon}_t) = \mathrm{diag}(\widetilde{\mathbf{D}} - \mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}^{'}, (\widetilde{\mathbf{D}} - \mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}^{'}) \otimes \mathbf{I}_K, \mathbf{I}_K)$. If $\gamma(L)$ and $\mathbf{C}(L)$ are coprime, then $p^* = M-1$ and $q^* = M+r-2$, and the identification problem is completely solved.

2.2.4 The MSI(M, r)-VARMA(p, q) Model

Firstly, we consider the following K-dimensional second-order stationary MSI(M, r)-VAR(p) process, with $r > 0$ and $p > 0$,

$$\boldsymbol{\phi}_{s_t}(L)\mathbf{y}_t = \sum_{j=0}^{r} \boldsymbol{\nu}_{j,s_{t-j}} + \boldsymbol{\Sigma}_{s_t}\mathbf{u}_t \tag{4.1}$$

where $\mathbf{u}_t \sim IID(\mathbf{0}, \mathbf{I}_K)$ and $\boldsymbol{\phi}_{s_t}(L) = \boldsymbol{\phi}_{0,s_t} + \boldsymbol{\phi}_{1,s_t}L + \cdots + \boldsymbol{\phi}_{p,s_t}L^p$ with $\boldsymbol{\phi}_{0,s_t} = \mathbf{I}_K$ and $\boldsymbol{\phi}_{p,s_t} \neq \mathbf{0}$. As usual, we assume that the polynomials $|\boldsymbol{\phi}_{s_t}(z)|$ have all their roots strictly outside the unit circle. Sufficient conditions ensuring second-order stationarity for Markov-switching VARMA models can be found, for example, in Francq and Zakoïan (2001) .

For every $j = 0, \ldots, r$, define $\boldsymbol{\Lambda}_j = (\boldsymbol{\nu}_{j1} \ldots \boldsymbol{\nu}_{jM})$. Then define $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1 \ldots \boldsymbol{\Sigma}_M)$ and

$$\boldsymbol{\phi}(L) = [\mathbf{I}_K + \boldsymbol{\phi}_{1,1}L + \cdots + \boldsymbol{\phi}_{p,1}L^p \ldots \mathbf{I}_K + \boldsymbol{\phi}_{1,M}L + \cdots + \boldsymbol{\phi}_{p,M}L^p].$$

The process in (4.1) has a first state space representation as follows

$$\begin{cases} \phi(L)(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{y}_t = \sum_{j=0}^{r} \boldsymbol{\Lambda}_j \boldsymbol{\xi}_{t-j} + \boldsymbol{\Sigma}(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{u}_t \\ \boldsymbol{\xi}_t = \mathbf{P}^{'}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t \end{cases} \tag{4.2}$$

Moreover, we can always obtain a second state space representation:

$$\begin{cases} \widetilde{\boldsymbol{\phi}}(L)(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{y}_t + \boldsymbol{\phi}(L)(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{y}_t \\ \qquad = (\sum_{j=0}^{r} \boldsymbol{\Lambda}_j)\boldsymbol{\pi} + \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j \boldsymbol{\delta}_{t-j} + \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_K)\mathbf{u}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_t \\ \boldsymbol{\delta}_t = \mathbf{F}\boldsymbol{\delta}_{t-1} + \mathbf{w}_t \end{cases} \tag{4.3}$$

where $\boldsymbol{\delta}_t$, $\widetilde{\boldsymbol{\Lambda}}_j$ and $\widetilde{\boldsymbol{\Sigma}}$ are defined as in the previous section, and

$$\widetilde{\boldsymbol{\phi}}(L) = [(\boldsymbol{\phi}_{1,1} - \boldsymbol{\phi}_{1,M})L + \cdots + (\boldsymbol{\phi}_{p,1} - \boldsymbol{\phi}_{p,M})L^p \ \dots \ (\boldsymbol{\phi}_{1,M-1} - \boldsymbol{\phi}_{1,M})L + \cdots + (\boldsymbol{\phi}_{p,M-1} - \boldsymbol{\phi}_{p,M})L^p].$$

Theorem 4.1. *Let $\mathbf{F}$ be non singular. For every $h \geq r > 0$, assume that the regime variable $\boldsymbol{\xi}_{t+h}$ is uncorrelated with $\mathbf{y}_t$. Then the autocovariance function of the second-order stationary process in (4.1) satisfies*

$$B(L)\boldsymbol{\Gamma}_{\mathbf{y}}(h) = \mathbf{A}^{'}\mathbf{F}^h\mathbf{B}$$

where $\mathbf{A}^{'} = \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j \mathbf{F}^{-j}$ and $\mathbf{B} = E(\boldsymbol{\delta}_t \boldsymbol{y}_t^{'})$, which are assumed to be nonzero matrices.

Now, applying Theorem 2.2 from Cavicchioli (2013) for $q = r$ and taking in mind that $\mathbf{F}$ is $(M-1) \times (M-1)$, we have the following main result for model (4.1):

Theorem 4.2. *Under the hypothesis that the regime variable is uncorrelated with the observable, the K-dimensional second-order stationary process in (4.1), with $r > 0$ and $p > 0$, admits a stable VARMA(p^*, q^*) representation with $p^* \leq M+p-1$ and $q* \leq M+r-2$. If we require that autoregressive lag polynomial of such a stable representation is scalar, then the bounds become $p^* \leq M + Kp - 1$ and $q* \leq M + (K-1)p + r - 2$. If the lag polynomials of the AR and MA parts of the former VARMA(p^*, q^*) representation have no roots in common, equalities hold in the previous relations, that is, $p^* = M + p - 1$ and $q^* = M + r - 2$.*

To end the section we compute explicitly a VARMA representation for a general MSI-VARMA. This gives a new proof of Theorem 3.3 (case $q = 0$) and extends Proposition 5 from Krolzig (1997), Section 10.2.2. We start with the more simple case in which the autoregressive lag polynomial of the initial process is state independent.

Theorem 4.3. *Let* $\mathbf{y} = (\mathbf{y}_t)$ *be an* K*–dimensional second-order stationary MSI(*M, r*)-VARMA(*p, q*) process, with* $r > 0$,

$$\mathbf{A}(L)\mathbf{y}_t = \sum_{j=0}^{r} \boldsymbol{\nu}_{j,s_{t-j}} + \sum_{i=0}^{q} \boldsymbol{\Theta}_{s_t}(L)\mathbf{u}_t$$

where $\mathbf{A}(L) = \sum_{\ell=0}^{p} \mathbf{A}_\ell L^\ell$ *with* $\boldsymbol{A}_0 = \mathbf{I}_K$, $|\mathbf{A}_p| \neq \mathbf{0}$, *and* $\boldsymbol{\Theta}_{s_t}(L) = \sum_{i=0}^{q} \boldsymbol{\Theta}_{s_t,i} L^i$, *with* $\boldsymbol{\Theta}_{s_t,0} = \boldsymbol{\Sigma}_{s_t}$ *(nonsingular symmetric* $K \times K$ *matrix) and* $|\boldsymbol{\Theta}_{s_t,q}| \neq \mathbf{0}$ *are full rank matrix lag polynomials. Under quite general regularity conditions, the dynamic process* $\mathbf{y} = (\mathbf{y}_t)$ *admits a stable VARMA(*p^*, q^**) representation, with* $p^* \leq M + Kp - 1$ *and* $q* \leq M + (K-1)p + \max\{r, q+1\} - 2$,

$$\gamma(L)(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = \mathbf{C}(L)\boldsymbol{\epsilon}_t$$

where $\gamma(L) = |F(L)||\mathbf{A}(L)|$ *is the scalar AR operator of degree* $M + Kp - 1$, $\mathbf{C}(L)$ *is a matrix lag polynomial of degree* $M + (K-1)p + \max\{r, q+1\} - 2$, *and* $\boldsymbol{\epsilon}_t$ *is a zero mean vector white noise process.* In the general case in which the autoregression part of the process in Theorem 4.3 is state dependent but the regime variable is uncorrelated with the observable, we can proceed as follows. By Theorem 4.1 the autocovariances of the process satisfy a finite difference equation of order $p^* = M + p - 1$ and rank $q^* + 1 = M + \max\{r, q+1\} - 1$. Then the process can be represented by a stable VARMA(p^*, q^*). Given the process $(\mathbf{y}_t)$, we can estimate the coefficients of the stable VARMA(p^*, q^*) with usual procedures. If there is no cancellation between the AR and MA part of the estimated VARMA(p^*, q^*), then we get the representation as in Theorem 4.3 with equalities.

2.2.5 Business Cycle Models

In this Section we show that business cycle models widely used in empirical works can be rewritten as MSI-VARMA. Therefore, we can formally test the number of regimes as well as lags in the intercept or autoregressive lags, given the above results. This avoids informal determination of regimes' number by the researcher and leads to correct estimation results and possibly to reliable forecasting conclusions.

(5.1) *The Lam-Hamilton-Kim model.* In modeling the time series behaviour of real Gross National Product (GNP) of United States, Hamilton (1989) considered the case in which real GNP is generated by the sum of two independent unobserved components, one following an autoregressive process with a unit root, and the other following a random walk with a Markov switching error term. Lam (1990) generalized the Hamilton model to the case in which the autoregressive component need not contain a unit root. Finally, Kim (1994) showed that the Lam-Hamilton model can be written in the following state space form, which we shall call *the Lam-Hamilton-Kim model* :

$$\begin{cases} y_t = \mathbf{H}\mathbf{x}_t + \beta_{s_t} \\ \mathbf{x}_t = \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{e}_t \end{cases} \tag{5.1}$$

where y_t is scalar ($K = 1$), $\mathbf{x}_t = (x_t\ x_{t-1} \cdots x_{t-r+1})'$ and $\mathbf{e}_t = (u_t\ 0 \cdots 0)'$ are $r \times 1$, with $u_t \sim IID(0, \sigma^2)$, $\mathbf{H} = (1\ \ -1\ 0 \cdots 0)$ is $1 \times r$ and

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_r \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

with $\phi_i \in \mathbb{R}$, $i = 1, \ldots, r$, $(r \geq 2)$. Here $\beta_{s_t} = \delta_0 + \delta_1 s_t$, where $\delta_j \in \mathbb{R}$, $j = 0, 1$, and (s_t) is an M-state Markov chain. Kim (1994) estimated this model by using suitable (filtering and smoothing) algorithms that he also constructed for more general state space representations with Markov switching. Now we are going to show that Model (5.1) can be interpreted as a model with distributed lags in the regime. Then we give a stable VARMA representation of it. This allows us to estimate the model via MLE, i.e., a very simple procedure which is alternative to that employed by Kim (1994). From the transition equation in (5.1) we get

$$\Phi(L)\mathbf{x}_t = \mathbf{e}_t$$

hence

$$|\Phi(L)|\mathbf{x}_t = \Phi(L)^*\mathbf{e}_t \tag{5.2}$$

where $\Phi(L) = \mathbf{I}_r - \mathbf{\Phi}L$. Substituting (5.2) into the measurement equation in (5.1) after pre-multiplying by the determinant of $\Phi(L)$ gives

$$|\Phi(L)|y_t = |\Phi(L)|\beta_{s_t} + \mathbf{H}\Phi(L)^*\mathbf{e}_t \tag{5.3}$$

which is an MSI(M, r)-VARMA(p, q) model in the sense of Section 4, where $p = r$ and $q = r - 1$ (recall $r \geq 2$). So Theorem 4.3 directly implies that the Lam-Hamilton-Kim model admits a stable VARMA(p^*, q^*) representation (whose autoregressive lag polynomial is scalar) with $p^* \leq M + r - 1$ and $q^* \leq M + \max\{p, p\} - 2 = M + p - 2$.

We now determine explicitly the final form of this stable VARMA and we need some notation. Define $\boldsymbol{\beta} = (\beta_1 \cdots \beta_M)$, where $\beta_m = \delta_0 + \delta_1 m$ for every $m = 1, \ldots, M$. Then we get $\beta_{s_t} = \boldsymbol{\beta}\boldsymbol{\xi}_t$. Substituting this relation into (5.3) yields

$$\begin{aligned} |\Phi(L)|y_t &= |\Phi(L)|\boldsymbol{\beta}\boldsymbol{\xi}_t + \mathbf{H}\Phi(L)^*\mathbf{e}_t \\ &= |\Phi(L)|\boldsymbol{\beta}(\boldsymbol{\xi}_t - \boldsymbol{\pi}) + |\Phi(1)|\boldsymbol{\beta}\boldsymbol{\pi} + \mathbf{H}\Phi(L)^*\mathbf{e}_t \\ &= |\Phi(L)|\widetilde{\boldsymbol{\beta}}\boldsymbol{\delta}_t + |\Phi(1)|\boldsymbol{\beta}\boldsymbol{\pi} + \mathbf{H}\Phi(L)^*\mathbf{e}_t \end{aligned}$$

where $\widetilde{\boldsymbol{\beta}} = (\beta_1 - \beta_M \cdots \beta_{M-1} - \beta_M)$. Then we get the state space representation

$$\begin{cases} |\Phi(L)|(y_t - \mu_y) = |\Phi(L)|\widetilde{\boldsymbol{\beta}}\boldsymbol{\delta}_t + \mathbf{H}\Phi(L)^*\mathbf{e}_t \\ \boldsymbol{\delta}_t = \mathbf{F}\boldsymbol{\delta}_{t-1} + \mathbf{w}_t \end{cases} \tag{5.4}$$

where $|\Phi(L)|\mu_y = |\Phi(1)|\boldsymbol{\beta}\boldsymbol{\pi}$ (in fact, $\mu_y = \boldsymbol{\beta}\boldsymbol{\pi}$). Substituting $\boldsymbol{\delta}_t = F(L)^{-1}\mathbf{w}_t$ into the measurement equation in (5.4) and premultiplying by $|F(L)|$ yield

$$|F(L)||\Phi(L)|(y_t - \mu_y) = \widetilde{\boldsymbol{\beta}}|\Phi(L)|F(L)^* \mathbf{w}_t + \mathbf{H}|F(L)|\Phi(L)^* \mathbf{e}_t$$

which is a stable VARMA(p^*, q^*) with $p^* \leq M + r - 1$ and $q^* \leq M + r - 2$, and whose autoregressive lag polynomial is scalar. Summarizing we have the following result:

Theorem 5.1. *The Lam-Hamilton-Kim model for US real GNP has a stable VARMA(p^*, q^*) representation*

$$\gamma(L)(y_t - \mu_y) = \mathbf{C}(L)\boldsymbol{\epsilon}_t$$

with $p^* \leq M + r - 1$ and $q* \leq M + r - 2$. Under quite general regularity conditions, $\gamma(L) = |F(L)||\Phi(L)|$ is the scalar AR operator of degree $M + r - 1$ in the lag operator L and $\mathbf{C}(L)$ is a $[1 \times (M + r - 1)]$-dimensional matrix lag polynomial of degree $M + r - 2$ in L given by

$$\mathbf{C}(L) = [\mathbf{H}|F(L)|\Phi(L)^* \qquad \widetilde{\boldsymbol{\beta}}|\Phi(L)|F(L)^*]$$

and $\boldsymbol{\epsilon}_t = (\mathbf{e}_t' \quad \mathbf{w}_t')'$ is a zero mean $(M + r - 1) \times 1$ vector white noise process. If $\gamma(L)$ and $\mathbf{C}(L)$ are coprime, then equalities hold in the previous relations, hence $p^* = M + r - 1$ and $q^* = p^* - 1$.

(5.2) *The Friedman-Kim-Nelson model.* Friedman's plucking model (1964) of business fluctuations suggests that output cannot exceed a ceiling level, and it is occasionally plucked downward by recession. The model implies that business fluctuations are asymmetric, that recessions have only a temporary effect on output, and that recessions are duration dependent while expansions are not. Subsequent literature has provided copious empirical support for these statements. See, for example, De Simone and Clarke (2007) and its references. Kim and Nelson (1998) showed that the Friedman model can be written in the following state space form, which we shall call *the Friedman-Kim-Nelson model* :

$$\begin{cases} y_t = \mathbf{H}\mathbf{x}_t \\ \mathbf{x}_t = \boldsymbol{\mu}_{s_t} + \boldsymbol{\Phi}\mathbf{x}_{t-1} + \boldsymbol{\Sigma}_{s_t}\mathbf{e}_t \end{cases} \tag{5.5}$$

where y_t is scalar ($K = 1$), $\mathbf{x}_t$, $\boldsymbol{\mu}_{s_t}$ and $\mathbf{e}_t$ are 4×1 with $\mathbf{e}_t \sim IID(\mathbf{0}, \mathbf{I}_4)$, $\boldsymbol{\Sigma}_{s_t}$ is a 4×4 diagonal matrix, $\mathbf{H} = (1\ 1\ 0\ 0)$, and

$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & \phi_1 & \phi_2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

with $\phi_i \in \mathbb{R}$, $i = 1, 2$; here (s_t) is an M-state Markov chain ($M = 2$ in the quoted papers). Kim and Nelson (1998) estimated the model by using Kim's approximate MLE. See also Kim (1994). We show that Model (5.5) can be viewed as a model with distributed lags in the regime and then we give a stable VARMA representation of it. From the transition equation in (5.5) we get

$$\Phi(L)\mathbf{x}_t = \boldsymbol{\mu}_{s_t} + \boldsymbol{\Sigma}_{s_t}\mathbf{e}_t$$

where

$$\Phi(L) = \mathbf{I}_4 - \boldsymbol{\Phi}L = \begin{pmatrix} 1-L & 0 & 0 & -L \\ 0 & 1-\phi_1 L & -\phi_2 L & 0 \\ 0 & -L & 1 & 0 \\ 0 & 0 & 0 & 1-L \end{pmatrix}$$

hence $|\Phi(L)| = (1-L)^2(1-\phi_1 L - \phi_2 L^2)$. Premultiplying by the adjoint matrix $\Phi(L)^*$ we obtain

$$|\Phi(L)|\mathbf{x}_t = \Phi(L)^*\boldsymbol{\mu}_{s_t} + \Phi(L)^*\boldsymbol{\Sigma}_{s_t}\mathbf{e}_t \tag{5.6}$$

where

$$\Phi(L)^* = \begin{pmatrix} (1-L)(1-\phi_1 L-\phi_2 L^2) & 0 & 0 & L(1-\phi_1 L - \phi_2 L^2) \\ 0 & (1-L)^2 & \phi_2 L(1-L)^2 & 0 \\ 0 & L(1-L)^2 & (1-\phi_1 L)(1-L)^2 & 0 \\ 0 & 0 & 0 & (1-L)(1-\phi_1 L - \phi_2 L^2) \end{pmatrix}.$$

Substituting (5.6) into the measurement equation in (5.5) gives

$$|\Phi(L)|y_t = \mathbf{H}\Phi(L)^*\boldsymbol{\mu}_{s_t} + \mathbf{H}\Phi(L)^*\boldsymbol{\Sigma}_{s_t}\mathbf{e}_t$$

which is an MSI(M,r)-VARMA(p,q) with $p = 4$ and $r = q = 3$. So Theorem 4.3 implies that such a model has a stable VARMA(p^*, q^*) representation, with $p^* \leq M + p - 1 = M + 3$ and $q^* \leq M + \max\{r, q+1\} - 2 = M + 2$. We now determine explicitly the final form of this stable VARMA. Model (5.5) has the following state space representation

$$\begin{cases} y_t = \mathbf{H}\mathbf{x}_t \\ \mathbf{x}_t = \widetilde{\boldsymbol{\mu}}\boldsymbol{\delta}_t + \boldsymbol{\mu}\boldsymbol{\pi} + \boldsymbol{\Phi}\mathbf{x}_{t-1} + \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_4)\mathbf{e}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_4)\mathbf{e}_t \\ \boldsymbol{\delta}_t = \mathbf{F}\boldsymbol{\delta}_{t-1} + \mathbf{w}_t \end{cases} \tag{5.7}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 \cdots \boldsymbol{\mu}_M)$ is $4 \times M$, $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1 \cdots \boldsymbol{\Sigma}_M)$ is $4 \times (4M)$, and $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$ are defined in the usual way. Furthermore, we have $E(y_t) = \mathbf{H}E(\mathbf{x}_t)$ and $\Phi(L)E(\mathbf{x}_t) = \boldsymbol{\mu}\boldsymbol{\pi}$, hence $y_t - E(y_t) = \mathbf{H}(\mathbf{x}_t - E(\mathbf{x}_t))$. From (5.7) we get

$$\begin{aligned} \Phi(L)(\mathbf{x}_t - E(\mathbf{x}_t)) &= \widetilde{\boldsymbol{\mu}}\boldsymbol{\delta}_t + \widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\delta}_t \otimes \mathbf{I}_4)\mathbf{e}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_4)\mathbf{e}_t \\ &= \widetilde{\boldsymbol{\mu}}F(L)^{-1}\mathbf{w}_t + \widetilde{\boldsymbol{\Sigma}}(F(L)^{-1}\mathbf{w}_t \otimes \mathbf{I}_4)\mathbf{e}_t + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_4)\mathbf{e}_t \end{aligned}$$

hence

$$|F(L)|\Phi(L)(\mathbf{x}_t - E(\mathbf{x}_t)) = \widetilde{\boldsymbol{\mu}}F(L)^*\mathbf{w}_t + \widetilde{\boldsymbol{\Sigma}}(F(L)^*\mathbf{w}_t \otimes \mathbf{I}_4)\mathbf{e}_t + \boldsymbol{\Sigma}|F(L)|(\boldsymbol{\pi} \otimes \mathbf{I}_4)\mathbf{e}_t.$$

Premultiplying by the determinant of $\Phi(L)$ yields

$$|F(L)||\Phi(L)|(\mathbf{x}_t - E(\mathbf{x}_t)) = \Phi(L)^*\widetilde{\boldsymbol{\mu}}F(L)^*\mathbf{w}_t + \Phi(L)^*\widetilde{\boldsymbol{\Sigma}}(F(L)^*\mathbf{w}_t \otimes \mathbf{I}_4)\mathbf{e}_t + \Phi(L)^*\boldsymbol{\Sigma}|F(L)|(\boldsymbol{\pi} \otimes \mathbf{I}_4)\mathbf{e}_t.$$

Assuming $\mu_y = E(y_t)$ time invariant, we obtain

$$\begin{aligned} &|F(L)||\Phi(L)|(y_t - \mu_y) = \mathbf{H}|F(L)||\Phi(L)|(\mathbf{x}_t - E(\mathbf{x}_t)) \\ &= \mathbf{H}\Phi(L)^*\widetilde{\boldsymbol{\mu}}F(L)^*\mathbf{w}_t + \mathbf{H}\Phi(L)^*\widetilde{\boldsymbol{\Sigma}}(F(L)^* \otimes \mathbf{I}_4)(\mathbf{w}_t \otimes \mathbf{I}_4)\mathbf{e}_t + \mathbf{H}\Phi(L)^*\boldsymbol{\Sigma}|F(L)|(\boldsymbol{\pi} \otimes \mathbf{I}_4)\mathbf{e}_t \end{aligned}$$

which is a stable VARMA(p^*, q^*) with $p^* \leq M+3$ and $q^* \leq M+2$, and whose autoregressive lag polynomial is scalar. Summarizing we have the following result:

Theorem 5.2. *The Friedman-Kim-Nelson model of business fluctuations admits a stable VARMA(p^*, q^*) representation*

$$\gamma(L)(y_t - \mu_y) = \mathbf{C}(L)\boldsymbol{\epsilon}_t$$

with $p^* \leq M+3$ and $q* \leq M+2$. Under quite general regularity conditions, $\gamma(L) = |F(L)||\Phi(L)|$ is the scalar AR operator of degree $M+3$ in the lag operator L and $\mathbf{C}(L)$ is a $[1 \times (5M-1)]$-dimensional matrix lag polynomial of degree $M+2$ in L given by

$$\mathbf{C}(L) = [\mathbf{H}\Phi(L)^* \widetilde{\boldsymbol{\mu}} F(L)^* \qquad \mathbf{H}\Phi(L)^* \widetilde{\boldsymbol{\Sigma}}(F(L)^* \otimes \mathbf{I}_4) \qquad \mathbf{H}\Phi(L)^* \boldsymbol{\Sigma}|F(L)|(\boldsymbol{\pi} \otimes \mathbf{I}_4)]$$

and $\boldsymbol{\epsilon}_t = (\mathbf{w}_t^{'} \quad \mathbf{e}_t^{'}(\mathbf{w}_t^{'} \otimes \mathbf{I}_4) \quad \mathbf{e}_t^{'})^{'}$ is a zero mean $(5M-1) \times 1$ vector white noise process. If $\gamma(L)$ and $\mathbf{C}(L)$ are coprime, then equalities hold in the previous relations, hence $p^* = M+3$ and $q^* = p^* - 1$.

2.2.6 Empirical Application

In our study we consider the Gross Domestic Product (GDP) from FRED at a quarterly frequency for the United States (US), from 1951:1 to 2012:4, and from EUROSTAT for the European Union (EU 12), from 1973:1 to 2012:4 [1]. The presence of unit root in the data has been checked by augmented Dickey-Fuller (ADF) test which points out the non-stationarity of both series. For the null hypothesis of unit roots, the test statistic gives 1.724 (with $p = 12$) for US GDP and -0.4037 (with p=5) for EU GDP. In both cases the null hypothesis of a unit root cannot be rejected. For differenced time series, the ADF test rejects the unit root hypothesis on the 1% significance level (with test statistics of -4.4853 for US and -5.3384 for EU). In the following analysis we consider the growth rate of quarterly real GDP data for US and EU and the series are plotted in Figure 1 and Figure 2.

[1] Data are taken from Fred website: $< research.stlouisfed.org >$ and from Eurostat website: $< http : //epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/ >$. Data are seasonally adjusted.

	$ARMA(\hat{p}^*, \hat{q}^*)$	$\hat{M}$	$\hat{r}$
US	ARMA(1,1)	2	1
EU	ARMA(3,2)	4	0

Table 2.8: Estimated number of regimes and distributed lags in the intercept for US and EU GDP. The time lags are from 1951:1 to 2012:4 for US GDP and from 1973:1 to 2012:4 for EU GDP. The procedure uses the bounds obtained in Section 3.

We model both series as a MSI(M,r)-AR(0), with state-dependent mean and variance and no autoregressive part. This follows from several empirical studies which find that most part of the forecast errors is due to time changes in some parameters of the prediction models. In particular, we follow Krolzig (2000) and Anas et al. (2008) where only intercept and volatility are assumed to be driven by a regime-switching variable. In fact, with regards to the Euro area, Anas et al. (2008) and Billio et al. (2013) find that allowing regime switching-autoregressive coefficients deteriorates the detection of the business cycle turning points. Note that we can now apply the bounds proposed in Section 3 which simultaneously define number of regimes and lags in the intercept. Those can be obtained having estimates of the stable VARMA orders $\hat{p}^*$ and $\hat{q}^*$ as follows

$$\begin{cases} \hat{M} = \hat{p}^* + 1 \\ \hat{r} = 1 - \hat{p}^* + \hat{q}^*. \end{cases}$$

For the computation of the orders of the stable VARMA we use the 3-pattern method (TPM) proposed by Choi (1992). This gives the results reported in Table 1.

