Markov-Regime Switching in Economic Variables: Part I. Modelling, Estimating and Testing. - Part II. A Selective Survey

Sylvia Kaufmann
Martin Scheicher

INSTITUT FÜR HÖHERE STUDIEN
INSTITUTE FOR ADVANCED STUDIES
Vienna

Institut für Höhere Studien (IHS), Wien
Institute for Advanced Studies, Vienna

Markov-Regime Switching in Economic Variables: Part I. Modelling, Estimating and Testing. - Part II. A Selective Survey

Sylvia Kaufmann and Martin Scheicher

Markov-Regime Switching in Economic Variables: Part I. Modelling, Estimating and Testing. - Part II. A Selective Survey

Sylvia Kaufmann and Martin Scheicher

November 1996

Sylvia Kaufmann
Department of Economics
University of Vienna
Hohenstaufengasse 9
A-1010 Vienna, AUSTRIA
Phone: ++43/1/40103-3365
e-mail: Sylvia.Kaufmann@univie.ac.at

Martin Scheicher
Department of Economics
University of Vienna BWZ
Bruenner Str. 72
A-1210 Vienna, AUSTRIA
Phone: ++43/1/29128-571
e-mail: Martin.Scheicher@univie.ac.at

Institut für Höhere Studien (IHS), Wien
Institute for Advanced Studies, Vienna

The Institute for Advanced Studies in Vienna is an independent center of postgraduate training and research in the social sciences. The **Economics Series** presents research done at the Economics Department of the Institute for Advanced Studies. Department members, guests, visitors, and other researchers are invited to contribute and to submit manuscripts to the editors. All papers are subjected to an internal refereeing process.

Abstract

Modelling the growth rate of economic time series with a Markov switching process in their mean and/or their variance allows to take account of two facts that are often encountered in such series, namely that the periods in which each mean is prevailing differ in their duration and that the variance of the time series differ in each period. In a first part, we will motivate the class of regime switching models, and revue the estimating and testing procedures. In the second part, we will present a brief survey of the literature on regime switching models and their applications, and also present first results of actual own research.

Keywords

JEL-Classifications

1   Introduction

The present working paper is a summary of several presentations we held at the Institute for Advanced Studies in Vienna. The purpose was to introduce the new field of estimating and testing Markov switching models in applied econometric work. The main part of the paper will be a condensed exposition of review material. So it can be used by researchers, who wish to acquaint with this new econometric field and who look for a brief summary of the literature. It is clear that we cannot survey the literature entirely. Rather, we selected out the papers according to our own research, and enabling us to give an overview of what we thought to be milesteps in the field.

The seminal papers of Hamilton[23, 24] established the method of modelling time series with changes in regime following a Markov process as one alternative of the nonlinear modelling methods. His approach is based on Goldfeld and Quandt's[19] Markov switching regression to characterize the change in the parameters of the time series process. He presents a nonlinear filter and smoother to obtain statistical estimates of the regime in which the unobserved state variable is in a specific period in time. This inference is based on the time path of an observed variable. The second step of his procedure is to solve for the marginal likelihood function of the observed variable, to maximize this likelihood function with