Our results suggest that US real GDP is sufficiently good described with two regimes and one lag in the intercept, which is in line with the estimated Markov-switching State Space models of the previous section and with several empirical works studying US business cycle. Then we are not going to indagate further. On the contrary, when modeling GDP of the Euro Area four regimes are more appropriate. In order to identify regimes for the European economy, we proceed with the estimation of a MSI(4,0)-AR(0) model, as suggested from the above step. Tables 2 and 3 report estimated parameters and their standard errors, the transition matrix and the expected duration of the regimes. Moreover, Figure 2 plots the smoothed probabilities

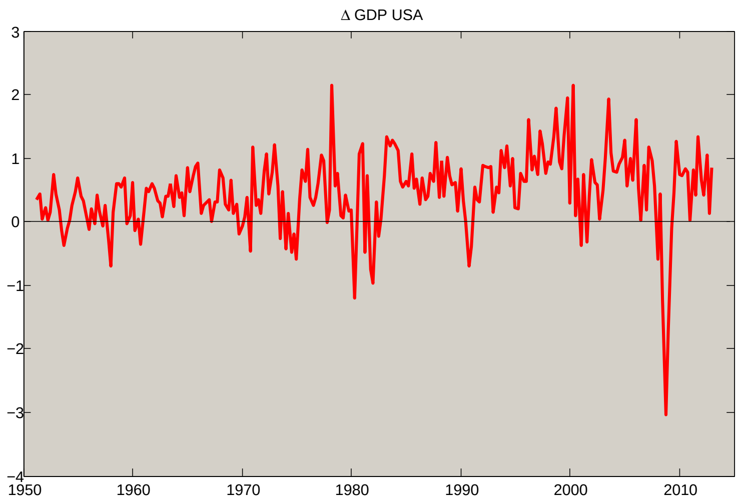

Figure 2.1: US quarterly growth rate of real GDP for the period 1951:1 - 2012:4. Data are taken from Fred website.

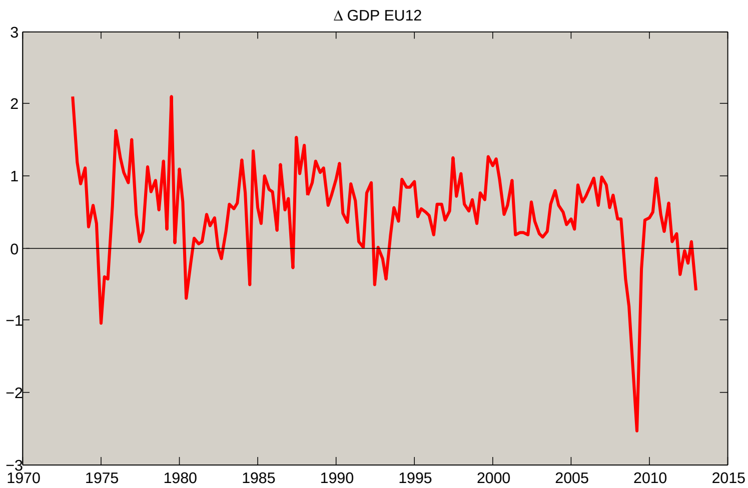

Figure 2.2: European (EU12) quarterly growth rate of real GDP for the period 1973:1 - 2012:4. Data are taken from Eurostat website.

Mean		St.Deviation	
μ_1	-0.437	σ_1	1.274
	(0.604)		(0.322)
μ_2	-0.180	σ_2	0.260
	(0.132)		(0.048)
μ_3	0.394	σ_3	0.184
	(0.058)		(0.042)
μ_4	0.893	σ_4	0.289
	(0.038)		(0.031)

Table 2.9: Estimated coefficients of the MSI(4,0)-AR(0) model for EU quarterly growth rate of real GDP for the period 1973:1 - 2012:4. Standard errors are in parenthesis. The log-likelihood value is 268.4987.

	Regime 1	Regime 2	Regime 3	Regime 4
Regime 1	0.49	0.00	0.09	0.00
	(0.45)	(0.01)	(0.06)	(0.01)
Regime 2	0.51	0.58	0.02	0.05
	(0.33)	(0.26)	(0.02)	(0.03)
Regime 3	0.00	0.20	0.68	0.18
	(0.01)	(0.12)	(0.16)	(0.07)
Regime 4	0.00	0.22	0.22	0.76
	(0.01)	(0.12)	(0.09)	(0.13)

Table 2.10: Transition probability matrix of the estimated MSI(4,0)-AR(0) model for EU quarterly growth rate of real GDP for the period 1973:1 - 2012:4. Standard errors are in parenthesis.

of the four regimes. All regimes can be matched with a plausible economic interpretation. In fact, recession phases are identified with Regime 1 and 2, being Regime 1 a stronger recession than Regime 2 (contraction), while Regime 3 is moderate/normal growth and Regime 4 is high growth. The expansionary regimes are more persistent than the others, in fact probabilities of staying in those regimes are 0.68 and 0.76. Note also that all volatility coefficients are significant at 1% level. The expected duration of recession phases is 4/5 quarters, while expansionary regimes cover about 7 quarters. Here we detect turning points as the last quarter of each regime phase and the following recession periods can be inferred from the estimation: 1974:1 - 1975:3, 1979:1 - 1981:3, 1982:1 - 1983:1, 1991:1 - 1993:3, 2008:1 - 2010:1 and 2011:1 - 2012:4. These conclusions are in line with well-recognised recession phases, see, for instance, Anas et al.(2007).

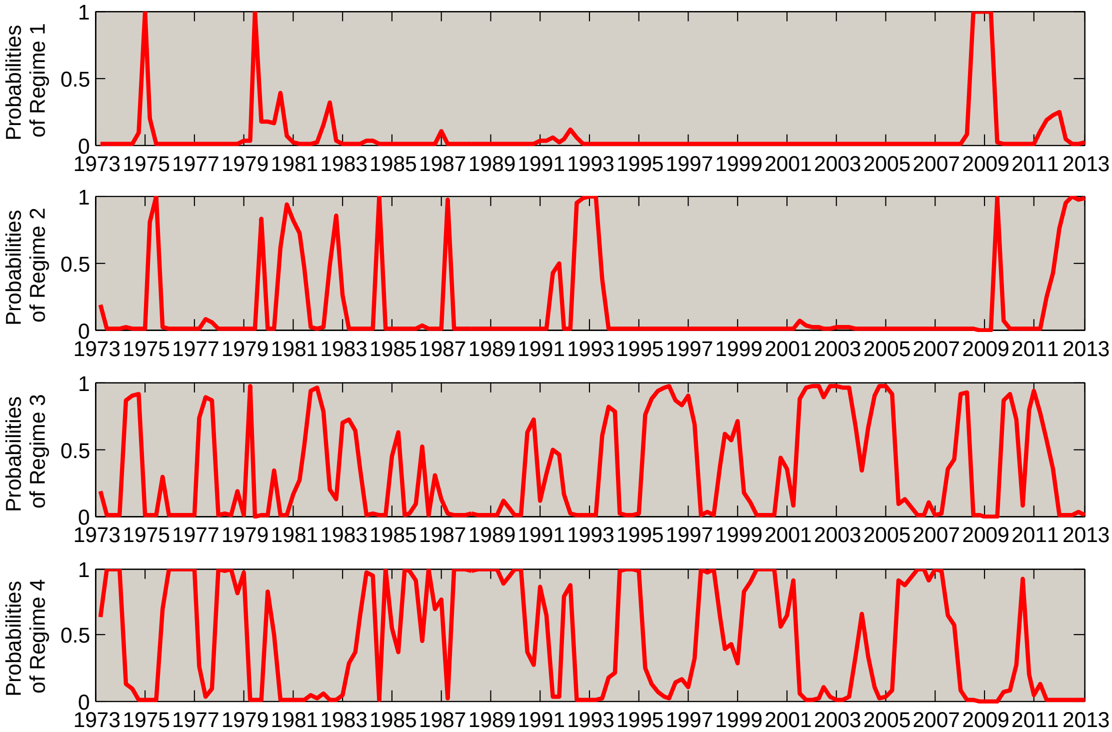

Figure 2.3: Smoothed probabilities from the estimated MSI(4,0)-AR(0) model for EU quarterly growth rate of real GDP for the period 1973:1 - 2012:4. Data are taken from Eurostat. Regime 1 corresponds to strong recession, Regime 2 to contraction, Regime 3 to moderate/normal growth and Regime 4 to high growth.

2.2.7 Conclusion

In this paper, for a general class of Markov-switching VARMA models with distributed lags in the regime, in symbols MSI(M, r)- VARMA(p, q), we give finite order VARMA(p^*, q^*) representations where the parameters can be determined by evaluating the autocovariance function of the Markov-switching models. It turns out that upper bounds for p^* and q^* are elementary functions of the dimension K of the process, the number M of regimes, the number of regimes r on the intercept and the orders p and q. If there is no cancellation, the bounds become equalities, and this solves the identification problem. This result produces an easy method for setting a lower bound on the number of regimes from the estimated autocovariance function. Of particular interest is how some well-known state space systems, introduced in the literature for business cycle analysis, are shown to be comprised in this general MSI-VARMA model, such as the Lam-Hamilton-Kim and the Friedman-Kim-Nelson models of business fluctuations. In the application we determine the number of regimes which turns out to be more appropriate for the description of US and EU economic systems by using the bounds obtained in this work. In particular, US real GDP is better described with two regimes, as is usually assumed in the estimation of such state space systems. However, EU business cycle exhibits strong non-linearities and more regimes are necessary. This is taken into account when performing estimation and regime identification.

References

[1] Anas, J., Billio, M., Ferrara, L., LoDuca, M. (2007) *Business Cycle Analysis with Multivariate Markov Switching Models*, in Growth and Cycle in the Euro Zone, Mazzi GL and Savio G (eds), Palgrave Macmillan: New York, 249–260.

[2] Anas, J., Billio, M., Ferrara, L., LoDuca, M. (2008) *A system for dating and detecting turning points in the Euro area*, The Manchester School, **76**, 549–577.

[3] Billio, M., and Casarin, R. (2010) *Identifying business cycle turning points with sequential monte carlo methods: an online and real-time application to Euro area*, Journal of Forecasting, **29**, 145–167.

[4] Billio, M., Casarin, R., Ravazzolo, F., and Van Dijk, H.K. (2012) *Combination scheme for turning point predictions*, The Quartely Review of Economics and Finance, **52**, 402–412.

[5] Billio, M., Ferrara, L., Gugan, D., and Mazzi, G.L. (2013) *Evaluation of regime switching models for real-time business cycle analysis of the Euro area*, Journal of Forecasting, DOI: 10.1002/for.2260, 2013.

[6] Burns, A.F. and Mitchell, W.C. (1946), *Measuring Business Cycles*, New York: National Bureau of Economic Research.

[7] Cavicchioli, M. (2013) *Determining the number of regimes in Markov-Switching VAR and VMA models*, forthcoming in Journal of Time Series Analysis.

[8] Cavicchioli, M. (2013) *Determining the number of regimes in Markov-Switching VAR and VMA models*, Working Papers from Department of Economics, University of Venice "Cá Foscari", N.03-2013.

[9] Choi, B. S. (1992) ARMA Model Identification, New York, Springer-Verlag.

[10] DeLong B., and Summers, L. (1988) *How does macroeconomic policy affect output?*, Brookings paper on economic activity **2**, 433–494.

[11] Francq, C. and Zakoïan, J.M. (2001) *Stationarity of Multivariate Markov-Switching ARMA Models*, Journal of Econometrics **102**, 339–364.

[12] Francq, C. and Zakoïan, J.M. (2002) *Autocovariance Structure of Powers of Switching-Regime ARMA Processes*, ESAIM: Probability and Statistics **6**, 259–270.

[13] Friedman, M. (1964) *Monetary Studies of the National Bureau* , The National Bureau Enters Its 45th Year, 44th Annual Report, 7-25; Reprinted in Milton Friedman (1969). The Optimum Quantity of Money and Other Essays. Chicago: Aldine, Chp. **12**, 261–284.

[14] Friedman, M. (1993) *The "plucking model" of business fluctuations Revisited* , Economic Inquiry **31**, 171–177.

[15] Hamilton, J. (1989) *A new approach to the economic analysis of nonstationary time series and the business cycle*, Econometrica **57**, 357–384.

[16] Kim, C.J. (1994) *Dynamic linear models with Markov-switching*, Journal of Econometrics **60**, 1–22.

[17] Kim, C.J. and Nelson, C.R. (1999) *Friedman's plucking model of business fluctuations: tests and estimates of permanent and transitory components*, Journal of Money, Credit, and Banking **31** (3), 317–334.

[18] Kim, C.J. and Nelson, C.R. (1999) *State-Space models with regime switching: classical and Gibbs-sampling approaches with applications*, The MIT Press.

[19] Krolzig, H.M. (1997) *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference and Application to Business Cycle Analysis*, Springer Verlag, Berlin-Heidelberg-New York.

[20] Krolzig, H.M. (2000) *Predicting Markov Switching vector autoregressive processes*, Nuffield College Economics Working Papers, 2000-WP31.

[21] Krolzig, H.M. (2001) *Markov Switching procedures for dating the Euro-zone business cycle*, Quarterly Journal of Economic Research, **3**, 339–351.

[22] Krolzig, H.M. (2004) *Constructing turning point chronologies with Markov-switching vector autoregressive models: the euro-zone business cycle*, in Mazzi G.L. and Savio G. (eds), Monographs of Official Statistics, Luxembourg: European Commission, 147–190.

[23] Lam, P.S. (1990) *The Hamilton model with a general autoregressive component: estimation and comparison with other models of economic time series*, Journal of Monetary Economics **26**, 409–432.

[24] Sichel, D.E. (1994) *Inventories and the three phases of business cycle*, Journal of Business and Economic Statistics, **12** (3), 269–277.

2.2.8 Appendix

Proof of Theorem 3.1. Using the measurement equation in (3.2) , we can easily compute

$$E(\mathbf{y}_t) = \sum_{j=0}^{r} \mathbf{\Lambda}_j E(\boldsymbol{\xi}_{t-j}) = (\sum_{j=0}^{r} \mathbf{\Lambda}_j)\boldsymbol{\pi}$$

and

$$E(\mathbf{y}_t)E(\mathbf{y}_{t+h}^{'}) = (\sum_{j=0}^{r} \mathbf{\Lambda}_j)\boldsymbol{\pi}\boldsymbol{\pi}^{'}(\sum_{j=0}^{r} \mathbf{\Lambda}_j^{'}) = (\sum_{j=0}^{r} \mathbf{\Lambda}_j)\mathbf{D}\mathbf{P}_{\infty}(\sum_{j=0}^{r} \mathbf{\Lambda}_j^{'})$$

where $\mathbf{D} = diag(\pi_1, \ldots, \pi_M)$. For every $h \geq r > 0$, we get

$$\begin{aligned} E(\mathbf{y}_t\mathbf{y}_{t+h}^{'}) &= E[\sum_{j=0}^{r} \mathbf{\Lambda}_j \boldsymbol{\xi}_{t-j} + \mathbf{\Sigma}(\boldsymbol{\xi}_t \otimes \mathbf{I}_K)\mathbf{u}_t, \sum_{i=0}^{r} \boldsymbol{\xi}_{t+h-i}^{'}\mathbf{\Lambda}_i^{'} + \mathbf{u}_{t+h}^{'}(\boldsymbol{\xi}_{t+h}^{'} \otimes \mathbf{I}_K)\mathbf{\Sigma}^{'}] \\ &= \sum_{i=0}^{r}\sum_{j=0}^{r} \mathbf{\Lambda}_j E(\boldsymbol{\xi}_{t-j}\boldsymbol{\xi}_{t+h-i}^{'})\mathbf{\Lambda}_i^{'} \\ &= \sum_{i=0}^{r}\sum_{j=0}^{r} \mathbf{\Lambda}_j E(\boldsymbol{\xi}_{t-j}\boldsymbol{\xi}_{t-j+h-i+j}^{'})\mathbf{\Lambda}_i^{'} \end{aligned}$$

$$\begin{aligned} &= \sum_{i=0}^{r}\sum_{j=0}^{r} \mathbf{\Lambda}_j \mathbf{D}\mathbf{P}^{h-i+j}\mathbf{\Lambda}_i^{'} \\ &= \sum_{i=0}^{r}\sum_{j=0}^{r} \mathbf{\Lambda}_j \mathbf{D}\mathbf{Q}^{h-i+j}\mathbf{\Lambda}_i^{'} + \sum_{i=0}^{r}\sum_{j=0}^{r} \mathbf{\Lambda}_j \mathbf{D}\mathbf{P}_{\infty}\mathbf{\Lambda}_i^{'} \end{aligned}$$

because $h+r \geq h-i+j \geq h-r \geq 0$ for every $i, j = 0, \ldots, r$. Here we have used the well-known property $E(\boldsymbol{\xi}_t\boldsymbol{\xi}_{t+h}^{'}) = \mathbf{D}\mathbf{P}^h$ for every $h \geq 0$ (where we set $\mathbf{P}^h = \mathbf{I}_M$ for $h = 0$). Thus

$$\mathbf{\Gamma}_{\mathbf{y}}(-h) = cov(\mathbf{y}_t, \mathbf{y}_{t+h}) = E(\mathbf{y}_t\mathbf{y}_{t+h}^{'}) - E(\mathbf{y}_t)E(\mathbf{y}_{t+h}^{'}) = (\sum_{j=0}^{r} \mathbf{\Lambda}_j\mathbf{D}\mathbf{P}^j)\mathbf{Q}^h(\sum_{i=0}^{r} \mathbf{P}^{-i}\mathbf{\Lambda}_i^{'})$$

for every $h \geq r > 0$, and taking the transpose gives the result. Here we have used the relations $\mathbf{Q}^{h-i+j} = \mathbf{P}^{h-i+j} - \mathbf{P}_{\infty} = \mathbf{P}^j(\mathbf{P}^h - \mathbf{P}_{\infty})\mathbf{P}^{-i} = \mathbf{P}^j\mathbf{Q}^h\mathbf{P}^{-i}$ as $\mathbf{P}^n\mathbf{P}_{\infty} = \mathbf{P}_{\infty}\mathbf{P}^n = \mathbf{P}_{\infty}$ and $\mathbf{Q}^n = \mathbf{P}^n - \mathbf{P}_{\infty}$ for every $n \geq 1$. □

Proof of Theorem 3.2. For every $h \geq r > 0$, we get

$$\mathbf{\Gamma}_{\mathbf{y}}(-h) = E(\boldsymbol{y}_t\boldsymbol{y}_{t+h}^{'}) - (\sum_{j=0}^{r} \mathbf{\Lambda}_j)\mathbf{D}\mathbf{P}_{\infty}(\sum_{i=0}^{r} \mathbf{\Lambda}_i^{'})$$

and

$$E(\boldsymbol{y}_t\boldsymbol{y}_{t+h}^{'}) = (\sum_{j=0}^{r}\boldsymbol{\Lambda}_j)\mathbf{D}\mathbf{P}_\infty(\sum_{j=0}^{r}\boldsymbol{\Lambda}_j^{'}) + \sum_{j=0}^{r}\sum_{i=0}^{r}\widetilde{\boldsymbol{\Lambda}}_j E(\boldsymbol{\delta}_{t-j}\boldsymbol{\delta}_{t-j+h-i+j}^{'})\widetilde{\boldsymbol{\Lambda}}_i^{'},$$

by using the State-Space representation in (3.3). Then it follows that

$$\boldsymbol{\Gamma}_{\mathbf{y}}(-h) = \sum_{j=0}^{r}\sum_{i=0}^{r}\widetilde{\boldsymbol{\Lambda}}_j\widetilde{\mathbf{D}}(\mathbf{F}^{'})^{h-i+j}\widetilde{\boldsymbol{\Lambda}}_i^{'} = (\sum_{j=0}^{r}\widetilde{\boldsymbol{\Lambda}}_j\widetilde{\mathbf{D}}(\mathbf{F}^{'})^j)(\mathbf{F}^{'})^h(\sum_{i=0}^{r}(\mathbf{F}^{'})^{-i}\widetilde{\boldsymbol{\Lambda}}_i^{'})$$

hence

$$\boldsymbol{\Gamma}_{\mathbf{y}}(h) = \mathbf{A}^{'}\mathbf{F}^h\mathbf{B}$$

where $\mathbf{A}^{'} = \sum_{i=0}^{r}\widetilde{\boldsymbol{\Lambda}}_i\mathbf{F}^{-i}$ and $\mathbf{B} = \sum_{j=0}^{r}\mathbf{F}^j\widetilde{\mathbf{D}}\widetilde{\boldsymbol{\Lambda}}_j^{'}$, which we assume to be nonzero matrices. Now apply Theorem 2.2 from Cavicchioli (2013) with $p = 0$, $q = r > 0$ and $M-1$ instead of M as $\mathbf{F}$ is $(M-1)\times(M-1)$. □

Proof of Theorem 3.3. The stable VAR(1) process $(\boldsymbol{\delta}_t)$ possesses the vector MA(∞) representation $\boldsymbol{\delta}_t = F(L)^{-1}\mathbf{w}_t$. Since the inverse matrix polynomial can be reduced to the inverse of the determinant, that is, $|F(L)|^{-1}$, and the adjoint matrix $F(L)^*$, we have $\boldsymbol{\delta}_t = |F(L)|^{-1}F(L)^*\mathbf{w}_t$. Inserting this transformed state equation into the measurement equation in (3.3) and multiplying by the determinant of $F(L)$ yield

$$|F(L)|(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = \sum_{j=0}^{r}\widetilde{\boldsymbol{\Lambda}}_j F(L)^* L^j\mathbf{w}_t + \widetilde{\boldsymbol{\Sigma}}(F(L)^*\otimes\mathbf{I}_K)(\mathbf{w}_t\otimes\mathbf{I}_K)\mathbf{u}_t + \ |F(L)|\boldsymbol{\Sigma}(\boldsymbol{\pi}\otimes\mathbf{I}_K)\mathbf{u}_t$$

which is a stable VARMA whose autoregressive lag polynomial is scalar, and where the orders of the stable VARMA are as in the statement. □

Proof of Theorem 4.1. Set $\mathbf{x}_t = \sum_{j=0}^{r}\boldsymbol{\Lambda}_j\boldsymbol{\xi}_{t-j} + \boldsymbol{\Sigma}(\boldsymbol{\xi}_t\otimes\mathbf{I}_K)\mathbf{u}_t$. For every $h \geq r > 0$, we have

$$(A.1)\qquad \begin{aligned} cov(\mathbf{x}_{t+h},\mathbf{y}_t) &= cov(\boldsymbol{\phi}(L)(\boldsymbol{\xi}_{t+h}\otimes\mathbf{I}_K)\mathbf{y}_{t+h},\mathbf{y}_t) \\ &= \boldsymbol{\phi}(L)[E(\boldsymbol{\xi}_{t+h})\otimes cov(\mathbf{y}_{t+h},\mathbf{y}_t)] \\ &= \boldsymbol{\phi}(L)(\boldsymbol{\pi}\otimes\mathbf{I}_K)[1\otimes cov(\mathbf{y}_{t+h},\mathbf{y}_t)] \\ &= B(L)\boldsymbol{\Gamma}_{\mathbf{y}}(h) \end{aligned}$$

where $B(L) = \boldsymbol{\phi}(L)(\boldsymbol{\pi}\otimes\mathbf{I}_K)$ is a $K\times K$ matrix lag polynomial of degree p. Since the process is second-order stationary, the above formula implies that $cov(\mathbf{x}_{t+h}, \mathbf{y}_t)$ is time invariant. Using the unrestricted State-Space representation (4.3) and the relation $\boldsymbol{\delta}_{t+h} = \mathbf{F}^h \boldsymbol{\delta}_t + \sum_{j=0}^{h-1} \mathbf{F}^j \mathbf{w}_{t+h-j}$, we have

$$
(A.2) \quad
\begin{aligned}
\mathbf{x}_{t+h} =& (\sum_{j=0}^{r} \boldsymbol{\Lambda}_j)\boldsymbol{\pi} + \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j \mathbf{F}^h \boldsymbol{\delta}_{t-j} + \sum_{i=0}^{h-1}\sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j \mathbf{F}^i \mathbf{w}_{t+h-i-j} + \widetilde{\boldsymbol{\Sigma}}[(\mathbf{F}^h \boldsymbol{\delta}_t) \otimes \mathbf{I}_K]\mathbf{u}_{t+h} \\
&+ \sum_{i=0}^{h-1} \widetilde{\boldsymbol{\Sigma}}[(\mathbf{F}^i \mathbf{w}_{t+h-i}) \otimes \mathbf{I}_K]\mathbf{u}_{t+h} + \boldsymbol{\Sigma}(\boldsymbol{\pi} \otimes \mathbf{I}_K)\mathbf{u}_{t+h}
\end{aligned}
$$

By (A.2), for every $h \geq r > 0$, we obtain

$$
(3.6) \quad
\begin{aligned}
cov(\mathbf{x}_{t+h}, \mathbf{y}_t) &= cov((\sum_{j=0}^{r} \boldsymbol{\Lambda}_j)\boldsymbol{\pi} + \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j \mathbf{F}^h \boldsymbol{\delta}_{t-j}, \mathbf{y}_t) = \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j \mathbf{F}^h cov(\boldsymbol{\delta}_{t-j}, \mathbf{y}_t) \\
&= \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j \mathbf{F}^{h-j} E(\boldsymbol{\delta}_t \mathbf{y}_t^{'}) = \mathbf{A}'\mathbf{F}^h \mathbf{B}
\end{aligned}
$$

where $\mathbf{A}^{'} = \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j \mathbf{F}^{-j}$ and $\mathbf{B} = E(\boldsymbol{\delta}_t \boldsymbol{y}_t^{'})$, as requested. Now we see that $E(\boldsymbol{\delta}_t \mathbf{y}_t^{'})$ is time invariant as $cov(\mathbf{x}_{t+h}, \mathbf{y}_t)$ is. Collecting formulae (A.1) and (A.3) gives the result. □

Proof of Theorem 4.3. We have $\boldsymbol{\delta}_t = F(L)^{-1}\mathbf{w}_t$ as usual. Substituting the last formula in the State-Space representation of the initial process obtained in the same manner as in (4.3), we get

$$
(A.4) \quad
\begin{aligned}
\mathbf{A}(L)(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) =& \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j F(L)^{-1} L^j \mathbf{w}_t + \sum_{i=0}^{q} \widetilde{\boldsymbol{\Theta}}_i (F(L)^{-1} \otimes \mathbf{I}_K)(\mathbf{w}_t \otimes \mathbf{I}_K) L^i \mathbf{u}_t \\
&+ \sum_{i=0}^{q} \boldsymbol{\Theta}_i (\boldsymbol{\pi} \otimes \mathbf{I}_K) L^i \mathbf{u}_t
\end{aligned}
$$

where $\widetilde{\boldsymbol{\Theta}}_i$ is given by the usual construction applied to $\boldsymbol{\Theta}_i = (\boldsymbol{\Theta}_{1i} \ldots \boldsymbol{\Theta}_{Mi})$ for $i = 0, \ldots, q$. Premultiplying (A.4) by $|F(L)|$ yields

$$
(A.5) \quad
\begin{aligned}
|F(L)|\mathbf{A}(L)(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) =& \sum_{j=0}^{r} \widetilde{\boldsymbol{\Lambda}}_j F(L)^* L^j \mathbf{w}_t + \sum_{i=0}^{q} \widetilde{\boldsymbol{\Theta}}_i (F(L)^* \otimes \mathbf{I}_K)(\mathbf{w}_t \otimes \mathbf{I}_K) L^i \mathbf{u}_t \\
&+ |F(L)| \sum_{i=0}^{q} \boldsymbol{\Theta}_i (\boldsymbol{\pi} \otimes \mathbf{I}_K) L^i \mathbf{u}_t.
\end{aligned}
$$

Now the regularity conditions of the statement mean that $\mathbf{A}(L)$ is invertible, that is, $\mathbf{A}(L)^*\mathbf{A}(L) = |\mathbf{A}(L)|\mathbf{I}_K$. Premultiplying (A.5) by $\mathbf{A}(L)^*$, we get the VARMA(p^*, q^*) representation, with $p^* \leq M + Kp - 1$ and $q^* \leq M + (K-1)p + \max\{r, q+1\} - 2$ (use the fact that the degree of $|\mathbf{A}(L)|$ is Kp):

$$|F(L)||\mathbf{A}(L)|(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y}}) = \sum_{j=0}^{r} \mathbf{A}(L)^* \widetilde{\boldsymbol{\Lambda}}_j F(L)^* L^j \mathbf{w}_t$$

$$+ \sum_{i=0}^{q} \mathbf{A}(L)^* \widetilde{\boldsymbol{\Theta}}_i (F(L)^* \otimes \mathbf{I}_K)(\mathbf{w}_t \otimes \mathbf{I}_K) L^i \mathbf{u}_t + |F(L)|\mathbf{A}(L)^* \sum_{i=0}^{q} \boldsymbol{\Theta}_i (\boldsymbol{\pi} \otimes \mathbf{I}_K) L^i \mathbf{u}_t$$

which is a stable model as required in the statement. □

Chapter 3

Markov Switching Models for Volatility: Filtering, Approximation and Duality

Abstract. *This paper is devoted to show duality in the estimation of Markov Switching (MS) processes for volatility. It is well-known that MS-GARCH models suffer of path dependence which makes the estimation step unfeasible with usual Maximum Likelihood procedure. However, by rewriting the MS-GARCH model in a suitable linear State Space representation, we are able to give a unique framework to reconcile the estimation obtained by the Kalman Filter and with some auxiliary models proposed in the literature. Reasoning in the same way, we present a linear Filter for MS-Stochastic Volatility (MS-SV) models on which different conditioning sets yield more flexibility in the estimation. Estimation on simulated data and on short-term interest rates shows the feasibility of the proposed approach.[JEL Classification: C01, C13, C58]*

Keywords: Markov Switching, MS-GARCH model, MS-SV model, estimation, auxiliary model, Kalman Filter.

3.1 Introduction

Time varying volatility is one of the main property of economic time series, common especially to many financial time series. Moreover, describing and, where possible, forecasting volatility is a key aspect in financial economics and econometrics. It is not only a statistical exercise but it has also important impacts in terms of asset allocation, asset pricing as well as value-at-risk computation and thus for risk management. A lot of work has been done on two popular classes of models which describe time-varying volatility: Generalized Autoregressive Conditional Heteroschedasticity (GARCH)-type models and Stochastic Volatility (SV)-type models. GARCH models (Bollerslev (1986), Nelson (1990), Lamoureux and Lastrapes (1990)) are commonly known as observation-driven models (see Shephard (1996)). In fact, they describe the variance as a linear function of the squares of past observations and then one type of shock alone drives both the series itself and its volatility. On the contrary, SV models (Taylor (1986), Harvey, Ruiz and Shephard (1994)) belong to the class of parameter-driven models since these models are driven by two type of shocks, one of which influences the volatility. The presence of unobserved or latent components makes SV models harder to estimate and to handle statistically, while GARCH parameters can easily be estimated using maximum likelihood procedure. In the latter models, one potential source of misspecification is that the structural form of conditional means and variances is relatively inflexible and it is held fixed throughout the sample period. In this sense, they are called *single-regime* models since a single structure is assumed for the conditional mean and variance.

In order to allow more flexibility, the assumption of a single regime could be relaxed in favour of a *regime-switching* model. The coefficients of this model are different in each regime to account for the possibility that the economic mechanism that generates the financial serie undergoes a finite number of changes over the sample period. These coefficients are unknown and must be estimated, and, althought the regimes are never observed, probabilistic statements can be made about the relative likelihood of their occurrence, conditional on an information set.

A well-known problem to face when dealing with the estimation of Markov Switching GARCH models is the path dependence. Cai (1994) and Hamilton and Susmel (1994) have argued that MS-GARCH models are essentially intractable and impossible to estimate due to the dependence of conditional variance on the entire path history of the data. That is, the distribution at time t, conditional on the current state and on available information, is directly dependent of the current state but also indirectly dependent on all past states due to the path dependence inherent in MS-GARCH models. This is because the conditional variance at time t depends upon the conditional variance at time $t-1$, which depends upon the regime at time $t-1$ and on the conditional variance at time $t-2$, and so on. Hence, the conditional variance at time t depends on the entire sequence of regimes up to time t.

In the first part of this paper, we will consider the univariate version of MS-GARCH and some methods proposed to bypass the problem of path dependence. The trick is mainly found in adopting different specifications of the original MS-GARCH model. Some authors propose Quasi Maximum Likelihood (QML) procedures of a model which allow similar effects of the original one. Models which elude in this way the path dependence problem are proposed by Gray (1996), Dueker (1997) and Klaassen (2002), among others. Gray (1996) proposes a model in which path dependence is removed by aggregating the conditional variances from the regimes at each step. This aggregated conditional variance (conditional on available information, but aggregated over the regimes) is then all that is required to compute the conditional variance at the next step. The same starting idea is used in Dueker (1997), with a slightly different approach. He extends the information set including also current information on the considered series. Furthermore, Klaassen (2002) puts further this idea. Particularly, when integrating out the unobserved regimes, he uses all available information, whereas Gray uses only part of it. Another method to deal with MS-GARCH models has been proposed by Haas, Mittnik and Paolella (2004) for which the variance is disaggregated in independent processes; this is a simple generalization of the GARCH process to a multi-regime setting. Finally, Bayesian approach based on Markov Chain Monte Carlo (MCMC) Gibbs technique for estimating MS-GARCH can be found in Bauwens, Preminger and Rombouts (2010) and Bauwens, Dufays and Rombouts (2011), Henneke, Rachev, Fabozzi and Metodi (2011) or Billio, Casarin and Osuntuyi (2012). Other approaches based on both Monte Carlo methods combined with expectation-maximization algorithm and importance sampling to evaluate ML estimators can be found in Augustyniak (2013) and Billio, Monfort and Robert (1998a and 1998b).

In the second part of the paper, we will consider the extension of univariate SV model with regime-switching features. If SV models are difficult to estimate due to the latent variable, MS-SV are even more complicated because there are two hidden levels in the latent structure. So, MS-SV models have been studied and estimated mainly with Bayesian techniques. For example, So, Lam and Li (1998) adopt MCMC method and they construct Bayesian estimators by Gibbs sampling. Another Bayesian approach is sequential simulation based filtering (Particle Filter). See, for instance, Casarin (2004) and Carvalho and Lopes (2007).

The main contribution of the present paper is to give a unique framework to reconcile the estimation obtained by the above auxiliary models from one side, and Kim's (1994) filtering algorithm for Markov switching state space from the other. Kim's algorithm can be used, under some regularity conditions, to obtain inferences about any dynamic time series model with Markov switching that can be put in a state space form. It is a very flexible approach and allows the estimation of a broad class of models. However, to make the filter operable, at each iteration it collapses M^2 posteriors (where M is the number of states) in M of it, employing an approximation. Finally, Quasi Maximum Likelihood estimation of the model recovers the unknown parameters. Then our first contribution is to show duality in the estimation of Markov Switching processes for volatility. In particular, having a suitable linear state space representation for the MS-GARCH model, we are able to prove the equivalence in the estimation obtained by Kim's Filter and through auxiliary models proposed in the literature. The second contribution relates instead to MS-SV models. In fact, we are able to extend the approach previously used for MS-GARCH to MS-SV models. In particular, we parallel the model with the gaussian state space model and we propose a linear Filter on which different conditioning information sets yield more flexibility in the estimation. Numerical and empirical applications show the feasibility of these approaches.

The paper is structured as follows. In Section 2 we specify the MS-GARCH model of interest and introduce some concepts and notations. Section 3 reviews the main auxiliary models for MS-GARCH which are proposed in the literature to overpass the path dependence problem. In Section 4 we present a linear state space representation associated to the MS-GARCH and determine the algorithm for the linear filter. This serves to prove our duality results discussed in Section 5. In Section 6 we write a linear approximated filter for MS-SV models. In Section 7 we compare estimation of the parameters using different approximations in the proposed filter for simulated data and short-term interest rates. Section 8 concludes. Finally, Appendix A describes in details some Formulae and in Appendix B we recall the main results about the stationarity of Markov Switching models and particulary applied to our specifications.

3.2 Markov Switching GARCH

Let ϵ_t be the observed univariate[1] time series variable (as for instance, returns on a financial asset) centered on its mean and let s_t be a discrete, unobserved state variable with M-states. The Markov Switching GARCH(1,1) model is defined as

$$\begin{cases} \epsilon_t = \sigma_t(\Psi_{t-1}, \theta(s_t))u_t \\ \sigma_t^2(\Psi_{t-1}, s_t) = \omega(s_t) + \alpha(s_t)\epsilon_{t-1}^2 + \beta(s_t)\sigma_{t-1}^2(\Psi_{t-2}, s_{t-1}) \end{cases} \tag{1}$$

where $u_t \sim IID(0,1)$, $\omega(s_t) > 0$, $\alpha(s_t), \beta(s_t) \geq 0$ and $\theta(s_t)$ is the parameter vector defined as $\theta(s_t) = (\omega(s_t), \alpha(s_t), \beta(s_t))'$. Here $\Psi_{t-1} = \{\epsilon_{t-1}, \ldots, \epsilon_1\}$ denotes the information set of observations available up to time $t-1$. Moreover, s_t is a M-state first order Markov chain with *transition probabilities*, which are assumed time invariant[2],

$$\pi_{ij,t} = p(s_t = j | s_{t-1} = i)$$

where

$$\sum_{j=1}^{M} \pi_{ij,t} = 1$$

for every $i = 1, \ldots, M$.

Let us introduce the following concepts and notations:

- $p(s_t = j|\Psi_{t-1}) = p_{j,t|t-1}$ which is the *prediction probability*;

[1] The proposed setting can be easily extended to a multivariate framework. This can be done on the line of multivariate GARCH models to regime-switching framework proposed by Billio and Caporin (2005) and Pellittier (2006). However, note that multivariate volatility models in the context of single regime switching are the Constant Conditional Correlation (CCC) model of Bollerslev (1990) and the Dynamic Conditional Correlation (DCC) model of Engle (2002).

[2] If the information variables that govern time-variation in the transition probabilities is conditionally uncorrelated with the state of the Markov process, which holds in general, Hamilton's (1989) filtering method is still valid also with time-varying transition probabilities.

- $p(s_t = j|\Psi_t) = p_{j,t|t}$ which is the *filtered probability.*

From these we can compute the *augmented filtered probability* as

$$p(s_{t-1} = i|s_t = j, \Psi_{t-1}) = \frac{\pi_{ij,t}\ p_{i,t-1|t-1}}{p_{j,t|t-1}} = p_{ij,t-1|t,t-1}.$$

Note that the filtering algorithm computes $p_{t|t-1,t} = p(s_t|s_{t-1}, \Psi_t)$ in terms of $p_{t|t-1,t-1}$ and the conditional density of ϵ_t which depends on the current regime s_t and all past regimes, i.e, $f(\epsilon_t|s_1, \ldots, s_t, \Psi_{t-1})$. Computation details are shown in Appendix A1.

3.3 Auxiliary Models for MS-GARCH

As argued in the Introduction, the main problem to face when dealing with the estimation of Markov Switching GARCH model is the path dependence, which is the dependence of the conditional variance on the entire sequence of regimes. The common approach to eliminate path dependence is to replace the lagged conditional variance derived from the original MS-GARCH model with a proxy. Various authors have proposed different auxiliary models which differ only by the content of the information used to define such a proxy. In general, different auxiliary models can be obtained by approximating the conditional variance of the MS-GARCH process

$$\sigma_t^2(\Psi_{t-1}, s_t) = \omega(s_t) + \alpha(s_t)\,{}^{(SP)}\epsilon_{t-1}^2 + \beta(s_t)\,{}^{(SP)}\sigma_{t-1}^2. \tag{2}$$

In the literature there are different specifications (in short, SP) of ${}^{(SP)}\epsilon_{t-1}^2$ and ${}^{(SP)}\sigma_{t-1}^2$ which in turn define different approximations of the original process. In this Section we give a detailed description of four auxiliary models presented in the literature, specifying the superscript in (2) with the initial letter of the author who proposed that specification.

3a. Gray's Model

The first attempt to eliminate the path dependence is proposed by Gray (1996). He approximates the original model by replacing the lagged conditional variance σ_{t-1}^2 with a proxy ${}^{(G)}\sigma_{t-1}^2$

as follows:

$$
\begin{aligned}
{}^{(G)}\sigma^2_{t-1} &= E[\sigma^2_{t-1}(\Psi_{t-2}, s_{t-1})|\Psi_{t-2}] \\
&= \sum_{i=1}^{M} \sigma^2_{t-1}(\Psi_{t-2}, s_{t-1}=i)\; p(s_{t-1}=i|\Psi_{t-2}) \qquad (3)\\
&= \sum_{i=1}^{M} {}^{(G)}\sigma^2_{i,t-1|t-2}\; p_{i,t-1|t-2}
\end{aligned}
$$

where, according to the model, ${}^{(G)}\sigma^2_{t-1|t-2}$ turns out to be a function of Ψ_{t-2} and $s_{t-1} = i$. Note that the model originally proposed by Gray is not centered as in our case, but this can always be assumed without loss of generality.

3b. Dueker's Model

In the previous approximation, the information coming from ϵ_{t-1} is not used. Dueker (1997) proposes to change the conditioning scheme including ϵ_{t-1} while assuming that σ^2_{t-1} is a function of Ψ_{t-2} and s_{t-2}. Hence

$$
\begin{aligned}
{}^{(D)}\sigma^2_{t-1} &= E[\sigma^2_{t-1}(\Psi_{t-2}, s_{t-2})|\Psi_{t-1}] \\
&= \sum_{k=1}^{M} \sigma^2_{t-1}(\Psi_{t-2}, s_{t-2}=k)\; p(s_{t-2}=k|\Psi_{t-1}) \qquad (4)\\
&= \sum_{k=1}^{M} {}^{(D)}\sigma^2_{k,t-1|t-2}\; p_{k,t-2|t-1}
\end{aligned}
$$

so that ${}^{(D)}\sigma^2_{t-1|t-2}$ is a function of Ψ_{t-2} and $s_{t-2} = k$, and $p_{k,t-2|t-1}$ is one-period ahead smoothed probability which, shifting one period, can be computed as

$$
p_{i,t-1|t} = p(s_{t-1} = i|\Psi_t) = p_{i,t-1|t-1} \sum_{j=1}^{M} \frac{\pi_{ij,t}\; p_{j,t|t}}{p_{j,t|t-1}}.
$$

3c. Simplified Klaassen's Model

The approximation proposed by Klaassen (2002) is similar to that from Dueker (1997) but it assumes that σ^2_{t-1} is a function of Ψ_{t-2} and s_{t-1}. So it results computationally simpler. In fact, we have

$$
\begin{aligned}
{}^{(SK)}\sigma^2_{t-1} &= E[\sigma^2_{t-1}(\Psi_{t-2}, s_{t-1})|\Psi_{t-1}] \\
&= \sum_{i=1}^{M} \sigma^2_{t-1}(\Psi_{t-2}, s_{t-1}=i)\ p(s_{t-1}=i|\Psi_{t-1}) \\
&= \sum_{i=1}^{M} {}^{(SK)}\sigma^2_{i,t-1|t-2}\ p_{i,t-1|t-1}.
\end{aligned} \tag{5}
$$

Then from the considered model, ${}^{(SK)}\sigma^2_{t-1|t-2}$ results to be a function of Ψ_{t-2} and $s_{t-1}=i$.

3d. Klaassen's Model

Finally, Klaassen (2002) generalizes the previous auxiliary model including in the conditioning set the information coming also from the current regime s_t. So σ^2_{t-1} turns out to be approximated as

$$
\begin{aligned}
{}^{(K)}\sigma^2_{t-1} &= E[\sigma^2_{t-1}(\Psi_{t-2}, s_{t-1})|\Psi_{t-1}, s_t=j] \\
&= \sum_{i=1}^{M} \sigma^2_{t-1}(\Psi_{t-2}, s_{t-1}=i)\ p(s_{t-1}=i|\Psi_{t-1}, s_t=j) \\
&= \sum_{i=1}^{M} {}^{(K)}\sigma^2_{i,t-1|t-2}\ p_{ij,t-1|t,t-1}
\end{aligned} \tag{6}
$$

where $p_{ij,t-1|t,t-1}$ is the augmented filtered probability as defined in Section 2. Consequently, here ${}^{(K)}\sigma^2_{t-1|t-2}$ becomes a function of Ψ_{t-2} and $s_{t-1}=i$.

3.4 State Space Representation and Filtering

In order to develop a theory of linear filtering for MS-GARCH models, we need to associate to the model some linear state space representations. In this Section we propose a state space representation and write the associated Kalman Filter. For this purpose, we use notations from Kim (1994) and Kim and Nelson (1999) which study Markov switching state space models. They propose basic filtering and smoothing algorithms, along with maximum likelihood estimation, for a broad class of Markov switching models which can be written in state space form. This linear filter can be used, under some regularity conditions, to obtain approximate inferences. In fact, it introduces an approximation by collapsing information on the regimes story at each iteration. Such an approximation will be presented hereafter.

Consider the model as in (1). For every $s_t = j$ and $s_{t-1} = i$, let us define $\epsilon_t^2 = \sigma_{j,t}^2 + v_t$, where $\sigma_{j,t}^2 = \sigma_t^2(\Psi_{t-1}, s_t = j)$ and $v_t = \sigma_{j,t}^2(u_t^2 - 1)$. Then v_t is a white noise with zero mean and variance σ_{vj}^2 and $v_t \in [-\sigma_{j,t}^2, +\infty[$. Now we have

$$\begin{aligned}\epsilon_t^2 &= \sigma_{j,t}^2 + v_t \\ &= \omega_j + \alpha_j \epsilon_{t-1}^2 + \beta_j \sigma_{i,t-1}^2 + v_t \\ &= \omega_j + \alpha_j \epsilon_{t-1}^2 + \beta_j(\epsilon_{t-1}^2 - v_{t-1}) + v_t,\end{aligned}$$

where ω_j, α_j and β_j are the elements obtained by replacing s_t by j in ω_{s_t}, α_{s_t} and β_{s_t}, respectively.

So we can write the MS-ARMA(1,1) representation of the process in (1) as

$$(1 - \delta_j L)\epsilon_t^2 = \omega_j + (1 - \beta_j L)v_t \tag{7}$$

where $\delta_j = \alpha_j + \beta_j$ for $j = 1, \ldots, M$. See, for example, Gourieroux and Monfort (1997). For stationarity conditions concerning with such a process we refer to Appendix B.

Setting $B_t = \begin{pmatrix} \epsilon_{t-1}^2 \\ v_{t-1} \end{pmatrix}$, we get

$$\epsilon_t^2 = \omega_j + (\delta_j \quad -\beta_j) \begin{pmatrix} \epsilon_{t-1}^2 \\ v_{t-1} \end{pmatrix} + v_t = \omega_j + (\delta_j \quad -\beta_j) B_t + v_t$$

for every $j = 1, \ldots, M$. In order to simplify notations, let us define

$$y_t = \epsilon_t^2, \qquad H_{s_t} = (\delta_{s_t} \quad -\beta_{s_t}), \qquad F_{s_t} = \begin{pmatrix} \delta_{s_t} & -\beta_{s_t} \\ 0 & 0 \end{pmatrix}, \qquad G = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \mu_{s_t} = \begin{pmatrix} \omega_{s_t} \\ 0 \end{pmatrix}.$$

Then, for every s_t, we obtain the following state space representation [1]:

$$\text{(8)} \qquad \begin{cases} y_t = \omega_{s_t} + H_{s_t} B_t + v_t \\ B_t = \mu_{s_t} + F_{s_t} B_{t-1} + G v_{t-1} \end{cases}$$

Conditional on $s_{t-1} = i$ and $s_t = j$, the Kalman Filter is:

Prediction

- $B^{(i,j)}_{t|t-1} = \mu_j + F_j B^i_{t-1|t-1}$
- $P^{(i,j)}_{t|t-1} = F_j P^i_{t-1|t-1} F'_j + GG' \sigma^2_{vj}$
- $\eta^{(i,j)}_{t|t-1} = y_t - y^{(i,j)}_{t|t-1} = y_t - H_j B^{(i,j)}_{t|t-1} - \omega_j$
- $f^{(i,j)}_{t|t-1} = H_j P^{(i,j)}_{t|t-1} H'_j + \sigma^2_{vj}$

Updating

- $B^{(i,j)}_{t|t} = B^{(i,j)}_{t|t-1} + K^{(i,j)}_t \eta^{(i,j)}_{t|t-1}$
- $P^{(i,j)}_{t|t} = P^{(i,j)}_{t|t-1} - K^{(i,j)}_t H_j P^{(i,j)}_{t|t-1}$

where $K^{(i,j)}_t = P^{(i,j)}_{t|t-1} H'_j [f^{(i,j)}_{t|t-1}]^{-1}$ is the Kalman gain

Initial Conditions

- $B^j_{0|0} = (I_2 - F_j)^{-1} \mu_j = \begin{pmatrix} (1-\delta_j)^{-1}\omega_j \\ 0 \end{pmatrix}$
- $\text{vec}(P^j_{0|0}) = \sigma^2_{vj} (I_4 - F_j \otimes F_j)^{-1} \text{vec}(GG') = \sigma^2_{vj} \begin{pmatrix} (1-\delta_j^2)^{-1}(1 - 2\delta_j\beta_j + \beta_j^2) \\ 1 \\ 1 \\ 1 \end{pmatrix}$

[1] Note that other state space representations can be associated to the model in (1). For instance, following the line of Kim and Nelson (1999), Example 2, Chapter 3. Our choice tends to be the less restrictive in term of stationarity conditions, hence more general.

- $p(s_0 = i) = \pi_i$ (steady-state probability) .

So $Y_{t-1} = \{y_{t-1}, \ldots, y_1\}$ is the information set up to time $t-1$, $B^i_{t-1|t-1} = E(B_t|Y_{t-1}, s_{t-1} = i)$ is an inference on B_t based on Y_{t-1} given $s_{t-1} = i$; $B^{(i,j)}_{t|t-1} = E(B_t|Y_{t-1}, s_t = j, s_{t-1} = i)$ is an inference on B_t based on Y_{t-1}, given $s_t = j$ and $s_{t-1} = i$; $P^i_{t-1|t-1}$ is the mean squared error matrix of $B^i_{t-1|t-1}$ conditional on $s_{t-1} = i$; $P^{(i,j)}_{t|t-1}$ is the mean squared error matrix of $B^{(i,j)}_{t|t-1}$ conditional on $s_t = j$ and $s_{t-1} = i$; $\eta^{(i,j)}_{t|t-1}$ is the conditional forecast error of y_t based on information up to time $t-1$, given $s_t = j$ and $s_{t-1} = i$; and $f^{(i,j)}_{t|t-1}$ is the conditional variance of forecast error $\eta^{(i,j)}_{t|t-1}$. Each iteration of the Kalman Filter produces an M-fold increase in the number of cases to consider. It is necessary to introduce some approximations to make the filter operable. The key is to collapse the $(M \times M)$ posteriors $B^{(i,j)}_{t|t}$ and $P^{(i,j)}_{t|t}$ into M posteriors $B^j_{t|t}$ and $P^j_{t|t}$. Hence, we consider the approximation proposed by Kim and Nelson (1999) and Kim (1994) applied to this state space representation (explicit computations are in Appendix A2) . Let $B^j_{t|t}$ be the expectation based not only on Y_t but also conditional on the random variable s_t taking on the value j. Then

$$B^j_{t|t} = \sum_{i=1}^{M} p_{ij,t-1|t,t}\ B^{(i,j)}_{t|t}. \tag{9}$$

3.5 Duality Results

Having such a convenient switching state space form associated to the initial MS-GARCH, gives us the possibility to reconcile in an unique framework the estimation through linear filter as described in Section 4 or via auxiliary models presented in Section 3. Duality exists when modifying the approximation described in (9) with different conditioning sets. From the measurement equation in (8) and using (9), we get

$$\begin{aligned} y^j_{t|t} &= E(y_t|s_t = j, Y_t) = \omega_j + H_j\ B^j_{t|t} \\ &= \omega_j + H_j\ \sum_{i=1}^{M} p_{ij,t-1|t,t}\ B^{(i,j)}_{t|t} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{M} p_{ij,t-1|t,t} \ (\omega_j + H_j \ B_{t|t}^{(i,j)}) \\
&= \sum_{i=1}^{M} p_{ij,t-1|t,t} \ y_{t|t}^{(i,j)}
\end{aligned}
$$

as $\sum_{i=1}^{M} p_{ij,t-1|t,t} = 1$. Here the expectation operator is meant in the sense of Kim and Nelson's book (1999). In the same way we can obtain

$$
y_{t|t-1}^{j} = E(y_t|Y_{t-1}, s_t = j) = \sum_{i=1}^{M} p_{ij,t-1|t,t-1} \ y_{t|t-1}^{(i,j)}
$$

and

$$
\begin{aligned}
y_{t-1|t-1}^{j} &= E(y_{t-1}|Y_{t-1}, s_t = j) = E(\sigma_{t-1}^2|Y_{t-1}, s_t = j) \\
&= \sigma_{j,t-1|t-1}^2 = \sum_{i=1}^{M} p_{ij,t-1|t,t-1} \ \sigma_{ij,t-1|t-2}^2.
\end{aligned}
$$

In particular, if the conditional variance is not a function of $s_t = j$, we get

$$
\begin{aligned}
y_{t-1|t-1} &= E(\epsilon_{t-1}^2|Y_{t-1}) = E(\sigma_{t-1}^2|Y_{t-1}) \\
&= \sigma_{t-1|t-1}^2 = \sum_{i=1}^{M} p_{ij,t-1|t,t-1} \ \sigma_{i,t-1|t-2}^2
\end{aligned}
\tag{10}
$$

which coincides with ${}^{(K)}\sigma_{t-1}^2$ in Formula (6). Here ${}^{(K)}\sigma_{t-1}^2$ is only a function of $s_{t-1} = i$. Thus the approximation of the Kalman Filter is dual to the one used as auxiliary model from Klaassen (2002). This also means that if we change the conditioning scheme in (10), we obtain others auxiliary models. In fact, if we assume probabilities to be only function of $s_{t-1} = i$ and if still σ_{t-1}^2 is a function of s_{t-1} , we have the Simplified Klaassen's model (2002). This gives the expression in (5), in fact:

$$
{}^{(SK)}\sigma_{t-1}^2 = \sum_{i=1}^{M} {}^{(SK)}\sigma_{i,t-1|t-2}^2 \ p_{i,t-1|t-1} \, .
$$

Moreover, if we assume instead that σ_{t-1}^2 is a function of $s_{t-2} = k$ and also considering prediction probabilities of $s_{t-2} = k$, we get the auxiliary model proposed by Dueker (1997):

$$^{(D)}\sigma^2_{t-1} = \sum_{i=1}^{M} {}^{(D)}\sigma^2_{k,t-1|t-2}\ p_{k,t-2|t-1}$$

which is Formula (4). Finally, if we consider the conditioning set up to Y_{t-2} rather than Y_{t-1}, we obtain

$$^{(G)}\sigma^2_{t-1} = \sum_{i=1}^{M} {}^{(G)}\sigma^2_{i,t-1|t-2}\ p_{i,t-1|t-2}$$

which is Formula (3) and corresponds to Gray's model. Hence, if we slightly change the conditioning set, we can obtain different specifications of the auxiliary models moving from the state space form in (8). To conclude, this proves ambivalence in the estimation via Kalman Filter and via approximated models. In Section 7, we will show the feasibility of the filtering procedure through numerical and empirical applications.

3.6 Markov Switching Stochastic Volatility

When we consider Markov Switching Stochastic Volatility model and in general parameter-driven models, we are facing a double level of latency which makes estimation and statistical analysis harder. However, there are very good reason to investigate this kind of models, as for instance, easier properties or generalization to the multivariate case as well as continuous time counterpart. Then, we consider the following MS-SV model

$$\begin{cases} \epsilon_t = \exp\{\frac{1}{2}h_t\}u_t \\ h_t = \mu_{s_t} + \rho_{s_t}h_{t-1} + v_t \end{cases} \tag{11}$$

where $u_t \sim IIN(0,1)$ and $v_t \sim IIN(0,\sigma^2_{v_{s_t}})$. Here the error terms are assumed to be independent of one other. To discuss stationarity conditions of the process, we will later rewrite the model in MS-ARMA form and stationarity conditions are discussed at the end of Appendix B.

Following Harvey, Ruiz and Shephard (1994), we easily obtain a linear state space form. In fact, squaring (11) and taking logs, we have

$$\log \epsilon_t^2 = \alpha + h_t + e_t$$

where $\alpha = E(\log u_t^2) \cong -1.270$ and $e_t = \log u_t^2 - E(\log u_t^2)$. Thus $e_t \sim IIN(0, \frac{\pi^2}{2})$, where $\frac{\pi^2}{2} \cong 4.935$, and higher moments of (e_t) are known. Now, replacing $\log \epsilon_t^2$ with y_t, the MS-SV can be written as

$$\begin{cases} y_t = \alpha + h_t + e_t \\ h_t = \mu_{s_t} + \rho_{s_t} h_{t-1} + v_t \end{cases} \qquad (12)$$

Hence, it is natural to propose the Kalman filter for model (12) following the line of Kim and Nelson (1999, Chapter 5). In this case, conditional on $s_t = j$ and $s_{t-1} = i$, we get

Prediction

- $h_{t|t-1}^{(i,j)} = \mu_j + \rho_j h_{t-1|t-1}^i$
- $P_{t|t-1}^{(i,j)} = \rho_j^2 P_{t-1|t-1}^i + \sigma_{vj}^2$
- $\eta_{t|t-1}^{(i,j)} = y_t - y_{t|t-1}^{(i,j)} = y_t - h_{t|t-1}^{(i,j)} - \alpha$
- $f_{t|t-1}^{(i,j)} = P_{t|t-1}^{(i,j)} + \frac{\pi^2}{2}$

Updating

- $h_{t|t}^{(i,j)} = h_{t|t-1}^{(i,j)} + P_{t|t-1}^{(i,j)} [f_{t|t-1}^{(i,j)}]^{-1} \eta_{t|t-1}^{(i,j)}$
- $P_{t|t}^{(i,j)} = P_{t|t-1}^{(i,j)} - P_{t|t-1}^{(i,j)} [f_{t|t-1}^{(i,j)}]^{-1} P_{t|t-1}^{(i,j)}$

where $K_t^{(i,j)} = P_{t|t-1}^{(i,j)} [f_{t|t-1}^{(i,j)}]^{-1}$ is the Kalman gain.

Initial Conditions

- $h_{0|0}^j = \mu_j (1 - \rho_j)^{-1}$
- $P_{0|0}^j = \sigma_{vj}^2 (1 - \rho_j^2)^{-1}$

- $p(s_0 = i) = \pi_i$ (steady-state probability).

If we apply the approximation proposed by Kim and Nelson (1999) to this state space representation, we can write

$$h^j_{t|t} = \sum_{i=1}^{M} h^{(i,j)}_{t|t} \; p_{ij,t-1|t,t}.$$

At this point, different conditioning sets can be applied to the above approximation, mimic the same ideas used to obtain different auxiliary models in the MS-GARCH model. In the sequel we propose four approximations for the MS-SV in (12). Approximation 1 denotes Kim and Nelson's approximation as specified above. As done for the MS-GARCH, if we change the conditioning set we can obtain different and possibily more precise estimates. Approximation 2 changes the conditioning set on the volatility up to $t-1$:

$$h^j_{t|t} = \sum_{i=1}^{M} h^{(i,j)}_{t|t-1} \; p_{ij,t-1|t,t}.$$

Approximation 3 considers the information set up to Y_{t-1} only for the augmented filtered probabilities:

$$h^j_{t|t} = \sum_{i=1}^{M} h^{(i,j)}_{t|t} \; p_{ij,t-1|t,t-1}.$$

The last approximation 4 simultaneously has the features of 2 and 3, conditioning both volatility and probabilities at $t-1$:

$$h^j_{t|t} = \sum_{i=1}^{M} h^{(i,j)}_{t|t-1} \; p_{ij,t-1|t,t-1}.$$

In the next Section we will test these specifications in a simulated study in order to investigate differences in the implementation of the Filter.

3.7 Numerical and Empirical Applications

In this Section we apply the methods described above both to Monte Carlo experiment and real data. In particular, the aim of these applications is to show the feasibility of the proposed approaches via linear filtering for both Markov switching GARCH and SV models. Note that this method has the advantage of avoiding fine-tuning procedures implemented in most Bayesian estimation techniques. In fact, giving some initial conditions, the only duty of the researcher is to decide which approximation to adopt in the filtering procedure.

(I) Simulation study

In this Subsection, we draw some comparisons from a simulation study performed by So, Lam and Li (1998). In that paper, they simulate a Markov switching stochastic volatility model with three states and parameters described hereafter and estimate the model through MCMC procedure and Gibbs sampler. Thus the model is a MS(3)-SV as in (11) with fixed ρ equal to 0.5, $v_t \sim N(0, 0.2)$ and the intercept equal to

$$\mu_{s_t} = \begin{cases} -1 & \text{if} \quad s_t = 1 \\ -2 & \text{if} \quad s_t = 2 \\ -5 & \text{if} \quad s_t = 3 \end{cases} .$$

The state variables are generated by a first order Markov process with transition probability matrix

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.9 & 0 & 0.05 \\ 0 & 0.95 & 0.05 \\ 0.1 & 0.05 & 0.9 \end{pmatrix}$$

which implies high persistence in each regime. A dataset of $n = 400$ observations has been simulated from the model. We estimate the model with the filter proposed in Section 6 and 2,000 iterations are considered.

Results are summarized in Table 4.1 where means and standard deviations are given, together with the Bayesian estimators of So, Lam and Li and true values. As point estimate, all the approximated filters give close values to the corresponding true one. The persistence parameter ρ is better captured by Approximation 3 or 4, which is also the faster; those seem to be the best choices. Finally, our estimates obtained via Kalman filters give closer result to the true values with respect to the Bayesian counterpart. In fact, the Mean Square Error (MSE) value for the third approximation is equal to 0.00203 and the MSE computed by So, Lam and Li is

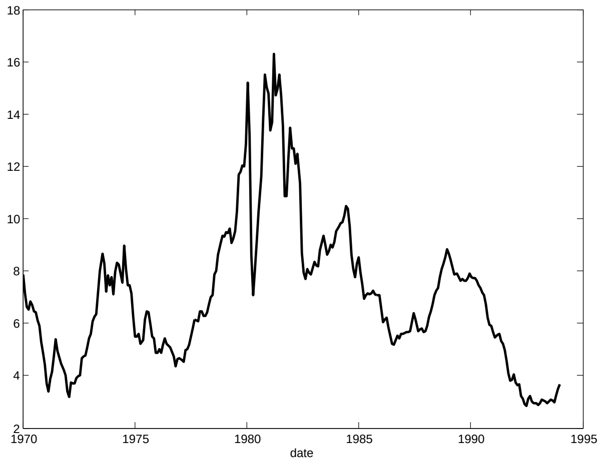

Figure 3.1: The panel contains a time series plot of monthly one-month US Treasury bill rates (in annualized percentage term). The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained from FRED database.

0.00330.

(II) Real Data: an application on US Treasury Bill rates

As a second application, we use real data and the same dataset as in Gray (1996). The data are one-month US Treasury bill rates obtained from FRED for the period January 1970 trought April 1994. Figure 4.1 plots the data. It is immediate the dramatic increase in interest rates that occurred during the Fed experiment and the OPEC oil crisis, which leads us to consider a 2 regimes model.

Then, we fit the model in (1) as MS(2)-GARCH and in (11) as MS(2)-SV with both changes in regimes in the intercept term and in the persistence parameters of the volatility process. The values of the estimation are reported in Table 4.3 and 4.4, respectively. Table 4.3 describes the estimated values along with robust standard errors of model (1). In particular, the model estimated by linear filter with Kim's approximation is labelled with Approximation 1. The following approximations instead are those in Section 5, respectively. Note that Approximation 2 mimics the auxiliary model of Gray (1996) and values are in fact in line (see Gray(1996), Table 3, p.44). The high-volatility regime is characterized by more sentivity to recent shocks ($\alpha_2 > \alpha_1$) and less persistence ($\beta_2 < \beta_1$) than the low-volatility regime. Within each regime, the GARCH processes are stationary ($\alpha_i + \beta_i < 1$) and the parameter estimates suggest that the regimes

Estimation	$\hat{p}_{00}$	$\hat{p}_{10}$	$\hat{p}_{01}$	$\hat{p}_{11}$	$\hat{p}_{02}$	$\hat{p}_{12}$	$\hat{\rho}$	$\hat{\sigma}_v^2$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	**Time***
Approximation 1	0.9489	0.0000	0.0000	0.9535	0.0518	0.0520	0.4838	0.2083	-0.8836	-2.0194	-4.9357	1.86
	(0.1165)	(0.0001)	(0.0001)	(0.0167)	(0.0009)	(0.0023)	(0.0207)	(0.0034)	(0.1073)	(0.0226)	(0.1909)	
Approximation 2	0.9263	0.0010	0.0000	0.9197	0.0540	0.0598	0.4858	0.2486	-0.8299	-2.096	-5.0332	1.92
	(0.0647)	(0.0005)	(0.0019)	(0.0689)	(0.0113)	(0.0217)	(0.1105)	(0.0697)	(0.3499)	(0.3628)	(0.5486)	
Approximation 3	0.9529	0.0011	0.0000	0.9334	0.0560	0.0471	0.5016	0.2279	-0.8659	-2.0545	-4.9824	1.83
	(0.0256)	(0.0017)	(0.0076)	(0.0912)	(0.0253)	(0.0212)	(0.0463)	(0.1027)	(0.5143)	(0.6148)	(0.7895)	
Approximation 4	0.9529	0.0011	0.0000	0.9334	0.0560	0.0471	0.5016	0.2279	-0.8659	-2.0545	-4.9824	1.83
	(0.0256)	(0.0017)	(0.0076)	(0.0912)	(0.0253)	(0.0212)	(0.0463)	(0.1027)	(0.5143)	(0.6148)	(0.7895)	
So, Lam and Li	0.877	0.026	0.027	0.941	0.079	0.040	0.513	0.314	-0.916	-2.071	-5.005	–
	(0.049)	(0.040)	(0.031)	(0.040)	(0.031)	(0.024)	(0.059)	(0.160)	(0.231)	(0.167)	(0.332)	
True values	0.95	0	0	0.95	0.05	0.05	0.5	0.2	-1	-2	-5	–

Table 3.1: Estimation results of simulated data from model (11). Robust standard errors in parenthesis. [*]: Minutes for single iteration. The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained by the FRED database.

Approximations	$\hat{p}$	$\hat{q}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\omega}_1$	$\hat{\omega}_2$
Approximation 1	0.8467	0.9982	0.4113	0.0085	0.0184	0.4967	0.023	0.068
	(0.1402)	(0.0334)	(0.0407)	(0.0927)	(0.0338)	(0.0331)	(0.2626)	(0.0908)
Approximation 2	0.8018	0.9157	0.391	0.0062	0.0203	0.4801	0.045	0.0713
	(0.1146)	(0.0157)	(0.1624)	(0.2609)	(0.0640)	(0.1089)	(0.3591)	(0.2853)
Approximation 3	0.8467	0.9983	0.4112	0.0086	0.0184	0.4967	0.023	0.068
	(0.1406)	(0.0336)	(0.0420)	(0.0923)	(0.0342)	(0.0337)	(0.2651)	(0.0799)
Approximation 4	0.8119	0.9160	0.397	0.0064	0.0212	0.4831	0.039	0.0692
	(0.1140)	(0.0157)	(0.1629)	(0.2618)	(0.0642)	(0.1092)	(0.3603)	(0.2867)

Table 3.2: Estimation of the parameters in model (1) MS(2)-GARCH. Robust standard errors in parenthesis. The observables are one-month US Treasury bill rates (in annualized percentage term). The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained from FRED database.

Approximations	$\hat{p}$	$\hat{q}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\sigma}_v$	$\hat{\mu}_1$	$\hat{\mu}_2$
Approximation 1	0.8935	0.9542	0.5492	0.9256	0.2863	0.2460	0.4165
	(0.0345)	(0.3633)	(0.1252)	(0.1260)	(0.0208)	(2.5734)	(2.5797)
Approximation 2	0.8677	0.8676	0.5388	0.9852	0.3478	0.2952	0.4905
	(0.0347)	(0.2813)	(0.0651)	(0.0585)	(0.1148)	(1.3258)	(1.2977)
Approximation 3	0.8668	0.8435	0.6063	0.9730	0.2706	0.3143	0.5707
	(0.0197)	(0.2288)	(0.0863)	(0.0620)	(0.3524)	(1.7598)	(2.4015)
Approximation 4	0.8668	0.8435	0.6063	0.9730	0.2706	0.3143	0.5707
	(0.0197)	(0.2288)	(0.0863)	(0.0620)	(0.3524)	(1.7598)	(2.4015)

Table 3.3: Estimation of the parameters in model (11) MS(2)-SV. Robust standard errors in parenthesis. The observables are one-month US Treasury bill rates (in annualized percentage term). The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained from FRED database.

are very persistent, so the source of volatility persistence will be important. With regards to the MS-SV model, the approximated filters are presented in Section 6. Most of the masses in the transition probability matrix are concentrated in the diagonal, implying medium-high persistence in each regime. Moreover, the first regime is associated with a intermediate level of persistence in the volatility process while the second shows a highly-persistent volatility, with values close to one. In both models, however, the four approximations are not very dissimilar to the others.

Figure 4.2 contains plots of smoothed probabilities $Pr(s_t = 1|\Phi_T)$ which are of interest to determine if and when the regime switching occurs. The smoothed probability plots manage to identify crises periods that affected the market indices. The top panel of Figure 4.2 refers to

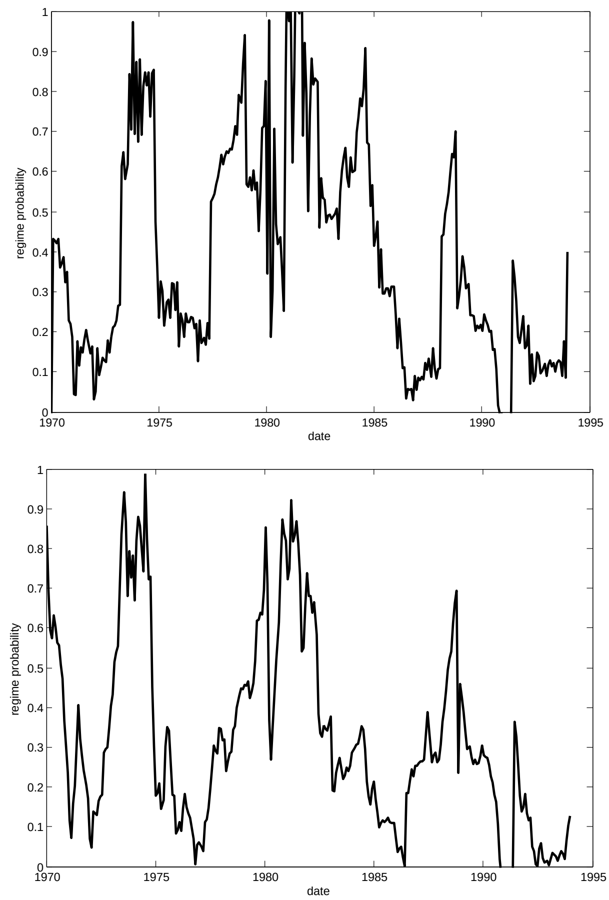

Figure 3.2: The top panel refers to MS-GARCH model and bottom panel to the MS-SV model. They represent smoothed probabilities being in high-volatility regime. Parameters estimates are based on a data set of one-month Treasury Bill rates, reported in annualized percentage terms. The sample period is from January 1970 to April 1994; a total of 1.267 observations. The data are obtained from FRED database.

the MS-GARCH and the bottom panel to the MS-SV model. In particular, both plots identify three periods of high-variance. The first (1973-1975) corresponds to the OPEC oil crisis. The second is shorter and more precise in the bottom panel and correspond to the Fed experiment (1979-1983). The third is a short period around 1987 after stock market crash.

3.8 Conclusion

In this paper we deal with Markov Switching models for volatility. In particular, we firstly consider MS-GARCH models which are known to suffer of path-dependence, i.e., dependence of the entire path history of the data. This makes Quasi Maximum Likelihood procedure unfeasible to apply. Hence, some solutions to overcome this problem have been proposed in the literature and particularly through the estimation of auxiliary models that allow similar effects of the original MS-GARCH. However, rewriting the model in a suitable state space representation, we propose an approximated linear filter following the line of Kim and Nelson (1999) and then we are able to prove duality in the estimation by Kalman filter and auxiliary models. Moreover, we introduce a linear filter also for MS-SV model on which different conditioning sets in the approximation step yield more flexibility in the estimation. We apply those methods to a simulation study and Treasury bill rates (the same dataset as in Gray (1996)). These applications show the feasibility of the linear filter for both MS models. In particular, this method has the advantage of avoiding fine-tuning procedures implemented in most Bayesian estimation techniques. In fact, giving some initial conditions, the only duty of the researcher is to decide which approximation to adopt in the filtering procedure. So, the proposed methods have a large applicability in financial and economics exercises and potential applications are those dealing with time varying volatility.

References

Augustyniak, M., (2013), "Maximum Likelihood Estimation of the Markov-switching GARCH model", forthcoming in *Computational Statistics and Data Analysis*.

Bauwens, L., Preminger, A., and Rombouts, J. (2010), "Theory and Inference for a Markov Switching GARCH Model", *Econometrics Journal*, Vol.13 **(2)**, 218–244.

Bauwens, L., Dufays, A., and Rombouts, J. (2011), "Marginal Likelihood for Markov-switching and Change-Point GARCH", *CORE discussion paper*, 2011/13.

Billio, M. and Caporin, M. (2005), "Multivariate Markov Switching dynamic conditional correlation GARCH representations for contagion analysis", *Statistical Methods and Applications*, **14**, 145–161.

Billio, M., Casarin, R., Osuntuy, A., (2012), "Efficient Gibbs Sampling for Markov Switching GARCH Models", Working Paper N. 35 , Dept. of Economics, University Ca' Foscari of Venice.

Billio, M., Monfort A., and Robert, C.P., (1998a), "A MCMC approach to maximum likelihood estimation", *Prague Stochastics '98*, Vol. 1, Ed. M. Huskova, P. Lachout and J.A. Visek, Union of Czech Mathematicians and Physicists, 49–54.

Billio, M., Monfort A., and Robert, C.P., (1998b), "The simulated likelihood ratio (SLR) method", Document de Travail du CREST 9828, Paris.

Bollerslev, T. (1986), "Generalized autoregressive conditional heteroskedasticity", *Journal of Econometrics*, **52**, 5–59.

Bollerslev, T. (1990), "Modelling the coherence in short-run nominal exchage rates: a multivariate generalized arch model", *Review of Economics and Statistics*, **72**, 498–505.

Cai, J. (1994), "A Markov model of unconditional variance in ARCH", *Journal of Business and Economic Statistics*, **12**, 309–316.

Casarin, R., (2004), Bayesian Inference for Generalised Markov Switching Stochastic Volatility Models, Cahier du CEREMADE N. 0414, University Paris Dauphine.

Dueker, M.J. (1997), "Markov Switching in GARCH Processes and Mean Reverting Stock Market Volatility", *Journal of Business and Economic Statistics*, Vol.15 **(1)**, 26–34.

Engle, R. (2002), "Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models", *Journal of Business and Economic Statistics*, Vol.20 **(3)**, 339–350.

Francq, C. and Zakoïan, J.M. (2001), "Stationarity of Multivariate Markov Switching ARMA Models", *Journal of Econometrics*, Vol.102, 339–364.

Francq, C. and Zakoïan, J.M. (2012), "Strict Stationarity and Estimation of Explosive and Stationary Generalized Autoregressive Conditional Heteroscedasticity Models", *Econometrica*, Vol.80 **(2)**, 821–861.

Gourieroux, C. and Monfort, A. (1997), "Time Series and Dynamic Models", *Cambridge University Press*, 1997.

Gray, S.F. (1996), "Modelling the Conditional Distribution of Interest Rates as a Regime-Switching Process", *Journal of Financial Economics*, Vol.42, 27–62.

Haas, M., Mittnik, S. and Paolella, M.S. (2004), "A new approach to Markov-Switching GARCH Models", *Journal of Financial Econometrics*, **2**, 493–530.

Hamilton, J.D. and Susmel, R. (1994), "Autoregressive conditional heteroskedasticity and changes in regime", *Journal of Econometrics*, **64**, 307–333.

Harvey, A.C., Ruiz, E. and Shephard, N. (1994), " Multivariate stochastic variance models", *Review of Economic Studies* **61**, 247–264.

Henneke, J.S., Rachev, S.T., Fabozzi, F.J. and Metodi, N. (2011), "MCMC-based estimation of Markov Switching ARMA–GARCH models", *Applied Economics*, 43 **(3)**, 259–271.

Klaassen, F. (2002), "Improving GARCH Volatility Forecasts with Regime-Switching GARCH", *Empirical Economics*, Vol.27 **(2)** , 363–394.

Kim, C.J. (1994), "Dynamic Linear Models with Markov Switching", *Journal of Econometrics*, Vol.64, 1–22.

Kim, C.J. and Nelson, C.R. (1999), "State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications", *The MIT Press*, 1999.

Lamoureux, C.G. and Lastrapes, W.D. (1990) "Persistence in variance, structural change, and the GARCH model", *Journal of Business and Economic Statistics*, **8**, 225–234.

Nelson, D.B. (1990), "Stationarity and persistence in the GARCH(1,1) model", *Econometric Theory*, **6**, 318–344.

Pelletier, D. (2006), "Regime switching for dynamic correlation", *Journal of Econometrics*, **131**, 445–473.

Shephard, N. (1996), "Statistical aspects of ARCH and stochastic volatility", in *Time Series Models with Econometric, Finance and Other Applications*, Cox, D.R., Hinkley, D.V., Barndorff-Nielsen, O.E. (eds.), Chapman and Hall, London, 1–677.

So, K.P., Lam, K. and Li, W.K. (1998), "A stochastic volatility model with Markov switching", *Journal of Business and Economic Statistics*, Vol.16 **(2)**, 244–253.

3.9 Appendix

Appendix A – Computation details of some Formulae

A1. We show that $p_{t|t-1,t} = p(s_t|s_{t-1}, \Psi_t)$ can be espressed in terms of $p_{t|t-1,t-1}$ and the conditional density of ϵ_t which depends on the current regime s_t and the past regimes, i.e, $f(\epsilon_t|s_1, \ldots, s_t, \Psi_{t-1})$. In fact,

$$\begin{aligned} p_{t|t-1,t} &= p(s_t|s_{t-1}, \Psi_t) = p(s_t|s_1, \ldots, s_{t-1}, \Psi_t) \\ &= p(s_t|s_1, \ldots, s_{t-1}, \epsilon_t, \Psi_{t-1}) \\ &= \frac{f(\epsilon_t|s_1, \ldots, s_t, \Psi_{t-1})p(s_t|s_1, \ldots, s_{t-1}, \Psi_{t-1})}{f(\epsilon_t|s_1, \ldots, s_{t-1}, \Psi_{t-1})} \\ &= \frac{f(\epsilon_t|s_1, \ldots, s_t, \Psi_{t-1})p(s_t|s_{t-1}, \Psi_{t-1})}{f(\epsilon_t|s_1, \ldots, s_{t-1}, \Psi_{t-1})} \\ &= \frac{f(\epsilon_t|s_1, \ldots, s_t, \Psi_{t-1})p_{t|t-1,t-1}}{f(\epsilon_t|s_1, \ldots, s_{t-1}, \Psi_{t-1})} \end{aligned}$$

where

$$\begin{aligned} f(\epsilon_t|s_1, \ldots, s_{t-1}, \Psi_{t-1}) &= \sum_{s_t=1}^{M} f(\epsilon_t|s_1, \ldots, s_t, \Psi_{t-1})p(s_t|s_{t-1}, \Psi_{t-1}) \\ &= \sum_{s_t=1}^{M} f(\epsilon_t|s_1, \ldots, s_t, \Psi_{t-1})\, p_{t|t-1,t-1}. \end{aligned}$$

A2. Here we derive the approximation of Kim and Nelson's Filter applied to model in (8), which is Formula (9):

$$\begin{aligned} B^j_{t|t} &= \frac{\sum_{i=1}^{M} B^{(i,j)}_{t|t} p(s_{t-1} = i, s_t = j|Y_t)}{p(s_t = j|Y_t)} \\ &= \sum_{i=1}^{M} \frac{p(s_{t-1} = i, s_t = j|Y_t)}{p(s_t = j|Y_t)} B^{(i,j)}_{t|t} \\ &= \sum_{i=1}^{M} p(s_{t-1} = i|s_t = j, Y_t)\ B^{(i,j)}_{t|t} \\ &= \sum_{i=1}^{M} p_{ij,t-1|t,t}\ B^{(i,j)}_{t|t}. \end{aligned}$$

Appendix B – Stationarity Conditions

Let us consider the MS-GARCH model in (1). Then we have

$$
\begin{aligned}
E(\epsilon_t^2) = E(\sigma_t^2) = E(E(\sigma_t^2|s_t)) &= \sum_{j=1}^{M} E(\sigma_t^2|s_t = j)p(s_t = j) \\
&= \sum_{j=1}^{M} \pi_j(\omega_j + \alpha_j E(\epsilon_{t-1}^2) + \beta_j E(\sigma_{t-1}^2)) \\
&= \sum_{j=1}^{M} \pi_j \omega_j + \sum_{j=1}^{M} \pi_j(\alpha_j + \beta_j) E(\sigma_{t-1}^2).
\end{aligned}
$$

For any $n \geq 1$, we have

$$E(\sigma_t^2) = a \sum_{i=0}^{n-1} b^i + b^n E(\sigma_{t-n}^2)$$

where $a = \sum_{j=1}^{M} \pi_j \omega_j$ and $b = \sum_{j=1}^{M} \pi_j(\alpha_j + \beta_j)$. This immediately implies that the MS-GARCH process in (1) is *covariance stationary* if and only if $b < 1$. Of course, if $\delta_j = \alpha_j + \beta_j < 1$, for every $j = 1, \ldots, M$, the above condition is satisfied. Conversely, if the MS-GARCH is covariance stationary, at least one of the regimes is covariance stationary. The above condition is sufficient but non necessary for strict stationarity. By iteration, we get

$$
\begin{aligned}
\sigma_t^2 &= \omega_{s_t} + \alpha_{s_t} \epsilon_{t-1}^2 + \beta_{s_t} \sigma_{t-1}^2 \\
&= \omega_{s_t} + \sigma_{t-1}^2(\alpha_{s_t} u_{t-1}^2 + \beta_{s_t}) \\
&= \omega_{s_t} + [\omega_{s_{t-1}} + \sigma_{t-2}^2(\alpha_{s_{t-1}} u_{t-2}^2 + \beta_{s_{t-1}})](\alpha_{s_t} u_{t-1}^2 + \beta_{s_t}) \\
&\vdots \\
&= \omega_{s_t} + \sum_{k=1}^{\infty} \omega_{s_{t-k}} \prod_{i=1}^{k} (\alpha_{s_{t-i+1}} u_{t-i}^2 + \beta_{s_{t-i+1}}).
\end{aligned}
$$

For every $n \geq 2$, define

$$\sigma_{t,n}^2 = \omega_{s_t} + \sum_{k=1}^{n-1} \omega_{s_{t-k}} \prod_{i=1}^{k} a_{s_{t-i+1}}(u_{t-i}^2)$$

where $a_{s_t}(x) = \alpha_{s_t} x^2 + \beta_{s_t}$. Now

$$\sum_{k=1}^{n-1} \log[\omega_{s_{t-k}} \prod_{i=1}^{k} a_{s_{t-i+1}}(u_{t-i}^2)] = \sum_{k=1}^{n-1} \{\log \omega_{s_{t-k}} + \sum_{i=1}^{k} \log a_{s_{t-i+1}}(u_{t-i}^2)\}$$

is monotone. Then the limit $n \to +\infty$ is finite whenever

$$E[\log(\alpha_{s_t} u_{t-1}^2 + \beta_{s_t})] < 0.$$

Here log denotes the natural logarithm as usual. But we have

$$E[\log(\alpha_{s_t} u_{t-1}^2 + \beta_{s_t})] = E[E[\log(\alpha_{s_t} u_{t-1}^2 + \beta_{s_t})|s_t]] = \sum_{j=1}^{M} \pi_j E[\log(\alpha_j u_{t-1}^2 + \beta_j)].$$

So we get that $\sigma_t^2 < +\infty$ a.s. (almost surely) and $\{\epsilon_t^2, \sigma_t^2\}$ is *strictly stationary* if

$$\sum_{j=1}^{M} \pi_j E[\log(\alpha_j u_{t-1}^2 + \beta_j)] < 0.$$

This extends the strictly stationarity condition given by Francq and Zakoïan (2012) for a GARCH(1,1) model to the case of changing in regime. See also Theorem 1 in Bauwens et al. (2010). Of course, the covariance stationarity condition implies strict stationarity, but the converse is not true in general.

The MS-GARCH(1,1) model in (1) can be represented by a MS-ARMA(1,1) process as in (7)

$$(1 - \delta_{s_t} L)\epsilon_t^2 = \omega_{s_t} + (1 + \theta_{s_t} L)v_t$$

where $\delta_{s_t} = \alpha_{s_t} + \beta_{s_t}$ and $\theta_{s_t} = -\beta_{s_t}$. The necessary and sufficient condition for second-order stationarity of univariate MS-ARMA(1,1) models was given by Francq and Zakoïan (2001), see Example 3 pag.351. We apply their result in our case. Let us consider the $M \times M$ matrix

$$\Omega = (a_{ij})_{i,j=1,\ldots,M}$$

where $a_{ij} = p_{ji}\delta_i^2$. Let $\rho(\Omega)$ be the spectral radius of the matrix, that is, its largest eigenvalue in modulus. From Francq and Zakoïan (2001), $\rho(\Omega) < 1$ if and only if the process (ϵ_t^2) in (1) is *second-order stationary* in the case where, for at least one regime, the AR and MA polynomials have no common roots. For our model, this means that $\delta_j \neq -\theta_j$, that is, $\alpha_j > 0$ for some $j = 1, \ldots, M$. Finally, note that the MA part in the process (ϵ_t^2) does not matter for the second-order stationarity condition.

Finally, with regards to the MS-SV model in (12), its MS-ARMA representation is easily obtained as follows

$$(1 - \rho_{s_t} L) y_t = \xi_{s_t} + (1 + \beta_{s_t} L) z_t$$

where $\xi_{s_t} = \alpha - \rho_{s_t}\alpha + \mu_{s_t}$ and $z_t + \beta_{s_t} z_{t-1} = v_t + e_t - \rho_{s_t} e_{t-1}$. Thus stationarity conditions as discussed above apply. More precisely, the process is second-order stationary if and only if $\rho(\widetilde{\Omega}) < 1$, where $\widetilde{\Omega}$ is the matrix obtained by replacing δ_j by ρ_j in the definition of Ω.

Chapter 4

Estimation and Spectral Representation

4.1 Analysis of the Likelihood Function for Markov Switching VAR(CH) Models

Abstract. *In this work we give simple matrix formulae for maximum likelihood estimates of parameters in a broad class of vector autoregressions subject to Markovian changes in regime. This allows us to determine explicitly the asymptotic variance-covariance matrix of the estimators, giving a concrete possibility for the use of the classical testing procedures. Our discussion is based on the fundamental work developed by Hamilton, Kim, and Krolzig, and is related with the analytical derivatives obtained by Gable, Van Norden and Vigfusson for univariate Markov-switching models without lags. Finally, in the context of multivariate ARCH models with changes in regime we show how analytic derivatives of the log likelihood can be successfully employed for estimation purposes.[JEL Classification: C01, C32, C51]*

Key words: Time series with changes in regime, Markov-switching VAR models, filtering, smoothing, MLE, asymptotic variance matrix.

4.1.1 Introduction

In this paper we derive maximum likelihood estimators and the asymptotic variance for a broad class of dynamic models with regime-switching parameters. The main contribution in this field is due to Hamilton (1990, 1993). In his work at the centre of interest is the Expectation Maximization (EM) algorithm. A limitation of this procedure is its implementation in the case of models with autoregressive dynamic, which conduces to some problems. A theoretical improvement on this side has been proposed by Krolzig (1997) who derived the normal equations of the Maximum Likelihood Estimation (MLE) for Markov-Switching Vector Autoregressive (MS-VAR) models. However, with regards to the estimation of asymptotic variance-covariance matrix, it was explicitly stated in his work that, for a generic MS-VAR model, it is often impracticable to evaluate the asymptotic information matrix analytically. The contribution of this paper is to provide simple matrix formulae for the MLE of parameters in the general case of MS-VAR models and to determine explicitly the asymptotic variance-covariance matrix, allowing for the implementation of classical testing procedures. As remarked in Gable, Van Norden and Vigfusson (1997), estimation for Markov switching models is usually done using numerical techniques that approximate the derivative by the change in the likelihood function for small changes in the parameter vector. This is not especially efficient, however, as such techniques typically require a lot of numerical evaluations (see Section 3 of the quoted paper). Using analytical gradients the number of calculations can be greatly reduced, and this in turn considerably speeds up MLE with no loss of accuracy. This gives a strong motivation for searching explicit formulae of ML estimates and their asymptotic variance-covariance matrix. The rest of the paper is organized as follows. In Section 2 we survey the basic definitions on Markov chains and the main steps of the Maximum Likelihood Estimation (MLE) for time series models of changes in regime. The contents of such a section are based on the arguments treated in the Hamilton book (1994), in the Krolzig book (1997), and in the Hamilton papers (1990, 1993). In Sections 3 and 4 we give explicit formulae for the ML estimates of the parameters for some Markov switching AR models. In Section 5 we determine the analytic derivatives of the log likelihood for multivariate ARCH models subject to Markovian changes in regime, and discuss the properties of the ML estimates. Our recursive matrix formulae are very simple and different in some cases to those listed in the known literature. Then we discuss the asymptotic properties of the ML estimators, and compute exactly the asymptotic variance-covariance matrix of them. This is also combined with classical specification testing procedures. A brief summary and conclusion follow in Section 6.

4.1.2 Time Series Models of Changes in Regime

2.1. The model. We consider time series models in which the parameters can change as a result of a regime-shift variable, described as the outcome of an unobserved Markov chain. The basic references are the Hamilton book (1994), Chp.22, Kim (1994), the Krolzig book (1997), Chps.5, 6 and 7, and the Hamilton papers (1990) and (1993). Let $\mathbf{y}_t$ be an $(K \times 1)$ vector of observed endogenous variables and let $\mathbf{Y}_t$ denote a vector containing all observations obtained through date t, that is, $\mathbf{Y}_t = (\mathbf{y}_t^{'}, \mathbf{y}_{t-1}^{'}, \dots)^{'}$. To take in account changes in the process $(\mathbf{y}_t)$, we assume it to be governed by an unobserved random variable, called *state* or *regime*, which is discrete-valued. The simplest time series model for a discrete-valued random variable is a Markov chain. Let $(s_t)_{t\geq 0}$ be an M-state, homogeneous, irreducible and ergodic *Markov chain*. Let $\mathbf{P} = (p_{ij})_{i,j=1,\dots,M}$ denote the transition matrix of the chain, where p_{ij} gives the probability that the state $s_{t-1} = i$ will be followed by the state $s_t = j$. As usual, we suppose that the probability that s_t equals some particular value depends on the past only through the most recent value s_{t-1}. Ergodicity implies the existence of a stationary vector of probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)^{'}$ satisfying $\mathbf{P}'\boldsymbol{\pi} = \boldsymbol{\pi}$ and $\mathbf{i}_M^{'}\boldsymbol{\pi} = 1$, where $\mathbf{i}_M$ denotes an $(M \times 1)$ vector of ones. Irreducibility implies that $\pi_i > 0$ for $i = 1, \dots, M$, meaning that all unobservable states are possible. An useful representation for a Markov chain is obtained by letting $\boldsymbol{\xi}_t$ denote a random $(M \times 1)$ vector whose jth element is equal to unity if $s_t = j$ and zero otherwise. We see that the conditional expectation of $\boldsymbol{\xi}_{t+1}$ satisfies the property $E(\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \dots) = E(\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t) = \mathbf{P}'\boldsymbol{\xi}_t$. This implies that it is possible to express a Markov chain in the AR(1) form

$$\boldsymbol{\xi}_{t+1} = \mathbf{P}'\boldsymbol{\xi}_t + \mathbf{v}_{t+1} \tag{2.1}$$

where the innovation $\mathbf{v}_{t+1} = \boldsymbol{\xi}_{t+1} - E(\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \cdots)$ is a zero mean martingale difference sequence. The vector $\boldsymbol{\pi}$ of ergodic probabilities can be described as the unconditional expectation of $\boldsymbol{\xi}_t$, that is, $\boldsymbol{\pi} = E(\boldsymbol{\xi}_t)$.

If the process is governed by regime $s_t = j$ at date t, then the *conditional density* of $\mathbf{y}_t$ is assumed to be given by

$$p(\mathbf{y}_t|s_t = j, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) \tag{2.2}$$

where $\boldsymbol{\theta}$ is a vector of parameters characterizing the conditional density. Then there are M different densities represented by (2.2) for $j = 1, \ldots, M$, which will be collected in an $(M \times 1)$ vector denoted by $\boldsymbol{\eta}_t(\boldsymbol{\theta})$, i.e.,

$$\boldsymbol{\eta}_t(\boldsymbol{\theta}) = \begin{pmatrix} p(\mathbf{y}_t|s_t = 1, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) \\ \vdots \\ p(\mathbf{y}_t|s_t = M, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) \end{pmatrix}. \tag{2.3}$$

Assumption A1. The conditional density in (2.3) is assumed to depend only on the current regime s_t and not on past regimes, i.e.,

$$p(\mathbf{y}_t|\mathbf{Y}_{t-1}, s_t = j; \boldsymbol{\theta}) = p(\mathbf{y}_t|\mathbf{Y}_{t-1}, s_t = j, s_{t-1} = i, s_{t-2} = k, \ldots; \boldsymbol{\theta})$$

though this is not really restrictive. See Hamilton (1994), Chp.22.

Assumption A2. The random variable s_t evolves according a Markov chain that is independent of past observations on $\mathbf{y}_t$, i.e.,

$$Pr(s_t = j|s_{t-1} = i, s_{t-2} = k, \ldots, \mathbf{Y}_{t-1}) = Pr(s_t = j|s_{t-1} = i) = p_{ij}$$

Let $\boldsymbol{\rho} = (p_{11}, p_{12}, \ldots, p_{MM})'$ denote the $(M^2 \times 1)$ vector of Markov transition probabilities. The population parameters that describe a time series governed by A1 and A2 consist of $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$ (here we always set, for simplicity, the initial state $\boldsymbol{\xi}_0$ equal to $\boldsymbol{\pi}$). We collect the unknown parameters to be estimated in a single vector $\boldsymbol{\lambda} = (\boldsymbol{\theta}', \boldsymbol{\rho}')'$. One important objective is to maximize the likelihood function of the observed data $p(\mathbf{Y}_T; \boldsymbol{\lambda}) = p(\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_1; \boldsymbol{\lambda})$ by choice of the population parameters vector $\boldsymbol{\lambda}$.

2.2. Optimal Inference for the Regime. Another objective will be to estimate the value of $\boldsymbol{\lambda}$ based on observation of $\mathbf{Y}_T$. Suppose for the moment that the value of $\boldsymbol{\lambda}$ is somehow known with certainty to the analyst. Nevertheless, we do not know which regime the process was in at every date in the sample. Instead the best we can do is to provide inference for $\boldsymbol{\xi}_t$ given a specified observation set $\mathbf{Y}_\tau$, $\tau \leq T$. The statistical tools are the *filter* and *smoother* recursions which reconstruct the time path of the regime ($\boldsymbol{\xi}_t$) under alternative information sets.

Collect the conditional probabilities $Pr(s_t = j|\mathbf{Y}_\tau)$, for $j = 1, \ldots, M$, in an $(M \times 1)$ vector denoted by

$$\widehat{\boldsymbol{\xi}}_{t|\tau} = \begin{pmatrix} Pr(s_t = 1|\mathbf{Y}_\tau) \\ \vdots \\ Pr(s_t = M|\mathbf{Y}_\tau) \end{pmatrix}. \tag{2.4}$$

Then $\widehat{\boldsymbol{\xi}}_{t|\tau}$ turns out to be the conditional mean of $\boldsymbol{\xi}_t$ given $\mathbf{Y}_\tau$, that is, $\widehat{\boldsymbol{\xi}}_{t|\tau} = E(\boldsymbol{\xi}_t|\mathbf{Y}_\tau)$.

The following theorem is well-known. It gives a fast algorithm for calculating the filtered and smoothed regime probabilities For the proof see Hamilton (1994), Chp.22, Krolzig (1997), Chp.5, and Kim (1994).

Theorem 2.1. i) *The optimal inference and forecast for each date t in the sample can be found by iterating on the following pair of recursive formulae*

$$\begin{aligned} \widehat{\boldsymbol{\xi}}_{t|t} &= \frac{\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t}{\mathbf{i}'_M(\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t)} \\ \widehat{\boldsymbol{\xi}}_{t+1|t} &= \mathbf{P}'\widehat{\boldsymbol{\xi}}_{t|t} \end{aligned} \tag{2.5}$$

where the symbol $\odot$ denotes the element-by-element multiplication. Furthermore, the conditional probability density of $\mathbf{y}_t$ based upon $\mathbf{Y}_{t-1}$ is given by

$$p(\mathbf{y}_t|\mathbf{Y}_{t-1};\boldsymbol{\theta}) = \mathbf{i}'_M(\widehat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t)$$

ii) *Smoothed inferences can be calculated using an algorithm which can be written, in vector* *form, as*

$$\widehat{\boldsymbol{\xi}}_{t|T} = \widehat{\boldsymbol{\xi}}_{t|t} \odot \{\mathbf{P}[\widehat{\boldsymbol{\xi}}_{t+1|T} (\div) \widehat{\boldsymbol{\xi}}_{t+1|t}]\} \tag{2.6}$$

where the symbol $(\div)$ *denotes element-by-element division.*

Remarks.

1) Given a starting value $\widehat{\boldsymbol{\xi}}_{1|0} = E(\boldsymbol{\xi}_1|\mathbf{Y}_0)$ (for example, $\widehat{\boldsymbol{\xi}}_{1|0} = \boldsymbol{\pi}$) and an assumed value for the population parameter vector $\boldsymbol{\lambda}$, one can iterate on (2.5) for $t = 1, \ldots, T$ to calculate the values of $\widehat{\boldsymbol{\xi}}_{t|t}$ and $\widehat{\boldsymbol{\xi}}_{t+1|t}$ for each date t in the sample. This gives the filtered regime probabilities $\widehat{\boldsymbol{\xi}}_{t|\tau}$, $t = \tau$ (*filtering*) and the predicted regime probabilities $\widehat{\boldsymbol{\xi}}_{t|\tau}$, $\tau < t$ (*forecasting*).

2) The smoothed regime probabilities $\widehat{\boldsymbol{\xi}}_{t|T}$ (*smoothing*) are found by iterating (2.6) backward for $t = T-1, T-2, \ldots, 1$. This iteration is started with $\widehat{\boldsymbol{\xi}}_{T|T}$ which is obtained from (2.5) for $t = T$. This algorithm is valid only when s_t follows a first-order Markov chain as in A2, when the conditional density (2.2) depends on $s_t, s_{t-1}, \ldots$ only through the current state s_t.

2.3. Maximum Likelihood Estimation of Parameters. In the iterations of Theorem 2.1 the parameter vector $\boldsymbol{\lambda}$ was taken to be a fixed known vector. Once the iteration has been completed for $t = 1, \ldots, T$ for a given fixed $\boldsymbol{\lambda}$, the value of the log likelihood implied by that value of $\boldsymbol{\lambda}$ is then known as follows (for these arguments we refer to the Krolzig book (1997), Chp.6):

$$\begin{aligned} L(\boldsymbol{\lambda}|\boldsymbol{Y}_T) := p(\mathbf{Y}_T|\boldsymbol{\lambda}) &= \int_{\boldsymbol{\xi}} p(\mathbf{Y}_T, \boldsymbol{\xi}|\boldsymbol{\lambda}) d\boldsymbol{\xi} \\ &= \int_{\boldsymbol{\xi}} p(\mathbf{Y}_T|\boldsymbol{\xi}, \boldsymbol{\theta}) Pr(\boldsymbol{\xi}|\boldsymbol{\xi}_0, \boldsymbol{\rho}) d\boldsymbol{\xi} \end{aligned} \tag{2.7}$$

where

$$p(\mathbf{Y}_T|\boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\boldsymbol{\xi}_t, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) \qquad Pr(\boldsymbol{\xi}|\boldsymbol{\xi}_0, \boldsymbol{\rho}) = \prod_{t=1}^{T} Pr(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1}; \boldsymbol{\rho}).$$

Then we have (see Krolzig (1997), formula (6.7))

$$\begin{aligned}\frac{\partial ln\, L(\boldsymbol{\lambda}|\mathbf{Y}_T)}{\partial \boldsymbol{\theta}} &= \sum_{t=1}^{T}\sum_{m=1}^{M} \frac{\partial ln\, p(\mathbf{y}_t|s_t = m, \mathbf{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} Pr(s_t = m|\mathbf{Y}_T; \boldsymbol{\lambda}) \\ &= \sum_{t=1}^{T}\sum_{m=1}^{M} \frac{\partial ln\, \eta_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \widehat{\xi}_{mt|T}(\boldsymbol{\lambda}).\end{aligned} \tag{2.8}$$

Thus the FOC condition for the ML estimates $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is given by (see Krolzig (1997), formula (6.8))

$$\sum_{t=1}^{T}\sum_{m=1}^{M} \frac{\partial ln\, \eta_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \widehat{\xi}_{mt|T}(\boldsymbol{\lambda}) = 0. \tag{2.9}$$

As usual, we assume that the transition probabilities are restricted by the conditions that $p_{ij} > 0$ and $\sum_{j=1}^{M} p_{ij} = 1$ for all i and j (recall that we set $\widehat{\xi}_{1|0} = \boldsymbol{\pi}$). Then the ML estimates $\widehat{\boldsymbol{\rho}}$ of $\boldsymbol{\rho}$ is given by the following formula, due to Hamilton (1990), formula (4.1) (see also Hamilton (1994), formula (22.4.16) or Krolzig (1997), formula (6.14)).

Theorem 2.2. *The maximum likelihood estimates for the elements of the transition probabilities vector $\boldsymbol{\rho} = (p_{11}, \ldots, p_{MM})'$ satisfy*

$$\widehat{p}_{ij} = \frac{\sum_{t=1}^{T} Pr(s_t = j, s_{t-1} = i|\mathbf{Y}_T; \widehat{\boldsymbol{\lambda}})}{\sum_{t=1}^{T} Pr(s_{t-1} = i|\mathbf{Y}_T; \widehat{\boldsymbol{\lambda}})} \tag{2.10}$$

where $\widehat{\boldsymbol{\lambda}}$ denotes the full vector of maximum likelihood estimates.

Thus the estimated transition probabilities $\widehat{p}_{ij}$ is essentially the number of times state i seems to have been followed by state j divided by the number of times the process was in state i.

We recall that estimation and inferences for multivariate Markov switching models are based on the *Expectation Maximization* EM algorithm, as described in Hamilton (1990). Maximization of the log-likelihood function within the current step of the EM algorithm is made faster by the fact that the FOC defining the ML estimators may often be written down in closed form. Finally, we mention that a state-space representation for VARMA models that enables ML estimation via EM algorithm can also be found in Metaxoglou and Smith (2007).

2.4. Consistency and Asymptotic Variance. Under quite general regularity conditions (such as identifiability, stability and the fact that the true parameter vector does not fall on the boundaries which we assume here), an ML estimator $\widehat{\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda}$ is *consistent* and *asymptotically normal* (see Lütkepohl (1991), Section C.4 and Krolzig (1997), Section 6.6.2):

$$\sqrt{T}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_a(\boldsymbol{\lambda})^{-1})$$

where $\mathcal{I}_a(\boldsymbol{\lambda})$ is the asymptotic information matrix

$$\mathcal{I}_a(\boldsymbol{\lambda}) = \lim_{T\to\infty} -T^{-1}E\Big[\frac{\partial^2 ln\, p(\mathbf{Y}_T;\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\Big] \tag{2.11}$$

Although other choices exist, i.e., either to use the conditional scores or a numerical evaluation of the second partial derivative of the log-likelihood function with respect to $\widehat{\boldsymbol{\lambda}}$, in applications it has become typical to employ the White sample estimator of $\mathcal{I}_a(\boldsymbol{\lambda})$ (see, for example, Krolzig (1997), formula (6.46)). The asymptotic normal distribution of the ML estimator $\widehat{\boldsymbol{\lambda}}$ ensures that standard inferential procedures (Wald test, Likelihood ratio LR and Lagrange multiplier LM) are available to test statistical hypothesis. However, notice that a necessary condition for the validity of such procedures is that the number M of regimes is unaltered under the null hypothesis. See Krolzig (1997), Section 7.4 for more details. The next sections are devoted to compute explicitly the ML estimates $\widehat{\boldsymbol{\lambda}}$ for some Markov-switching VAR models. Our matrix formulae are very simple and different in some cases to those obtained in the known literature. Then we determine the asymptotic variance of the estimators. For the basic identities and results on Matrix Calculus we refer for example to Fackler (2005), Greene (2008) and Petersen and Pedersen (2008).

4.1.3 The Basic Markov Switching Model

Suppose that the $(K \times 1)$ random vector $\mathbf{y}_t$ follows an M-regime Markov-switching autoregressive process without lags in the endogeneous variable, in short, an MS AR(p) model, for $p = 0$ (the case $p > 0$ will be treated in the next section). Such a model is also called the *Hidden*

Markov-chain process (see, for example, Krolzig (1997), Chp.3):

$$\mathbf{y}_t = \boldsymbol{\nu}_{s_t} + \boldsymbol{\Sigma}_{s_t}\mathbf{u}_t \tag{3.1}$$

where the innovations $(\mathbf{u}_t)$ represent a zero mean white noise process which we assume to be Gaussian, i.e., $\mathbf{u}_t \sim \text{NID}(\mathbf{0}, \mathbf{I}_K)$. As usual, s_t is a latent state variable driving all the matrices of parameters appearing in (3.1). The $(K \times 1)$ vector $\boldsymbol{\nu}_{s_t}$ collects the regime-dependent intercepts. The $(K \times K)$ matrix $\boldsymbol{\Sigma}_{s_t}$ represents the factor applicable to state s_t in a state-dependent Choleski factorization of the variance-covariance matrix $\boldsymbol{\Omega}_{s_t}$ of the variables of interest, i.e., $\boldsymbol{\Omega}_{s_t} = \boldsymbol{\Sigma}^2_{s_t}$. Thus we have

$$\mathbf{y}_t|_{s_t} \sim \text{NID}(\boldsymbol{\nu}_{s_t}, \boldsymbol{\Omega}_{s_t}) \tag{3.2}$$

The unknown parameter $\boldsymbol{\theta}$ consists of the elements of the intercept vectors $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$ and the variance-covariance matrices $(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_M)$, or, equivalently, their inverses. Notice that the independent parameters characterizing each matrix $\boldsymbol{\Omega}_m$, $m = 1, \ldots, M$, are at most $K(K+1)/2$. As usual, the vector $\boldsymbol{\rho} = (p_{11}, \ldots, p_{MM})'$ characterizes the unknown matrix $\mathbf{P}$, and it has $M(M-1)$ independent random variables. Hence Model (3.1) implies the estimation of a number of parameters which is equal to $\dim \Lambda_M = M[K + K(K+1)/2 + (M-1)]$ for the most general situation. The log of the conditional density in (2.2) becomes

$$ln\,\boldsymbol{\eta}_t(\boldsymbol{\theta}) = ln\,p(\mathbf{y}_t|s_t;\boldsymbol{\theta}) = -\frac{K}{2}ln(2\pi) - \frac{1}{2}ln|\boldsymbol{\Omega}_{s_t}| - \frac{1}{2}(\mathbf{y}_t - \boldsymbol{\nu}_{s_t})'\boldsymbol{\Omega}^{-1}_{s_t}(\mathbf{y}_t - \boldsymbol{\nu}_{s_t}).$$

The first derivatives are given by

$$\frac{\partial ln\,\boldsymbol{\eta}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\nu}_m} = \boldsymbol{\Omega}^{-1}_m(\mathbf{y}_t - \boldsymbol{\nu}_m) \qquad \frac{\partial ln\,\boldsymbol{\eta}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\Omega}^{-1}_m} = \frac{1}{2}\boldsymbol{\Omega}_m - \frac{1}{2}(\mathbf{y}_t - \boldsymbol{\nu}_m)(\mathbf{y}_t - \boldsymbol{\nu}_m)'$$

when $s_t = m$ and zero otherwise, for every $m = 1, \ldots, M$. So the FOC conditions in (2.9) give

$$\sum_{t=1}^{T} \boldsymbol{\Omega}^{-1}_m(\mathbf{y}_t - \boldsymbol{\nu}_m)\widehat{\xi}_{mt|T} = \mathbf{0} \tag{3.3}$$

and

$$\sum_{t=1}^{T}\big[\frac{1}{2}\mathbf{\Omega}_m-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\nu}_m)(\mathbf{y}_t-\boldsymbol{\nu}_m)^{'}\big]\widehat{\xi}_{mt|T}=\mathbf{0} \tag{3.4}$$

for every $m=1,\ldots,M$. Solving (3.3) and (3.4) for $\boldsymbol{\nu}_m$ and $\mathbf{\Omega}_m$, we get their ML estimates

$$\widehat{\boldsymbol{\nu}}_m=\big[\sum_{t=1}^{T}\widehat{\xi}_{mt|T}\big]^{-1}\big[\sum_{t=1}^{T}\mathbf{y}_t\widehat{\xi}_{mt|T}\big] \tag{3.5}$$

and

$$\widehat{\mathbf{\Omega}}_m=\big[\sum_{t=1}^{T}\widehat{\xi}_{mt|T}\big]^{-1}\big[\sum_{t=1}^{T}(\mathbf{y}_t-\widehat{\boldsymbol{\nu}}_m)(\mathbf{y}_t-\widehat{\boldsymbol{\nu}}_m)^{'}\widehat{\xi}_{mt|T}\big] \tag{3.6}$$

Substituting (3.5) into (3.6) yields

$$\widehat{\mathbf{\Omega}}_m=\big[\sum_{t=1}^{T}\widehat{\xi}_{mt|T}\big]^{-3}\big[\sum_{t=1}^{T}\sum_{\tau=1}^{T}\sum_{\sigma=1}^{T}(\mathbf{y}_t-\mathbf{y}_\tau)(\mathbf{y}_t-\mathbf{y}_\sigma)^{'}\widehat{\xi}_{m\tau|T}\widehat{\xi}_{m\sigma|T}\widehat{\xi}_{mt|T}\big]. \tag{3.7}$$

The ML estimates of p_{ij} are given by (2.10). However, we compute them explicitly for our case. Form the Lagrangean

$$\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\mu})=ln\,L(\boldsymbol{\lambda}|Y_T)-\sum_{t=1}^{M}\mu_i\big(\sum_{j=1}^{M}p_{ij}-1\big)$$

where $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_M)^{'}$ is the vector of Lagrange multipliers. Then we must have (see, for example Krolzig (1997), Section 6.3.2)

$$\frac{\partial\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\mu})}{\partial p_{ij}}=\frac{\partial ln\,L(\boldsymbol{\lambda}|Y_T)}{\partial p_{ij}}-\mu_i=0 \tag{3.8}$$

where

$$\frac{\partial ln\,L(\boldsymbol{\lambda}|Y_T)}{\partial p_{ij}}=\sum_{t=1}^{T}\sum_{n=1}^{M}\sum_{m=1}^{M}\frac{\partial ln\,Pr(s_t=m|s_{t-1}=n;\boldsymbol{\rho})}{\partial p_{ij}} \tag{3.9}$$
$$Pr(s_t=m,s_{t-1}=n|\mathbf{Y}_T;\boldsymbol{\lambda})$$

Since

$$\frac{\partial ln\, Pr(s_t = m|s_{t-1} = n; \boldsymbol{\rho})}{\partial p_{ij}} = p_{ij}^{-1}$$

when $m = j$ and $n = i$, and zero otherwise, (3.9) becomes

$$p_{ij}^{-1} \sum_{t=1}^{T} Pr(s_t = j, s_{t-1} = i|\mathbf{Y}_T; \boldsymbol{\lambda}) = \mu_i. \tag{3.10}$$

Summation over $j = 1, \ldots, M$ gives

$$\sum_{t=1}^{T}\sum_{j=1}^{M} Pr(s_t = j, s_{t-1} = i|\mathbf{Y}_T; \boldsymbol{\lambda}) = \mu_i\big(\sum_{j=1}^{M} p_{ij}\big) = \mu_i$$

that is

$$\mu_i = \sum_{t=1}^{T} Pr(s_{t-1} = i|\mathbf{Y}_T; \boldsymbol{\lambda}) = \sum_{t=1}^{T} \widehat{\xi}_{i,t-1|T}(\boldsymbol{\lambda}). \tag{3.11}$$

Substituting (3.11) into (3.10) and evaluating in $\widehat{\boldsymbol{\lambda}}$ gives (2.10), as requested. The consistency of the ML estimators follows easily from relations above. Substituting $\mathbf{y}_t$ by (3.1) for $s_t = m$ into (3.3) yields

$$\begin{aligned}\widehat{\boldsymbol{\nu}}_m &= \big[\sum_{t=1}^{T} \widehat{\xi}_{mt|T}\big]^{-1}\big[\sum_{t=1}^{T}(\boldsymbol{\nu}_m + \boldsymbol{\Sigma}_m \mathbf{u}_t)\widehat{\xi}_{mt|T}\big]\\ &= \boldsymbol{\nu}_m + \big[\sum_{t=1}^{T} \widehat{\xi}_{mt|T}\big]^{-1}\boldsymbol{\Sigma}_m\big[\sum_{t=1}^{T} \mathbf{u}_t\widehat{\xi}_{mt|T}\big]\end{aligned}$$

hence

$$\text{plim}_{T\to\infty}\, \widehat{\boldsymbol{\nu}}_m = \boldsymbol{\nu}_m + \text{plim}_{T\to\infty}\big[\frac{1}{T}\sum_{t=1}^{T} \widehat{\xi}_{mt|T}\big]^{-1}\boldsymbol{\Sigma}_m\, \text{plim}_{T\to\infty}\big[\frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t\widehat{\xi}_{mt|T}\big] = \boldsymbol{\nu}_m$$

for every $m = 1, \ldots, M$, as

$$\text{plim}_{T\to\infty}\, \frac{1}{T}\sum_{t=1}^{T} \widehat{\xi}_{mt|T} = E(\xi_{mt}) = \pi_m \tag{3.12}$$

and

$$\text{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t \widehat{\xi}_{mt|T} = E(\mathbf{u}_t \xi_{mt}) = E(\mathbf{u}_t)E(\xi_{mt}) = E(\mathbf{u}_t)\pi_m = \mathbf{0} \tag{3.13}$$

since $\boldsymbol{\xi}_t$ is independent of $\mathbf{u}_t$.

Using (3.1) for $s_t = m$, we get

$$\begin{aligned}(\mathbf{y}_t - \widehat{\boldsymbol{\nu}}_m)(\mathbf{y}_t - \widehat{\boldsymbol{\nu}}_m)' &= \big[(\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m) + \boldsymbol{\Sigma}_m \mathbf{u}_t\big]\big[(\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)' + \mathbf{u}_t' \boldsymbol{\Sigma}_m'\big] \\ &= (\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)(\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)' + (\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)\mathbf{u}_t' \boldsymbol{\Sigma}_m' \\ &+ \boldsymbol{\Sigma}_m \mathbf{u}_t (\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)' + \boldsymbol{\Sigma}_m \mathbf{u}_t \mathbf{u}_t' \boldsymbol{\Sigma}_m'\end{aligned}$$

hence

$$\begin{aligned}&\sum_{t=1}^{T}(\mathbf{y}_t - \widehat{\boldsymbol{\nu}}_m)(\mathbf{y}_t - \widehat{\boldsymbol{\nu}}_m)' \widehat{\xi}_{mt|T} = \big[\sum_{t=1}^{T} \widehat{\xi}_{mt|T}\big](\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)(\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)' \\ &+ (\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)\big[\sum_{t=1}^{T} \mathbf{u}_t' \widehat{\xi}_{mt|T}\big]\boldsymbol{\Sigma}_m' + \boldsymbol{\Sigma}_m\big[\sum_{t=1}^{T} \mathbf{u}_t \widehat{\xi}_{mt|T}\big](\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)' + \boldsymbol{\Sigma}_m\big[\sum_{t=1}^{T} \mathbf{u}_t \mathbf{u}_t' \widehat{\xi}_{mt|T}\big]\boldsymbol{\Sigma}_m'.\end{aligned}$$

From (3.6) we obtain

$$\begin{aligned}\widehat{\boldsymbol{\Omega}}_m =& (\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)(\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)' + \big[\frac{1}{T}\sum_{t=1}^{T} \widehat{\xi}_{mt|T}\big]^{-1}(\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)\big[\frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t' \widehat{\xi}_{mt|T}\big]\boldsymbol{\Sigma}_m' \\ &+ \big[\frac{1}{T}\sum_{t=1}^{T} \widehat{\xi}_{mt|T}\big]^{-1}\boldsymbol{\Sigma}_m\big[\frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t \widehat{\xi}_{mt|T}\big](\boldsymbol{\nu}_m - \widehat{\boldsymbol{\nu}}_m)' \\ &+ \big[\frac{1}{T}\sum_{t=1}^{T} \widehat{\xi}_{mt|T}\big]^{-1}\boldsymbol{\Sigma}_m\big[\frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t \mathbf{u}_t' \widehat{\xi}_{mt|T}\big]\boldsymbol{\Sigma}_m'.\end{aligned}$$

Using (3.12), (3.13) and the consistency of $\widehat{\boldsymbol{\nu}}_m$, we get

$$\begin{aligned}\text{plim}_{T\to\infty} \widehat{\boldsymbol{\Omega}}_m &= \text{plim}_{T\to\infty} \big[\frac{1}{T}\sum_{t=1}^{T} \widehat{\xi}_{mt|T}\big]^{-1}\boldsymbol{\Sigma}_m \, \text{plim}_{T\to\infty} \big[\frac{1}{T}\sum_{t=1}^{T} \mathbf{u}_t \mathbf{u}_t' \widehat{\xi}_{mt|T}\big]\boldsymbol{\Sigma}_m' \\ &= \pi_m^{-1}\boldsymbol{\Sigma}_m E(\mathbf{u}_t \mathbf{u}_t')E(\xi_{mt})\boldsymbol{\Sigma}_m' = \boldsymbol{\Sigma}_m^2 = \boldsymbol{\Omega}_m,\end{aligned}$$

as $\boldsymbol{\xi}_t$ is independent of $\mathbf{u}_t$, $E(\mathbf{u}_t \mathbf{u}_t') = \mathbf{I}_K$, $E(\xi_{mt}) = \pi_m$ and $\boldsymbol{\Sigma}_m$ is symmetric.

The consistency of $\widehat{p}_{ij}$ follows directly from (2.10) as

$$\begin{aligned}\mathrm{plim}_{T\to\infty}\,\widehat{p}_{ij} &= \mathrm{plim}_{T\to\infty}\,\big[\frac{1}{T}\sum_{t=1}^{T}\widehat{\xi}_{i,t-1|T}\big]^{-1}\\ &\mathrm{plim}_{T\to\infty}\,\big[\frac{1}{T}\sum_{t=1}^{T}\widehat{\xi}_{i,t-1|T}Pr(s_t=j|s_{t-1}=i,\mathbf{Y}_T)\big]\\ &= \pi_i^{-1}E(\xi_{i,t-1})p_{ij} = \pi_i^{-1}\pi_i p_{ij} = p_{ij}.\end{aligned}$$

For the asymptotic variance of $\widehat{\boldsymbol{\nu}}_m$ we have

$$\begin{aligned}&\mathrm{var}_a(\widehat{\boldsymbol{\nu}}_m) = \mathrm{plim}_{T\to\infty}\,TE((\widehat{\boldsymbol{\nu}}_m-\boldsymbol{\nu}_m)(\widehat{\boldsymbol{\nu}}_m-\boldsymbol{\nu}_m)^{'})\\ &= \mathrm{plim}_{T\to\infty}\,E\big[\frac{1}{T}\sum_{t=1}^{T}\widehat{\xi}_{mt|T}\big]^{-2}\boldsymbol{\Sigma}_m\,\mathrm{plim}_{T\to\infty}\,E\big[\frac{1}{T}\sum_{t=1}^{T}\sum_{\tau=1}^{T}\mathbf{u}_t\mathbf{u}_\tau^{'}\widehat{\xi}_{mt|T}\widehat{\xi}_{m\tau|T}\big]\boldsymbol{\Sigma}_m^{'}\\ &= \pi_m^{-2}\boldsymbol{\Sigma}_m E(\xi_{mt}^2)\boldsymbol{\Sigma}_m^{'} = \pi_m^{-1}\boldsymbol{\Sigma}_m^2 = \pi_m^{-1}\boldsymbol{\Omega}_m\end{aligned}$$

as $E(\xi_{mt}^2)=\pi_m$. Recall that $E(\boldsymbol{\xi}_t\boldsymbol{\xi}_t^{'}) = \mathbf{D} = \mathrm{diag}(\pi_1,\ldots,\pi_M)$.

Thus $\sqrt{T}(\widehat{\boldsymbol{\nu}}_m-\boldsymbol{\nu}_m)$ is asymptotically normal with zero mean and asymptotic variance $\pi_m^{-1}\boldsymbol{\Omega}_m$ for every $m=1,\ldots,M$ (see Subsection 2.4). One could proceed in this manner for the other estimators. Alternatively, we compute explicitly the asymptotic information matrix in (2.11). For the second derivatives, we have

$$\begin{aligned}\frac{\partial^2 ln\,\boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial\boldsymbol{\nu}_m^{'}\partial\boldsymbol{\nu}_m} &= -\boldsymbol{\Omega}_m^{-1}\\ \frac{\partial^2 ln\,\boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial\boldsymbol{\nu}_m\partial\boldsymbol{\Omega}_m^{-1}} &= \frac{1}{2}\big[(\mathbf{y}_t-\boldsymbol{\nu}_m)\otimes\mathbf{I}_K + \mathbf{I}_K\otimes(\mathbf{y}_t-\boldsymbol{\nu}_m)\big]\\ \frac{\partial^2 ln\,\boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial\boldsymbol{\Omega}_m^{-1}\partial\boldsymbol{\nu}_m^{'}} &= \mathbf{I}_K\otimes(\mathbf{y}_t-\boldsymbol{\nu}_m)^{'}\\ \frac{\partial^2 ln\,\boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial\boldsymbol{\Omega}_m^{-1}\partial\boldsymbol{\Omega}_m^{-1}} &= -\frac{1}{2}(\boldsymbol{\Omega}_m\otimes\boldsymbol{\Omega}_m)\end{aligned}$$

for every $m=1,\ldots,M$. Using the FOC conditions in (3.1) and (3.2) and taking the derivatives of (2.8) with respect to $\boldsymbol{\theta}$, we get the Hessian matrix

$$H(\boldsymbol{\theta}_m) = \left(\frac{\partial^2 ln\,p(\mathbf{Y}_T;\boldsymbol{\lambda})}{\partial\boldsymbol{\theta}_m\partial\boldsymbol{\theta}_m^{'}}\right) = \begin{pmatrix}-\boldsymbol{\Omega}_m^{-1}S_m & \mathbf{0}\\ \mathbf{0} & -\frac{1}{2}(\boldsymbol{\Omega}_m\otimes\boldsymbol{\Omega}_m)S_m\end{pmatrix}$$

where $S_m = \sum_{t=1}^{T} \widehat{\xi}_{mt|T}$ (> 0). Since the above matrix is negative definite for every $m = 1, \ldots, M$, also $H(\boldsymbol{\theta}) = \mathrm{diag}(H(\boldsymbol{\theta}_1), \ldots, H(\boldsymbol{\theta}_M))$ is. Then the full vector $\widehat{\boldsymbol{\theta}}$ (and whence $\widehat{\boldsymbol{\lambda}}$ from Theorem 2.2) maximizes the log likelihood function (for interior values in the parameter space). Furthermore, we have

$$\mathcal{I}_a(\boldsymbol{\theta}_m) = \lim_{T\to\infty} -T^{-1}E(H(\boldsymbol{\theta}_m)) = \begin{pmatrix} \boldsymbol{\Omega}_m^{-1}\pi_m & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}(\boldsymbol{\Omega}_m \otimes \boldsymbol{\Omega}_m)\pi_m \end{pmatrix}.$$

In particular, we see that $\mathrm{var}_a(\widehat{\boldsymbol{\nu}}_m) = \pi_m^{-1}\boldsymbol{\Omega}_m$ and $\mathrm{var}_a(\widehat{\boldsymbol{\Omega}}_m^{-1}) = 2\pi_m^{-1}(\boldsymbol{\Omega}_m^{-1} \otimes \boldsymbol{\Omega}_m^{-1})$. Analogously, taking the derivatives of (3.7) with respect to p_{ij} and taking asymptotic expectations give $\mathcal{I}_a(p_{ij}) = (p_{ij}^{-1}\pi_i)$, hence $\mathrm{var}_a(\widehat{p}_{ij}) = \pi_i^{-1}p_{ij}$, for every $i, j = 1, \ldots, M$.

The explicit expressions for the score vector and Hessian of the log likelihood and the derivation of the asymptotic information matrix are useful in applications, for instance, in the concrete formulation of Wald and Lagrange multiplier tests. For sake of conciseness, we present the Wald test in the case of linear null hypothesis (the general situation can be obtained by using the Jacobian matrix). Suppose that the parameter vector $\boldsymbol{\theta}_m$ is partitioned as $\boldsymbol{\theta}_m = (\boldsymbol{\theta}_{1m}, \boldsymbol{\theta}_{2m})$, where $\boldsymbol{\theta}_{1m} = \boldsymbol{\nu}_m$ and $\boldsymbol{\theta}_{2m} = \mathrm{vec}\,\boldsymbol{\Omega}_m^{-1}$, for $m \in \{1, \ldots, M\}$. Let us consider linear restrictions on $\boldsymbol{\theta}_{im}$, that is, $H_0 : \mathbf{A}_{im}\boldsymbol{\theta}_{im} = \mathbf{0}$ with $\mathrm{rk}\,\mathbf{A}_{im} = r_{im}$, while there is no constraint given for $\boldsymbol{\theta}_{3-i,m}$, for $i = 1, 2$, and $m \in \{1, \ldots, M\}$. Then the Wald statistic is given by

$$\widehat{\boldsymbol{\theta}}_{im}^{'}\mathbf{A}_{im}^{'}\big[\frac{1}{T}\mathbf{A}_{im}\widehat{\boldsymbol{\Sigma}}_{im}\mathbf{A}_{im}^{'}\big]^{-1}\mathbf{A}_{im}\widehat{\boldsymbol{\theta}}_{im} \xrightarrow{d} \chi^2(r_{im})$$

where $\widehat{\boldsymbol{\Sigma}}_{im}$ is the ML estimator of the asymptotic variance matrix of $\widehat{\boldsymbol{\theta}}_{im}$, that is, $\widehat{\boldsymbol{\Sigma}}_{1m} = \pi_m^{-1}\widehat{\boldsymbol{\Omega}}_m$ and $\widehat{\boldsymbol{\Sigma}}_{2m} = 2\pi_m^{-1}(\widehat{\boldsymbol{\Omega}}_m^{-1} \otimes \widehat{\boldsymbol{\Omega}}_m^{-1})$, for any $m \in \{1, \ldots, M\}$.

4.1.4 State-dependent Autoregressive Dynamics

Let now $\mathbf{y}_t$ be a $(K \times 1)$ random vector which follows an M-regime Markov-switching (MS) AR(p) process, with $p > 0$:

$$\mathbf{y}_t + \sum_{i=1}^{p} \boldsymbol{\Phi}_{s_t,i}\mathbf{y}_{t-i} = \boldsymbol{\nu}_{s_t} + \boldsymbol{\Sigma}_{s_t}\mathbf{u}_t \tag{4.1}$$

where $\mathbf{u}_t \sim \mathrm{NID}(\mathbf{0}, \mathbf{I}_K)$, and $\mathbf{\Phi}_{m,i}$ is a $(K \times K)$ matrix for every $m = 1, \ldots, M$. If we set $\mathbf{Y}_{t-1} = (\mathbf{y}_{t-1}^{'}, \ldots, \mathbf{y}_{t-p}^{'})^{'}$, then we have

$$\mathbf{y}_t|_{s_t, \mathbf{Y}_{t-1}} \sim \mathrm{NID}(\boldsymbol{\nu}_{s_t} - \sum_{i=1}^{p} \mathbf{\Phi}_{s_t,i}\mathbf{y}_{t-i}, \mathbf{\Omega}_{s_t}) \tag{4.2}$$

where $\mathbf{\Omega}_{s_t} = \mathbf{\Sigma}_{s_t}^2$, as before. Now the unknown parameter $\boldsymbol{\theta}$ consists of the elements of the intercept vectors $(\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$ and the variance-covariance matrices $(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_M)$ (or, equivalently, their inverses) and the matrices $\mathbf{\Phi}_{m,i}$ for $m = 1, \ldots, M$ and $i = 1, \ldots, p$. The vector $\boldsymbol{\rho}$ is as in Section 3. Hence Model (4.1) implies the estimation of a number of parameters which is equal to $\dim \Lambda_M = M[K + pK^2 + K(K+1)/2 + (M-1)]$ for the most general situation. In the case $s_t = m$, the log of the conditional density in (2.2) becomes

$$\begin{aligned} ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta}) &= ln\, p(\mathbf{y}_t|s_t = m, \mathbf{Y}_{t-1}; \boldsymbol{\theta}) = -\frac{K}{2} ln(2\pi) - \frac{1}{2} ln|\mathbf{\Omega}_m| \\ &- \frac{1}{2}(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p} \mathbf{\Phi}_{m,i}\mathbf{y}_{t-i})^{'} \mathbf{\Omega}_m^{-1} (\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p} \mathbf{\Phi}_{m,i}\mathbf{y}_{t-i}). \end{aligned}$$

The first derivatives are given by

$$\begin{aligned} \frac{\partial ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\nu}_m} &= \mathbf{\Omega}_m^{-1}(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p} \mathbf{\Phi}_{m,i}\mathbf{y}_{t-i}) \\ \frac{\partial ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \mathbf{\Omega}_m^{-1}} &= \frac{1}{2}\mathbf{\Omega}_m - \frac{1}{2}(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p} \mathbf{\Phi}_{m,i}\mathbf{y}_{t-i})(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p} \mathbf{\Phi}_{m,i}\mathbf{y}_{t-i})^{'} \\ \frac{\partial ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \mathbf{\Phi}_{m,i}} &= -\mathbf{\Omega}_m^{-1}(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{j=1}^{p} \mathbf{\Phi}_{m,j}\mathbf{y}_{t-j})\mathbf{y}_{t-i}^{'} \end{aligned}$$

when $s_t = m$ and zero otherwise, for every $m = 1, \ldots, M$ and $i = 1, \ldots, p$. So the FOC conditions in (2.9) give

$$\sum_{t=1}^{T} \mathbf{\Omega}_m^{-1}(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p} \mathbf{\Phi}_{m,i}\mathbf{y}_{t-i})\widehat{\xi}_{mt|T} = \mathbf{0} \tag{4.3}$$

$$\sum_{t=1}^{T} [\frac{1}{2}\mathbf{\Omega}_m - \frac{1}{2}(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p} \mathbf{\Phi}_{m,i}\mathbf{y}_{t-i})(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p} \mathbf{\Phi}_{m,i}\mathbf{y}_{t-i})^{'}]\widehat{\xi}_{mt|T} = \mathbf{0} \tag{4.4}$$

and

$$-\sum_{t=1}^{T}\mathbf{\Omega}_m^{-1}(\mathbf{y}_t-\boldsymbol{\nu}_m+\sum_{j=1}^{p}\mathbf{\Phi}_{m,j}\mathbf{y}_{t-j})\mathbf{y}_{t-i}^{'}\widehat{\xi}_{mt|T}=\mathbf{0} \tag{4.5}$$

for every $m=1,\dots,M$ and $i=1,\dots,p$.

Define the following matrices:

$$\mathbf{\Phi}_m=(\mathbf{\Phi}_{m,1}\cdots\mathbf{\Phi}_{m,p})\qquad \mathbf{A}_m=(\mathbf{A}_m(i,j))\qquad \mathbf{B}_m=(\mathbf{B}_m(i))$$

$$\mathbf{C}_m=(\mathbf{C}_m(i))\qquad S_m=(\sum_{t=1}^{T}\widehat{\xi}_{mt|T})\qquad \mathbf{T}_m=(-\sum_{t=1}^{T}\mathbf{y}_t\widehat{\xi}_{mt|T})$$

where

$$\mathbf{A}_m(i,j)=\sum_{t=1}^{T}\mathbf{y}_{t-i}\mathbf{y}_{t-j}^{'}\widehat{\xi}_{mt|T}\qquad \mathbf{B}_m(i)=\sum_{t=1}^{T}\mathbf{y}_{t-i}^{'}\widehat{\xi}_{mt|T}$$

and

$$\mathbf{C}_m(i)=-\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_{t-i}^{'}\widehat{\xi}_{mt|T}$$

for $m=1,\dots,M$ and $i=1,\dots,p$.

Using these matrices, Equations (4.3) and (4.5) become

$$\mathbf{\Phi}_m\mathbf{B}_m^{'}-\boldsymbol{\nu}_mS_m=\mathbf{T}_m$$

$$\mathbf{\Phi}_m\mathbf{A}_m-\boldsymbol{\nu}_m\mathbf{B}_m=\mathbf{C}_m$$

Multiplying the first (resp. the second) equation by $\mathbf{B}_m$ on the right (resp. by the scalar S_m), we get

$$\mathbf{\Phi}_m\mathbf{B}_m^{'}\mathbf{B}_m-\boldsymbol{\nu}_mS_m\mathbf{B}_m=\mathbf{T}_m\mathbf{B}_m$$

$$\mathbf{\Phi}_mS_m\mathbf{A}_m-\boldsymbol{\nu}_mS_m\mathbf{B}_m=S_m\mathbf{C}_m$$

Subtracting the last two equations gives

$$\mathbf{\Phi}_m\mathbf{X}_m=\mathbf{W}_m$$

where

$$\mathbf{X}_m=\mathbf{B}_m^{'}\mathbf{B}_m-S_m\mathbf{A}_m\qquad \mathbf{W}_m=\mathbf{T}_m\mathbf{B}_m-S_m\mathbf{C}_m$$

for every $m = 1, \dots, M$. Now solving (4.3), (4.4) and (4.5) for $\boldsymbol{\nu}_m$, $\boldsymbol{\Omega}_m$ and $\boldsymbol{\Phi}_m$ and using the above-defined matrices, we get their ML estimates

$$\widehat{\boldsymbol{\nu}}_m = S_m^{-1}(\mathbf{W}_m \mathbf{X}_m^{-1} \mathbf{B}_m^{'} - \mathbf{T}_m) \tag{4.6}$$

$$\widehat{\boldsymbol{\Phi}}_m = \mathbf{W}_m \mathbf{X}_m^{-1} \tag{4.7}$$

and

$$\widehat{\boldsymbol{\Omega}}_m = S_m^{-1}\big[\sum_{t=1}^{T}(\mathbf{y}_t - \widehat{\boldsymbol{\nu}}_m + \sum_{i=1}^{p}\widehat{\boldsymbol{\Phi}}_{m,i}\mathbf{y}_{t-i})(\mathbf{y}_t - \widehat{\boldsymbol{\nu}}_m + \sum_{i=1}^{p}\widehat{\boldsymbol{\Phi}}_{m,i}\mathbf{y}_{t-i})^{'}\widehat{\xi}_{mt|T}\big]. \tag{4.8}$$

Moreover, we obtain

$$\mathbf{X}_m = \big(\mathbf{X}_m(i,j)\big) \qquad \mathbf{W}_m = \big(\mathbf{W}_m(i)\big)$$

where

$$\mathbf{X}_m(i,j) = \big[\sum_{t=1}^{T}\mathbf{y}_{t-i}\widehat{\xi}_{mt|T}\big]\big[\sum_{t=1}^{T}\mathbf{y}_{t-j}^{'}\widehat{\xi}_{mt|T}\big] - \big[\sum_{t=1}^{T}\widehat{\xi}_{mt|T}\big]\big[\sum_{t=1}^{T}\mathbf{y}_{t-i}\mathbf{y}_{t-j}^{'}\widehat{\xi}_{mt|T}\big]$$

and

$$\mathbf{W}_m(i) = -\big[\sum_{t=1}^{T}\mathbf{y}_t\widehat{\xi}_{mt|T}\big]\big[\sum_{t=1}^{T}\mathbf{y}_{t-i}^{'}\widehat{\xi}_{mt|T}\big] + \big[\sum_{t=1}^{T}\widehat{\xi}_{mt|T}\big]\big[\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_{t-i}^{'}\widehat{\xi}_{mt|T}\big]$$

for every $m = 1, \dots, M$ and $i, j = 1, \dots, p$. The ML estimates of p_{ij} are given by (2.10). For the second derivatives, we have

$$\frac{\partial^2 ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\nu}_m^{'} \partial \boldsymbol{\nu}_m} = -\boldsymbol{\Omega}_m^{-1}$$

$$\frac{\partial^2 ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\nu}_m \partial \boldsymbol{\Omega}_m^{-1}} = \frac{1}{2}\big[(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p}\boldsymbol{\Phi}_{m,i}\mathbf{y}_{t-i}) \otimes \mathbf{I}_K + \mathbf{I}_K \otimes (\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p}\boldsymbol{\Phi}_{m,i}\mathbf{y}_{t-i})\big]$$

$$\frac{\partial^2 ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\nu}_m \partial \boldsymbol{\Phi}_{m,i}^{'}} = \boldsymbol{\Omega}_m^{-1} \otimes \mathbf{y}_{t-i}$$

$$\frac{\partial^2 ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\Omega}_m^{-1} \partial \boldsymbol{\nu}_m^{'}} = \mathbf{I}_K \otimes (\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{i=1}^{p}\boldsymbol{\Phi}_{m,i}\mathbf{y}_{t-i})^{'}$$

$$\frac{\partial^2 ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\Omega}_m^{-1} \partial \boldsymbol{\Omega}_m^{-1}} = -\frac{1}{2}(\boldsymbol{\Omega}_m \otimes \boldsymbol{\Omega}_m)$$

$$\frac{\partial^2 ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\Omega}_m^{-1} \partial \boldsymbol{\Phi}_{m,i}^{'}} = -\mathbf{I}_K \otimes \big[\mathbf{y}_{t-i}(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{j=1}^{p} \boldsymbol{\Phi}_{m,j}\mathbf{y}_{t-j})^{'}\big]$$

$$\frac{\partial^2 ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\Phi}_{m,i} \partial \boldsymbol{\nu}_m^{'}} = \boldsymbol{\Omega}_m^{-1} \otimes \mathbf{y}_{t-i}^{'}$$

$$\frac{\partial^2 ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\Phi}_{m,i} \partial \boldsymbol{\Omega}_m^{-1}} = -\frac{1}{2}\big[(\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{j=1}^{p} \boldsymbol{\Phi}_{m,j}\mathbf{y}_{t-j}) \otimes \mathbf{I}_K + \mathbf{I}_K \otimes (\mathbf{y}_t - \boldsymbol{\nu}_m + \sum_{j=1}^{p} \boldsymbol{\Phi}_{m,j}\mathbf{y}_{t-j})\big]$$
$$(\mathbf{y}_{t-i}^{'} \otimes \mathbf{I}_K)$$

$$\frac{\partial^2 ln\, \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\Phi}_{m,i} \partial \boldsymbol{\Phi}_{m,j}^{'}} = -\boldsymbol{\Omega}_m^{-1} \otimes (\mathbf{y}_{t-j}\mathbf{y}_{t-i}^{'})$$

for every $m = 1, \ldots, M$ and $i, j = 1, \ldots, p$. Using the FOC conditions in (4.3)–(4.5) and taking the second derivatives with respect to $\boldsymbol{\theta}$, we get the Hessian matrix (up to re–ordering)

$$H(\boldsymbol{\theta}_m) = \begin{pmatrix} -\frac{1}{2}(\boldsymbol{\Omega}_m \otimes \boldsymbol{\Omega}_m)S_m & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{\Omega}_m^{-1} \otimes \mathbf{R}_m \end{pmatrix}$$

where

$$\mathbf{R}_m = \begin{pmatrix} S_m & -\mathbf{B}_m \\ -\mathbf{B}_m^{'} & \mathbf{A}_m \end{pmatrix}$$

is positive definite, hence $H(\boldsymbol{\theta}_m)$ is negative definite, for every $m = 1, \ldots, M$. In fact, we have

$$\mathbf{R}_m = \sum_{t=1}^{T} \widehat{\xi}_{mt|T} \mathbf{W}\mathbf{W}^{'}$$

where $\mathbf{W} = (1 \quad -\mathbf{y}_{t-1}^{'} \cdots - \mathbf{y}_{t-p}^{'})^{'}$. Finally, we have

$$\mathcal{I}_a(\boldsymbol{\theta}_m) = \begin{pmatrix} \frac{1}{2}(\boldsymbol{\Omega}_m \otimes \boldsymbol{\Omega}_m)\pi_m & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_m^{-1} \otimes E(\mathbf{R}_m) \end{pmatrix}.$$

For $\mathcal{I}_a(p_{ij})$ see the previous section. The derivation of the asymptotic information matrix is useful in the explicit formulation of Wald and Lagrange multiplier tests, as shown in the previous section for the case of Wald test with linear null hypothesis.

4.1.5 State-dependent Multivariate ARCH Models

In this section we discuss the parameter estimation for a multivariate Markov-switching model with ARCH disturbances. Estimation of ARCH models has been considered by several authors as, for example, Weiss (1986), Bollerslev, Engle and Nelson (1993), and Hamilton (1994), Chp.21, p.657.

Let us consider the following M–regime Markov-switching regression model

$$\mathbf{y}_t = X_t \boldsymbol{\beta}_{s_t} + \mathbf{u}_t \tag{5.1}$$

where $\mathbf{y}_t$ is $(K \times 1)$, $\boldsymbol{\beta}_{s_t}$ is a $(n \times 1)$ parameter vector, and X_t denotes a $(K \times n)$ matrix of predetermined explanatory variables, which could include lagged values of $\mathbf{y}$. The error term $\mathbf{u}_t$ is a $(K \times 1)$ random variable which follows a stationary ARCH process of order p with $p > 0$. To specify the ARCH equation, let I_{t-1} be the information set containing information about the process up to and including time $t-1$. Then assume $E(\mathbf{u}_t | I_{t-1}) = \mathbf{0}$. Furthermore, the conditional variance–covariance matrix of $\mathbf{u}_t$ is given by

$$\boldsymbol{\Omega}_{s_t,t} = \boldsymbol{\Sigma}_{s_t} \boldsymbol{\Sigma}_{s_t}^{'} + \sum_{i=1}^{p} \boldsymbol{\Lambda}_{s_t,i} \mathbf{u}_{t-i} \mathbf{u}_{t-i}^{'} \boldsymbol{\Lambda}_{s_t,i}^{'}. \tag{5.2}$$

This formulation is taken from Bollerslev, Engle and Nelson (1993), Formula (6.5), p.50. It has the advantage that $\boldsymbol{\Omega}_{s_t,t}$ is guaranteed to be positive definite a.s. for all t. The process in (5.1) is the multivariate version of that considered in Hamilton (1994) , Formula [21.1.17], p.660. In this book the scalar conditional variance $\boldsymbol{\Omega}_t = h_t$ evolves according to $h_t = \xi + \sum_{i=1}^{p} \alpha_i u_{t-i}^2$, where ξ and α_i are scalar parameters and the error term u_t is scalar, too. If $\mathbf{u}_t|_{I_{t-1}} \sim \text{NID}(\mathbf{0}, \boldsymbol{\Omega}_{s_t,t})$ with $\mathbf{u}_t$ independent of s_t, X_t and I_{t-1}, then the conditional distribution of $\mathbf{y}_t$ is Gaussian with mean $X_t \boldsymbol{\beta}_{s_t}$ and conditional variance $\boldsymbol{\Omega}_{s_t,t}$, that is, we have

$$\begin{aligned} p(\mathbf{y}_t | s_t, X_t, I_{t-1}; \boldsymbol{\theta}) =& (2\pi)^{-K/2} |\boldsymbol{\Omega}_{s_t,t}|^{-1/2} \\ & \times \exp\left\{ -\frac{(\mathbf{y}_t - X_t \boldsymbol{\beta}_{s_t})' \boldsymbol{\Omega}_{s_t,t}^{-1} (\mathbf{y}_t - X_t \boldsymbol{\beta}_{s_t})}{2} \right\} \end{aligned}$$

where $\boldsymbol{\Omega}_{s_t,t}$ follows (5.2). In short, we denote the above dynamic process by VARCH(p) with

$p > 0$. The unknown parameter $\boldsymbol{\theta}$ consists of the $(n \times 1)$ vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M$ and the $(K \times K)$ matrices $\boldsymbol{\Sigma}_m$ and $\boldsymbol{\Lambda}_{m,i}$ for $m = 1, \ldots, M$ and $i = 1, \ldots, p$. The vector $\boldsymbol{\rho}$ is as in Section 3. Hence Model (5.1) implies the estimation of a number of parameters which is equal to $M[n + (p+1)K^2 + (M-1)]$ for the most general situation. In the case $s_t = m$, the log of the conditional density of $\mathbf{y}_t$ becomes

$$\begin{aligned}\ln \boldsymbol{\eta}_{mt}(\boldsymbol{\theta}) &= \ln p(\mathbf{y}_t | s_t = m, X_t, I_{t-1}; \boldsymbol{\theta}) \\ &= -\frac{K}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Omega}_{mt}| - \frac{1}{2}(\mathbf{y}_t - X_t\boldsymbol{\beta}_m)^{'}\boldsymbol{\Omega}_{mt}^{-1}(\mathbf{y}_t - X_t\boldsymbol{\beta}_m),\end{aligned}$$

where

$$\boldsymbol{\Omega}_{mt} = \boldsymbol{\Sigma}_m\boldsymbol{\Sigma}_m^{'} + \sum_{i=1}^{p}\boldsymbol{\Lambda}_{m,i}(\mathbf{y}_{t-i} - X_{t-i}\boldsymbol{\beta}_m)(\mathbf{y}_{t-i} - X_{t-i}\boldsymbol{\beta}_m)^{'}\boldsymbol{\Lambda}_{m,i}^{'}.$$

The first derivatives are given by

$$\frac{\partial \ln \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_m} = X_t^{'}\boldsymbol{\Omega}_{mt}^{-1}\mathbf{u}_t - \sum_{i=1}^{p}\left[X_{t-i}^{'}\boldsymbol{\Lambda}_{m,i}^{'}\left(\boldsymbol{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\boldsymbol{\Omega}_{mt}^{-1} - \boldsymbol{\Omega}_{mt}^{-1}\right)\boldsymbol{\Lambda}_{m,i}\mathbf{u}_{t-i}\right] \tag{5.3}$$

$$\frac{\partial \ln \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_m} = \frac{1}{2}\left\{\left(\boldsymbol{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\boldsymbol{\Omega}_{mt}^{-1} - \boldsymbol{\Omega}_{mt}^{-1}\right)\boldsymbol{\Sigma}_m^{'} + \boldsymbol{\Sigma}_m\left(\boldsymbol{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\boldsymbol{\Omega}_{mt}^{-1} - \boldsymbol{\Omega}_{mt}^{-1}\right)\right\} \tag{5.4}$$

and

$$\begin{aligned}\frac{\partial \ln \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\Lambda}_{m,i}} =& \frac{1}{2}\left(\boldsymbol{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\boldsymbol{\Omega}_{mt}^{-1} - \boldsymbol{\Omega}_{mt}^{-1}\right)\mathbf{u}_{t-i}\mathbf{u}_{t-i}^{'}\boldsymbol{\Lambda}_{m,i}^{'} \\ &+ \frac{1}{2}\boldsymbol{\Lambda}_{m,i}\mathbf{u}_{t-i}\mathbf{u}_{t-i}^{'}\left(\boldsymbol{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\boldsymbol{\Omega}_{mt}^{-1} - \boldsymbol{\Omega}_{mt}^{-1}\right)\end{aligned} \tag{5.5}$$

when $s_t = m$ and zero otherwise, for every $m = 1, \ldots, M$ and $i = 1, \ldots, p$. In the univariate case ($K = 1$), setting $M = 1$, $\boldsymbol{\Omega}_{mt} = h_t$, $\boldsymbol{\Lambda}_{m,i} = \sqrt{\alpha_i}$ and $X_t = \mathbf{x}_t^{'}$ (which is a $(1 \times n)$ vector), the first derivative in (5.3) becomes that obtained in Hamilton (1994), Formula [21.1.21], p.661, that is

$$\frac{\partial \ln \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_m} = \frac{u_t}{h_t}\mathbf{x}_t - \left(\frac{u_t^2}{h_t^2} - \frac{1}{h_t}\right)\sum_{i=1}^{p}\alpha_i u_{t-i}\mathbf{x}_{t-i}.$$

Of course, the first derivatives in (5.4) and (5.5) are slightly different from those listed in that page of the quoted book due to the definition of the conditional variance. In our univariate case, setting $M = 1$ and $\mathbf{\Sigma}_m = \sqrt{\xi}$, we get

$$\frac{\partial \ln \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \mathbf{\Sigma}_m} = \left(\frac{u_t^2}{h_t^2} - \frac{1}{h_t}\right)\sqrt{\xi} \qquad \frac{\partial \ln \boldsymbol{\eta}_{mt}(\boldsymbol{\theta})}{\partial \mathbf{\Lambda}_{m,i}} = \left(\frac{u_t^2}{h_t^2} - \frac{1}{h_t}\right) u_{t-i}^2\sqrt{\alpha_i}.$$

So the FOC conditions in (2.9) give

$$(5.6) \qquad \sum_{t=1}^{T} X_t^{'}\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\widehat{\xi}_{mt|T} = \sum_{t=1}^{T}\sum_{i=1}^{p}\left[X_{t-i}^{'}\mathbf{\Lambda}_{m,i}^{'}\left(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\right)\mathbf{\Lambda}_{m,i}\mathbf{u}_{t-i}\right]\widehat{\xi}_{mt|T}$$

$$(5,7) \qquad \sum_{t=1}^{T}\left(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\right)\mathbf{\Sigma}_m^{'}\widehat{\xi}_{mt|T} + \sum_{t=1}^{T}\mathbf{\Sigma}_m\left(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\right)\widehat{\xi}_{mt|T} = \mathbf{0}$$

and

$$(5.8) \qquad \begin{aligned} &\sum_{t=1}^{T}\left(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\right)\mathbf{u}_{t-i}\mathbf{u}_{t-i}^{'}\mathbf{\Lambda}_{m,i}^{'}\widehat{\xi}_{mt|T} \\ &\quad + \sum_{t=1}^{T}\mathbf{\Lambda}_{m,i}\mathbf{u}_{t-i}\mathbf{u}_{t-i}^{'}\left(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\right)\widehat{\xi}_{mt|T} = \mathbf{0} \end{aligned}$$

for every $m = 1, \dots, M$ and $i = 1, \dots, p$. Using the relations

$$\widehat{\mathbf{u}}_t = \mathbf{y}_t - X_t\widehat{\boldsymbol{\beta}}_m = X_t\boldsymbol{\beta}_m + \mathbf{u}_t - X_t\widehat{\boldsymbol{\beta}}_m,$$

we get from (5.6)

$$\widehat{\boldsymbol{\beta}}_m = \boldsymbol{\beta}_m + \mathcal{I}_{mT}^{-1}\mathcal{T}_{mT}$$

where

$$\mathcal{I}_{mT} = \sum_{t=1}^{T} X_t^{'}\mathbf{\Omega}_{mt}^{-1}X_t\widehat{\xi}_{mt|T}$$

and

$$\mathcal{T}_{mT} = \sum_{t=1}^{T} X_t^{'}\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\widehat{\xi}_{mt|T} - \sum_{t=1}^{T}\sum_{i=1}^{p}\left[X_{t-i}^{'}\mathbf{\Lambda}_{m,i}^{'}\left(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\right)\mathbf{\Lambda}_{m,i}\mathbf{u}_{t-i}\right]\widehat{\xi}_{mt|T}.$$

Then we have

$$\text{plim}_{T\to\infty}\widehat{\boldsymbol{\beta}}_m = \boldsymbol{\beta}_m + \left[\text{plim}_{T\to\infty}\frac{1}{T}\mathfrak{I}_{mT}\right]^{-1}\left[\text{plim}_{T\to\infty}\frac{1}{T}\mathfrak{T}_{mT}\right] = \boldsymbol{\beta}_m$$

for every $m = 1, \ldots, M$ as

$$\mathfrak{I}_m = \text{plim}_{T\to\infty}\frac{1}{T}\mathfrak{I}_{mT} = E(X_t'\boldsymbol{\Omega}_{mt}^{-1}X_t)\pi_m < +\infty \qquad\qquad \text{plim}_{T\to\infty}\frac{1}{T}\mathfrak{T}_{mT} = \mathbf{0}.$$

This gives the consistency of the ML estimators $\widehat{\boldsymbol{\beta}}_m$ for every $m = 1, \ldots, M$. For the conditional asymptotic variance of $\widehat{\boldsymbol{\beta}}_m$ we have

$$\begin{aligned}\text{var}_a(\widehat{\boldsymbol{\beta}}_m) &= \text{plim}_{T\to\infty} TE((\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m)(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m)') \\ &= \text{plim}_{T\to\infty} TE\left(\mathfrak{I}_{mT}^{-1}\mathfrak{T}_{mT}\mathfrak{T}_{mT}'\mathfrak{I}_{mT}^{-1}\right) = \mathfrak{I}_m^{-1}S_m\mathfrak{I}_m^{-1},\end{aligned}$$

where

$$S_m = \text{plim}_{T\to\infty}\frac{1}{T}\mathfrak{T}_{mT}\mathfrak{T}_{mT}'.$$

For example, in the univariate non state–dependent case with the above considered correspondences, we get

$$\widehat{\boldsymbol{\beta}}_m = \boldsymbol{\beta}_m + \big(\sum_{t=1}^{T}\frac{1}{h_t}\mathbf{x}_t\mathbf{x}_t'\big)^{-1}\big[\sum_{t=1}^{T}\frac{u_t}{h_t}\mathbf{x}_t - \sum_{t=1}^{T}\big(\frac{u_t^2}{h_t^2} - \frac{1}{h_t}\big)\sum_{i=1}^{p}\alpha_i u_{t-i}\mathbf{x}_{t-i}\big].$$

So the conditional asymptotic variance of $\widehat{\boldsymbol{\beta}}_m$ becomes

$$\text{var}_a(\widehat{\boldsymbol{\beta}}_m) = \big(1 + 2\sum_{i=1}^{p}\alpha_i^2\big)\mathfrak{I}_m^{-1}.$$

In particular, if $\alpha_i = 0$ for every $i = 1, \ldots, p$, that is, the conditional variance $\boldsymbol{\Omega}_t = h_t = \xi$ is time independent, we obtain

$$\text{var}_a(\widehat{\boldsymbol{\beta}}_m) = \xi[E(\mathbf{x}_t\mathbf{x}_t')]^{-1}$$

as expected.

From (5.7) we get

$$\sum_{t=1}^{T}[(\mathbf{\Sigma}_m \otimes \mathbf{I}_K) + T_{K,K}(\mathbf{\Sigma}_m \otimes \mathbf{I}_K)]\,\mathrm{vec}(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1})\widehat{\xi}_{mt|T} = \mathbf{0}$$

where $T_{K,K}$ is the usual $(K^2 \times K^2)$ permutation matrix which transforms $\mathrm{vec}(A)$ into $\mathrm{vec}(A^{'})$ for a given $(K \times K)$ matrix A. Here we have used the property $T_{K,K}(\mathbf{\Sigma}_m \otimes \mathbf{I}_K) = (\mathbf{I}_K \otimes \mathbf{\Sigma}_m)T_{K,K}$. Thus we obtain

$$(T_{K,K} + \mathbf{I}_{K^2})(\mathbf{\Sigma}_m \otimes \mathbf{I}_K)\sum_{t=1}^{T}\mathrm{vec}(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1})\widehat{\xi}_{mt|T} = \mathbf{0}$$

which gives another form of (5.7). In similar manner the FOC in (5.8) can be expressed as

$$(T_{K,K} + \mathbf{I}_{K^2})\big\{\sum_{t=1}^{T}\Big[(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1})\mathbf{u}_{t-i}\mathbf{u}_{t-i}^{'}\Big] \otimes \mathbf{I}_K\widehat{\xi}_{mt|T}\big\}\,\mathrm{vec}(\mathbf{\Lambda}_{mi}) = \mathbf{0}$$

for every $i = 1, \ldots, p$.

To end the section we observe that the conditional asymptotic variances of $\widehat{\mathbf{\Sigma}}_m$ and $\widehat{\mathbf{\Lambda}}_{m,i}$ can be obtained by making use of the matrices $[E(R_{mt}R_{mt}^{'})]\pi_m$ and $[E(T_{mt}T_{mt}^{'})]\pi_m$, respectively, where

$$\begin{aligned} R_{mt} = \frac{1}{2}\big[&\mathbf{I}_K \otimes \Big(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\Big)\,\mathrm{vec}(\mathbf{\Sigma}_m^{'}) \\ &+ \Big(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\Big) \otimes \mathbf{I}_K\,\mathrm{vec}(\mathbf{\Sigma}_m)\big] \end{aligned}$$

and

$$\begin{aligned} T_{mt} =& \frac{1}{2}\big\{\mathbf{I}_K \otimes \big[\Big(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\Big)\,\mathbf{u}_{t-i}\mathbf{u}_{t-i}^{'}\big]\,\mathrm{vec}(\mathbf{\Lambda}_{m,i}^{'}) \\ &+ \big[\Big(\mathbf{\Omega}_{mt}^{-1}\mathbf{u}_t\mathbf{u}_t^{'}\mathbf{\Omega}_{mt}^{-1} - \mathbf{\Omega}_{mt}^{-1}\Big)\,\mathbf{u}_{t-i}\mathbf{u}_{t-i}^{'}\big] \otimes \mathbf{I}_K\,\mathrm{vec}(\mathbf{\Lambda}_{m,i})\big\}. \end{aligned}$$

4.1.6 Conclusion

In this paper we derive maximum likelihood estimators and their asymptotic variance-covariance matrix for Markov switching vector autoregressive models. Our matrix formulae are very simple and different in some cases to those obtained in the known literature. This in turn permits to use easily the classical specification testing procedures. Our discussion is based on the fundamental work developed by Hamilton, Kim, and Krolzig, and is related with the analytical derivatives obtained by Gable, Van Norden and Vigfusson for univariate Markov-switching models without lags. As remarked by these last authors, estimation for Markov switching models is usually done using numerical techniques that approximate the derivative by the change in the likelihood function for small changes in the parameter vector. This is not especially efficient, however, as such techniques typically require a lot of numerical evaluations. Using analytical gradients the number of calculations can be greatly reduced, and this in turn considerably speeds up MLE with no loss of accuracy. This gives a strong motivation for searching explicit formulae of ML estimates and their asymptotic variance-covariance matrix. Finally, in the context of multivariate ARCH models with changes in regime we show how analytic derivatives of the log likelihood can be successfully employed for estimation purposes.

References

Bollerslev, T., Engle, R.F., and Nelson, D.B. 1994, ARCH models, in: *Handbook of Econometrics* (Engle, R.F., and McFadden, D.L., eds.), Vol.**4**, Elsevier Science Publishers, Amsterdam: North-Holland.

Fackler, P.L. 2005, *Notes on Matrix Calculus*, North Carolina State Univ., U.S.

Gable, J., Van Norden, S., and Vigfusson, R. 1997, Analytical derivatives for Markov switching models, *Computational Economics* **10**: 187-194.

Greene, W.H. 2008, *Econometric Analysis*, sixth edition. Pearson International Edition, Upper Saddle River, N.J.

Hamilton, J.D. 1990, Analysis of Time Series Subject to Changes in Regime, *Journal of Econometrics* **45**: 39-70.

Hamilton, J.D. 1993, Estimation, Inference and Forecasting of Time Series Subject to Changes in Regime, in: *Handbook of Statistics* (Maddala, G.S., Rao, C.R., and Vinod, H.D., eds.), Vol.11, Elsevier Science Publishers, Amsterdam: North-Holland.

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

Kim, C.J. 1994, Dynamic Linear Models with Markov-Switching, *Journal of Econometrics* **60**: 1-22.

Krolzig, H.M. 1997, *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference and Application to Business Cycle Analysis*, Springer Verlag, Berlin, Heidelberg, New York.

Lütkepohl, H. 1991, *Introduction to Multiple Time Series Analysis*, Springer Verlag, Berlin, Heidelberg, New York.

Metaxoglou, K., and Smith, A. 2007, Maximum Likelihood Estimation of VARMA models using a state-space EM algorithm, *Journal of Time Series Analysis*, **28** (5): 666-685.

Petersen, K.B., and Pedersen, M.S. 2003, *The Matrix Cookbook*, available in: http://matrixcookbook.com

Weiss, A.A. 1986, Asymptotic theory for ARCH models: Estimation and Testing, *Econometric Theory* **2** (1): 107-131.

4.2 Spectral Density of Regime Switching VAR Models

Abstract. *We consider multivariate AR models subject to Markov Switching and present results about the computation of their autocovariance function. We derive a close-form formula for the spectral density function of Markov Switching VAR models, and give stable VARMA representations of such processes. An example is proposed in order to illustrate the behavior of spectral density functions and the feasibility of the approach. Finally, we investigate whether S&P500 stock market returns suffer of structural change rather then long memory via spectral analysis.[JEL Classification: C01, C32, C50]*

Keywords: Markov Switching, VAR model, Spectral density, Stable representation, Long memory, financial returns.

4.2.1 Introduction

This paper studies multivariate autoregressive models which are subject to change in regime, described as an outcome of an unobserved Markov chain. Markov switching models play an important role in many financial and economic studies and constitute an useful method to model uncertainty preserving the tractabililty of linear framework. Moreover, in time series analysis, the primary interest is often to study the periodic behavior of the data and a useful tool is the Fourier transform. The frequency content of a time series can be analyzed through the spectral density function, which results from its autocovariance function values. Hence, we derive close-form formulae for the spectral representations of Markov switching VAR processes. Our results are related to the work of Krolzig (1997) in terms of state space representation and stable representation and to the paper of Pataracchia (2011) where a different Markovian representation has been considered.

4.2.2 Markov-switching VAR(0) process

Let us consider the model

$$\mathbf{y}_t = \nu_{s_t} + \Sigma_{s_t}\mathbf{u_t} \tag{4.1}$$

where $\mathbf{u}_t \sim IID(\mathbf{0}, \mathbf{I_k})$, $\mathbf{y_t}$, ν_{s_t} and $\mathbf{u_t}$ are $K \times 1$, Σ_{s_t} is $K \times K$ and (s_t) follows an M-state (irreducible and ergodic) Markov chain. Let $\mathbf{P} = (p_{ij})_{i,j=1,\dots,M}$ be the transition matrix of the chain, where $p_{ij} = Pr(s_t = j|s_{t-1} = i)$. Ergodicity implies the existence of a stationary vector of probabilities $\pi = (\pi_1 \dots \pi_M)^{'}$ satisfying $\pi = \mathbf{P}^{'}\pi$ and $\mathbf{i}_M^{'}\pi = 1$, where $\mathbf{i}_M$ denotes the $(M \times 1)$ vector of ones. Irreducibility implies that $\pi_m > 0$ for $m = 1, \dots, M$, meaning that all unobservable states are possibile. An useful representation for (s_t) is obtained by letting $\xi_\mathbf{t}$ denote a random $(M \times 1)$ vector whose mth element is equal to unity if $s_t = m$ and zero otherwise. Then the Markov chain follows a VAR(1) process

$$\xi_\mathbf{t} = \mathbf{P}^{'}\xi_\mathbf{t-1} + \mathbf{v_t}$$

where $\mathbf{v_t} = \xi_t - E(\xi_\mathbf{t}|\xi_{t-1})$ is a zero mean martingale difference sequence.

Consequently, we have the following standard properties $(h > 0)$:

$$E(\xi_\mathbf{t}) = \pi \qquad E(\xi_\mathbf{t}\xi_\mathbf{t}^{'}) = \mathbf{D} = \text{diag}(\pi_1 \dots \pi_M)$$
$$E(\xi_\mathbf{t}\xi_\mathbf{t+h}^{'}) = \mathbf{DP^h} \qquad \mathbf{v_t} \sim IID(\mathbf{0}, \mathbf{D} - \mathbf{P}^{'}\mathbf{DP})$$

Define $\Lambda = (\nu_1 \dots \nu_M)$ and $\Sigma = (\Sigma_1 \dots \Sigma_M)$. We get a first state space representation of (1)

$$\begin{cases} \mathbf{y}_t = \Lambda\xi_\mathbf{t} + \Sigma\xi_t \otimes \mathbf{I}_K)\mathbf{u_t} \\ \xi_\mathbf{t} = \mathbf{P}^{'}\xi_\mathbf{t-1} + \mathbf{v_t} \end{cases} \tag{4.2}$$

In fact, for $s_t = m$, $\xi_t = \mathbf{e_m}$ the mth column of the identity matrix $\mathbf{I}_M$. So we get

$$\mathbf{y}_t = \begin{pmatrix} \nu_1 & \dots & \nu_M \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \Sigma_1 & \dots & \Sigma_M \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{I}_K \\ \vdots \\ \mathbf{0} \end{pmatrix} \mathbf{u_t}$$

$$= \nu_m + \Sigma_m \mathbf{I}_K \mathbf{u_t} = \nu_m + \Sigma_m \mathbf{u_t}.$$

The transition equation in (2) differs from a stable linear VAR(1) process by the fact that one eigenvalue of $\mathbf{P}^{'}$ is equal to one, and the covariance matrix is singular due to the adding-up restriction. For analytical purposes, a slightly different formulation of the transition equation in (2) is more useful, where the identity $\mathbf{i}_M^{'}\xi_t = 1$ is eliminated. See Krolzig (1997), Chp.3. This procedure alters the state-space representation by using a new $(M-1)$-dimensional state vector

$$\delta_{\mathbf{t}} = \begin{pmatrix} \xi_{1,t} - \pi_1 \\ \vdots \\ \xi_{M-1,t} - \pi_{M-1} \end{pmatrix}.$$

The transition matrix $\mathbf{F}$ associated with $\delta_{\mathbf{t}}$ is given by

$$\mathbf{F} = \begin{pmatrix} p_{1,1} - p_{M,1} & \dots & p_{M-1,1} - p_{M,1} \\ \vdots & & \vdots \\ p_{1,M-1} - p_{M,M-1} & \dots & p_{M-1,M-1} - p_{M,M-1} \end{pmatrix}.$$

The eigenvalues of $\mathbf{F}$ are less than 1 in absolute value. Here the relations

$$\xi_{M,t} = 1 - \sum_{m=1}^{M-1} \xi_{mt} \qquad \pi_M = 1 - \sum_{m=1}^{M-1} \pi_m$$

have been used. Then we have

$$\xi_{\mathbf{t}} - \pi = \mathbf{P}^{'}(\xi_{\mathbf{t-1}} - \pi) + \mathbf{v_t}$$

hence

$$\delta_{\mathbf{t}} = \mathbf{F}\,\delta_{\mathbf{t-1}} + \mathbf{w_t}$$

where

$$\mathbf{w}_t = (\mathbf{I}_{M-1} \qquad - \mathbf{i}_{M-1})v_t.$$

This gives a second (unrestricted) state-space representation

$$\mathbf{y_t} = \Lambda\pi + \Lambda(\xi_{\mathbf{t}} - \pi) + \Sigma((\xi_{\mathbf{t}} - \pi)\otimes \mathbf{I}_K)\mathbf{u_t} + \Sigma(\pi \otimes \mathbf{I}_K)\mathbf{u_t}$$

hence

$$\begin{cases} \mathbf{y_t} = \Lambda\pi + \widetilde{\Lambda}\delta_{\mathbf{t}} + \widetilde{\Sigma}(\delta_t \otimes \mathbf{I}_K)\mathbf{u_t} + \Sigma(\pi\otimes \mathbf{I}_K)\mathbf{u_t} \\ \delta_{\mathbf{t}} = \mathbf{F}\,\delta_{\mathbf{t-1}} + \mathbf{w_t} \end{cases} \tag{4.3}$$

where

$$\widetilde{\Lambda} = (\nu_1 - \nu_M \dots \nu_{M-1} - \nu_M) \qquad\qquad \widetilde{\Sigma} = (\Sigma_1 - \Sigma_M \dots \Sigma_{M-1} - \Sigma_M).$$

We then have the following standard properties:

$$E(\delta_{\mathbf{t}}) = \mathbf{0} \qquad\qquad E(\delta_{\mathbf{t}}\delta_{\mathbf{t}}^{'}) = \widetilde{\mathbf{D}}$$

$$E(\delta_{\mathbf{t}}\delta_{\mathbf{t+h}}^{'}) = \widetilde{\mathbf{D}}(\mathbf{F}^{'})^{h}, \quad h > 0 \qquad \mathbf{w_t} \sim IID(\mathbf{0}, \widetilde{\mathbf{D}} - \mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}^{'})$$

where

$$\widetilde{\mathbf{D}} = \begin{pmatrix} \pi_1(1-\pi_1) & \dots & -\pi_1\pi_{M-1} \\ \vdots & & \vdots \\ -\pi_{M-1}\pi_1 & \dots & \pi_{M-1}(1-\pi_{M-1}) \end{pmatrix}.$$

The autocovariance function of the process $(\mathbf{y_t})$ in (3) is given by

$$\Gamma_{\mathbf{y}}(0) = \widetilde{\Lambda}\widetilde{\mathbf{D}}\widetilde{\Lambda}^{'} + \widetilde{\Sigma}(\widetilde{\mathbf{D}}\otimes \mathbf{I}_K)\widetilde{\Sigma}^{'} + \Sigma((\mathbf{DP}_\infty)\otimes \mathbf{I}_K)\Sigma^{'}$$

$$\Gamma_{\mathbf{y}}(h) = \widetilde{\Lambda}\mathbf{F}^h\widetilde{\mathbf{D}}\widetilde{\Lambda}^{'}, \qquad h > 0$$

where $\mathbf{DP}_\infty = \pi\pi^{'}$ and $\mathbf{P}_\infty = \lim_n \mathbf{P}^n = \mathbf{i}_M\pi^{'}$. For the proof see Cavicchioli (2013), Th.3.3.

The multivariate spectral matrix describes the spectral density functions of each element of the state vector in the diagonal terms. The off-diagonal terms are defined cross spectral density functions and they are typically complex numbers. Here we are only interested in the diagonal terms. Therefore, we can compute them, without loss of generality, considering the summation

$$F_{\mathbf{y}}(\omega) = \sum_{h=-\infty}^{+\infty} \Gamma_{\mathbf{y}}(|h|)e^{-i\omega h}$$

where the frequency ω belongs to $[-\pi, \pi]$. See also Pataracchia (2011) where a different spectral representation was obtained. Since the spectral radius $\rho(\mathbf{F})$ of $\mathbf{F}$ is less than 1, the spectral density matrix of the process $(\mathbf{y_t})$ in (3) is given by

$$F_{\mathbf{y}}(\omega) = Q + 2\widetilde{\Lambda}\mathbf{F}\Re e\{(\mathbf{I}_{M-1}e^{i\omega} - \mathbf{F})^{-1}\}\widetilde{\mathbf{D}}\widetilde{\Lambda}'$$

where $\Re e$ denotes the real part of the complex matrix $(\mathbf{I}_{M-1}e^{i\omega} - \mathbf{F})^{-1}$, and

$$Q = \widetilde{\Lambda}\widetilde{\mathbf{D}}\widetilde{\Lambda}' + \widetilde{\Sigma}(\widetilde{\mathbf{D}} \otimes \mathbf{I}_K)\widetilde{\Sigma}' + \Sigma((\mathbf{D}\mathbf{P}_\infty) \otimes \mathbf{I}_K)\Sigma'.$$

An alternative approach to the same problem is based on a stable representation of (3). Set $\mu_{\mathbf{y}} = \Lambda\pi$. From (3) we get

$$\delta_t = F(L)^{-1}\mathbf{w_t}$$

where $F(L) = \mathbf{I}_{M-1} - \mathbf{F}\,L$ (here L is the lag operator). Substituting this relation into the measurement equation in (3) yields

$$|F(L)|(\mathbf{y}_t - \mu_{\mathbf{y}}) = \widetilde{\Lambda}F(L)^*\mathbf{w_t} + \widetilde{\Sigma}(F(L)^*\mathbf{w_t} \otimes \mathbf{I}_K)\mathbf{u_t} + |F(L)|\Sigma(\pi \otimes \mathbf{I}_K)\mathbf{u_t}$$

where $F(L)^*$denotes the adjoint matrix of $F(L)$ and $|F(L)|$ is the deteminant of $F(L)$. Thus we get a stable VARMA(p^*, q^*) representation of the process $(\mathbf{y_t})$ in (3)

$$\phi(L)(\mathbf{y_t} - \mu_{\mathbf{y}}) = \theta(L)\epsilon_t \tag{4.4}$$

where $p^* = q^* \leq M-1$, $\phi(L) = |F(L)|$ is scalar and

$$\theta(L) = (\widetilde{\Lambda}F(L)^* \qquad \widetilde{\Sigma}(F(L)^* \otimes \mathbf{I}_K) \qquad |F(L)|\mathbf{I}_K).$$

See Cavicchioli (2013), Th. 3.5. The error term is also given by

$$\epsilon_t = (\mathbf{w_t}^{'} \qquad \mathbf{u}_t^{'}(\mathbf{w}_t^{'} \otimes \mathbf{I}_K) \qquad \mathbf{u}_t^{'}(\pi^{'} \otimes \mathbf{I}_K)\Sigma^{'})^{'}$$

with variance matrix

$$\Xi = Var(\epsilon_t) = \text{diag}(\widetilde{\mathbf{D}} - \mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}^{'}, (\widetilde{\mathbf{D}} - \mathbf{F}\widetilde{\mathbf{D}}\mathbf{F}^{'}) \otimes \mathbf{I}_K, \Sigma((\mathbf{D}\mathbf{P}_{\infty}) \otimes \mathbf{I}_K)\Sigma^{'}).$$

Using (4) the spectral density matrix of the process $(\mathbf{y_t})$ in (3) is also given by

$$F_{\mathbf{y}}(\omega) = \frac{\theta(e^{i\omega})\Xi\theta^{'}(e^{-i\omega})}{|\phi(e^{i\omega})|^2}.$$

In fact, we can apply a well-known result (see, for example, Gourieroux and Monfort (1997), Chp.8, Formula 8.3, p.257). The spectral density of a VARMA process

$$\Phi(L)\mathbf{y_t} = \mathbf{\Theta(L)}\epsilon_\mathbf{t},$$

with $Var(\epsilon) = \mathbf{\Omega}$, is given by

$$F_{\mathbf{y}}(\omega) = \frac{1}{2\pi}\Phi^{-1}(exp(i\omega))\Theta(exp(i\omega))\mathbf{\Omega}\overline{\Theta(exp(i\omega))'}\ \overline{\Phi^{-1}(exp(i\omega))'}$$

This formula can be applied when $\det\Phi(z)$ has all its roots outside the unit circle. Moreover, we can also write $F_{\mathbf{y}}(\omega)$ as

$$F_{\mathbf{y}}(\omega) = \frac{1}{2\pi}\frac{\Phi^*(exp(i\omega))\Theta(exp(i\omega))\mathbf{\Omega}\overline{\Theta(exp(i\omega))'}\ \overline{\Phi^*(exp(i\omega))'}}{|\det\Phi(exp(i\omega))|^2}$$

where Φ^* denotes the adjoint matrix of Φ. Here, we apply these formulae ignoring the coefficient. Written in this form $F_{\mathbf{y}}(\omega)$ is a matrix whose elements are rational functions of $exp(i\omega)$. This property is a characteristic of the VARMA process.

4.2.3 Markov-switching VAR(p) process

Let us consider the MS-VAR(p), $p > 0$, process

$$A(L)\mathbf{y_t} = \nu_{s_t} + \Sigma_{s_t}\mathbf{u_t} \tag{4.5}$$

where $A(L) = \mathbf{I}_K - \mathbf{A}_1 L - \cdots - \mathbf{A}_p L^p$ is a $(K \times K)$-dimensional lag polynomial. Assume that there are no roots on or inside the unit circle of the complex plane, i.e., $|A(z)| \neq 0$ for $|z| \leq 1$. Reasoning as above, the process $(\mathbf{y_t})$ in (5) admits a stable VARMA(p^*, q^*) with $p^* \leq M+p-1$ and $q^* \leq M-1$:

$$\Psi(L)(\mathbf{y_t} - \mu_{\mathbf{y}}) = \theta(L)\epsilon_t \tag{4.6}$$

where $\Psi(L) = |F(L)|A(L) = \phi(L)A(L)$ and $\theta(L)\epsilon_t$ is as in (4). If we want the autoregressive part of the stable VARMA in (6) to be scalar, we have to multiply (6) on the left with the adjoint $A(L)^*$ to give a stable VARMA($p^{'}, q^{'}$) representation, where the bounds satisfy $p^{'} \leq M+Kp-1$ and $q^{'} \leq M+(K-1)p-1$. Thus the spectral density matrix of the process $(\mathbf{y_t})$ in (6) is given by

$$\begin{aligned} F_{\mathbf{y}}(\omega) &= \frac{A^{-1}(e^{i\omega})\theta(e^{i\omega})\Xi\theta^{'}(e^{-i\omega})[A^{'}(e^{-i\omega})]^{-1}}{|\phi(e^{i\omega})|^2} \\ &= \frac{A^*(e^{i\omega})\theta(e^{i\omega})\Xi\theta^{'}(e^{-i\omega})A^{*'}(e^{-i\omega})}{|\phi(e^{i\omega})|^2|\det A(e^{i\omega})|^2}. \end{aligned}$$

From the above section we can also obtain the matrix expression

$$\begin{aligned} F_{\mathbf{y}}(\omega) = {} & A^{-1}(e^{i\omega})Q[A^{'}(e^{-i\omega})]^{-1} + 2A^{-1}(e^{i\omega})\widetilde{\Lambda}\mathbf{F} \\ & \times \Re e\{(\mathbf{I}_{M-1}e^{i\omega} - \mathbf{F})^{-1}\}\widetilde{\mathbf{D}}\widetilde{\Lambda}^{'}[A^{'}(e^{-i\omega})]^{-1}. \end{aligned} \tag{4.7}$$

A similar result can be obtained for a Markov switching VAR(p), $p > 0$, process

$$A_{s_t}(L)\mathbf{y_t} = \nu_{s_t} + \Sigma_{s_t}\mathbf{u_t} \tag{4.8}$$

where we assume that the state variable is independent of the observables.

Define

$$\mathbf{A}(L) = (A_1(L) \dots A_M(L))$$

where

$$A_m(L) = \mathbf{I}_K - \mathbf{A}_{1,m} L - \cdots - \mathbf{A}_{p,m} L^p$$

for $m = 1, \dots, M$. Recall that $s_t \in \{1, \dots, M\}$. Then (8) can be written in the form

$$\mathbf{A}(L)(\xi_t \otimes \mathbf{I}_K)\mathbf{y_t} = \Lambda \xi_t + \Sigma(\xi_t \otimes \mathbf{I}_K)\mathbf{u_t}.$$

Assume that $B(L) = \mathbf{A}(L)(\pi \otimes \mathbf{I}_K)$ is invertible. Then the spectral density matrix of the process $(\mathbf{y_t})$ in (8) is given by

$$F_{\mathbf{y}}(\omega) = \frac{B^{-1}(e^{i\omega})\theta(e^{i\omega})\Xi\theta^{'}(e^{-i\omega})[B^{'}(e^{-i\omega})]^{-1}}{|\phi(e^{i\omega})|^2}.$$

Finally, we can also obtain the matrix expression

$$\begin{aligned} F_{\mathbf{y}}(\omega) = {} & B^{-1}(e^{i\omega})Q[B^{'}(e^{-i\omega})]^{-1} + 2B^{-1}(e^{i\omega})\widetilde{\Lambda}\mathbf{F} \\ & \times \mathcal{R}e\{(\mathbf{I}_{M-1}e^{i\omega} - \mathbf{F})^{-1}\}\widetilde{\mathbf{D}}\widetilde{\Lambda}^{'}[B^{'}(e^{-i\omega})]^{-1}. \end{aligned}$$

4.2.4 A numerical example

Let us consider the MS(2)-AR(1) model defined as

$$\begin{cases} y_t = a_1\, y_{t-1} + \nu_1 + \sigma_1\, u_t & s_t = 1 \\ y_t = a_2\, y_{t-1} + \nu_2 + \sigma_2\, u_t & s_t = 2 \end{cases} . \tag{4.9}$$

In Figure 1 we plot the spectral density of model (9), where $\sigma_1 = \sigma_2 = 1$, $\nu_1 = 1, \nu_2 = 2$, $p = 0.3$ and $q = 0.8$. When the AR(1) coefficients are both positive (top left panel: $a_1 = 0.4$ and $a_2 = 0.8$) or negative (top right panel: $a_1 = -0.4$ and $a_2 = -0.8$), the shape is similar to the typical spectral representation of an AR(1) with positive/negative coefficients. When the sign is opposite (bottom left panel: $a_1 = 0.4$ and $a_2 = -0.8$; bottom right panel: $a_1 = -0.4$

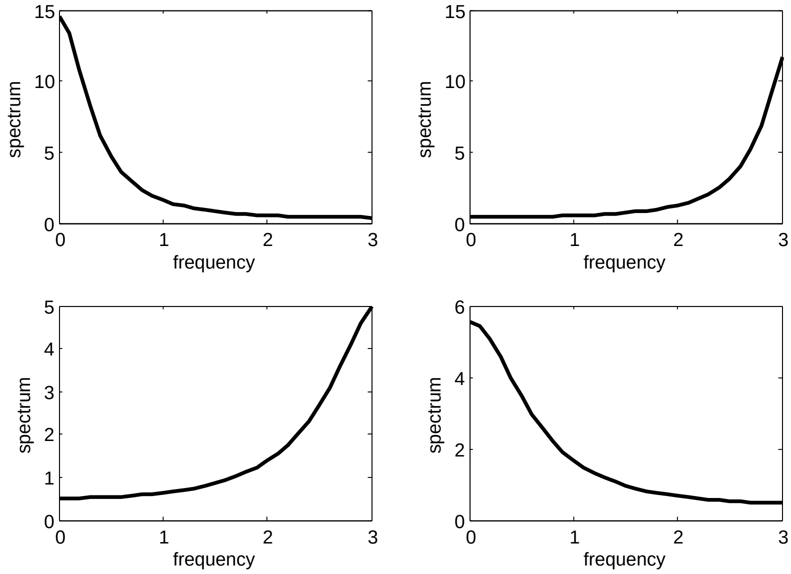

Figure 4.1: Plot of spectral density functions of MS(2)-AR(1)

and $a_2 = 0.8$) the prevailing shape depends on which model dominates in terms of absolute value of the coefficients and underlying probabilities.

4.2.5 Long memory or Regime Switching?

Since the empirical work of Ding, Engle and Granger (1993) and Ding and Granger (1996), a long debate on theoretical and empirical econometrics of long memory and fractional integration has been developed. A common finding is that returns themselves contain little serial correlation, while absolute reruns and their power transformations are highly correlated. In Ding et al. (1993) a long memory property for the absolute returns of S&P 500 daily stock market is established empirically. However, in more recent works more attention has been paid to the possibility of confusing long memory and structural change. Mikosch and Starica (2000) find structural change in asset return dynamics and argue that it could be responsible for evidence of long memory. Moreover, Diebold and Inoue (2001) show analytically that stochastic regime switching is easily confused with long memory, even asymptotically, so long as a small amount of regime switching occurs. This claim is supported with Monte Carlo analysis and there stochastic regime switching produce realizations that appear to have long memory in particular

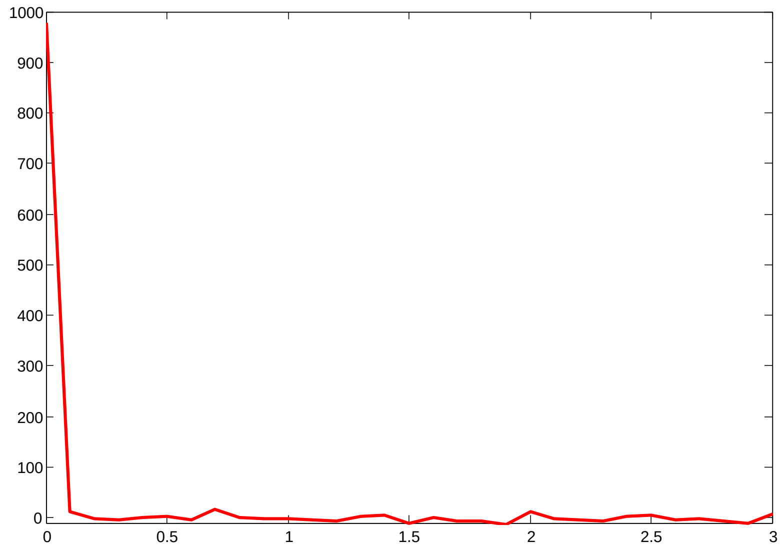

Figure 4.2: Spectra of S&P500 daily absolute log-return

cases. Granger and Hyung (2004) studied occasional structural breaks and the empirical results suggest a possibility such that, at least, part of the long memory can be caused by the presence of neglected breaks in the series. In the context of diffusion models, Chen, Hansen and Carrasco (2010) show how nonlinearity induces temporal dependence in continuous-time Markov models. Traditionally, long memory has been defined in the time domain in terms of decay rate of long-lag autocorrelations, or in the frequency domain in terms of rates of explosion of low-frequency spectra. In our work we obtain close form formulae for the spectra of Markov Switching VAR models and those can be used to investigate the "long memory or Markov switching" debate from the frequency domain prospective.

In particular, using the sample periodogram, we obtain the sample spectral density of absolute log-returns from S&P 500 daily stock market for the period January 2nd, 1957 September 30th, 2013. See Figure 2.

As expected, we recognize the explosion at low-frequency in the spectra. Then we follow a Monte Carlo experiment of Diebold and Inoue (2001) (Example 4.3, page 149) in which they analize the finite-sample property of the following Markov Switching model (in symbols MS(2)-AR(0)):

$$y_t = \mu_{s_t} + \epsilon_t$$

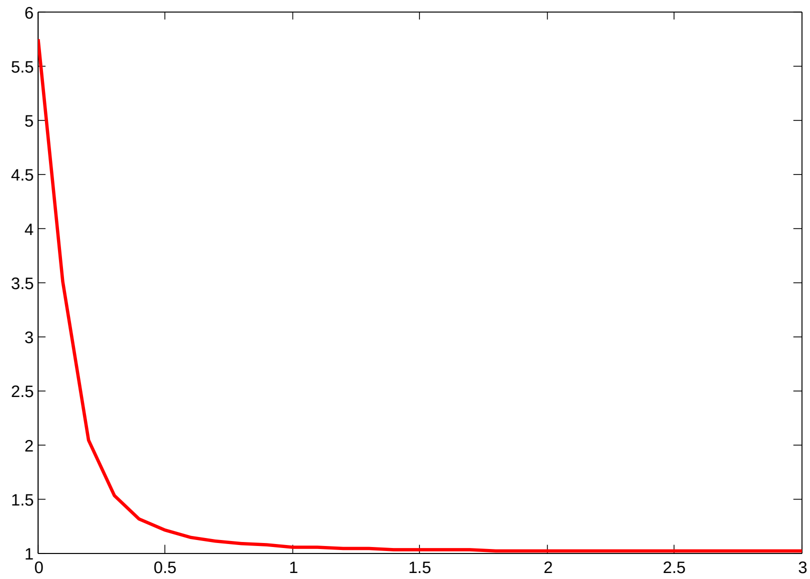

Figure 4.3: Spectra MS(2)-AR(0) with $p_{00} = p_{11} = 0.95$

where $\epsilon_t \sim IIDN(0, \sigma^2)$, and s_t and ϵ_τ are independent for all t and τ. The intercept term take values $\mu_0 = 0$ in regime 0 and $\mu_1 = 1$ in regime 1. For a short $T = 400$ or long $T = 10,000$ the following results remain the same. Diebold and Inoue (2001) observe the behavior of the data for different values of probabilities p_{00} and p_{11} and show that for equal probabilities close to one (say, 0.9995), the regime does not change with positive probability so that it does a good job of mimicking long memory as opposed to the equal probabilities case well away from unity (say, 0.95). We are able to confirm those conclusions from a different prospective through the analysis of the spectra of the two described cases. Figure 3 show the spectral density of the MS(2)-AR(0) model with equal probabilities which are away from unity; here the spectrum shows a smooth behaviour which vanishes only at frequency equal to 1. On the contrary, when probabilities are close to one (Figure 4) we witness an explosion at low-frequency in the spectra, which is exactly what we see in Figure 2. Therefore, stochastic regime switching is intimately related to long memory and easily confused with it, so long as only a small amount of regime switching occurs in an observed sample path.

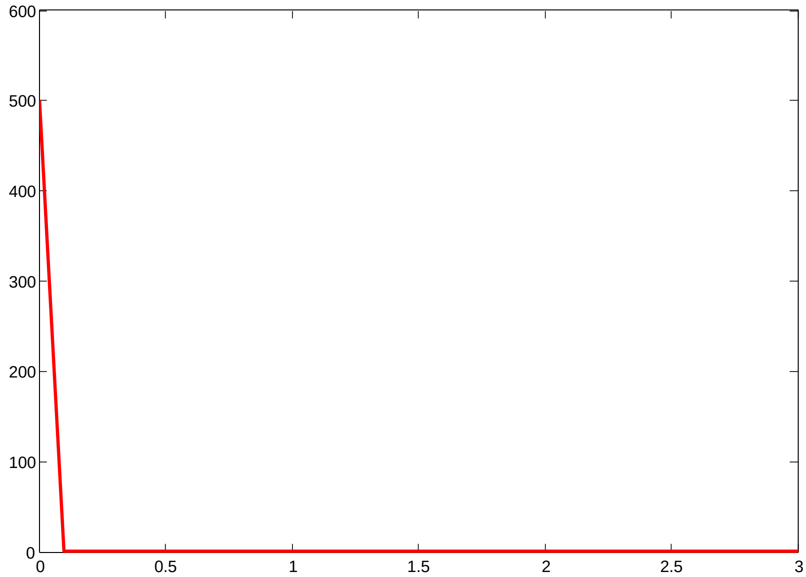

Figure 4.4: Spectra MS(2)-AR(0) with $p_{00} = p_{11} = 0.9995$

4.2.6 Conclusion

In this paper we derive close-form formulae for the spectral density function of MS-VAR models. These results are related to the work of Krolzig (1997) in terms of state space representation and stable representation and to the paper of Pataracchia (2011) where a different Markovian representation has been considered. Due to the simple tractability of this framework, we investigate via spectral analysis whether S&P500 stock market returns suffer of structural changes rather than long memory. These arguments will be developed in the next future for further research.

References

[1] Cavicchioli, M. (2013) *Determining the number of regimes in Markov-Switching VAR and VMA models*, forthcoming in Journal of Time Series Analysis.

[2] Cavicchioli, M. (2013) *Determining the number of regimes in Markov-Switching VAR and VMA models*, Working Papers from Department of Economics, University of Venice "Cá Foscari", N.03-2013.

[3] Chen, X., Hansen, L.P., Carrasco, M. (2010) *Nonlinearity and temporal dependence*, Journal of Econometrics **155**, 155–169.

[4] Diebold, F.X., Inoue, A. (2001) *Long memory and regime switching* Journal of Econometrics **105**, 131–159.

[5] Ding, Z., Granger, C.W.J., Engle, R.F. (1993) *A long memory property of stock market returns and a new model*, Journal of Empirical Finance **1**, 83–106.

[6] Ding, Z., Granger, C.W.J. (1996) *Modeling volatility persistence of speculative returns: a new approach*, Journal of Econometrics **73**, 185–215.

[7] Gourieroux, C., Monfort, A. (1997) Time Series and Dynamic Modelsn, *Cambridge University Press.*

[8] Granger, C.W.J., Hyung, N. (2004) *Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns* Journal of Empirical Finance **11**, 399–421.

[9] Mikosch, T., Stărică, C. (2000) *Is it really long memory we see in financial returns?* Extremes and Integrated Risk Management **12**, 149–168.

[10] Krolzig, H.M. (1997) Markov-Switching Vector Autoregressions: Modelling, Statistical Inference and Application to Business Cycle Analysis, *Springer Verlag*, Berlin-Heidelberg-New York.

[11] Pataracchia, B., (2011) *The Spectral Representation of Markov Switching ARMA Models*, Economics Letters **112**, 11–15.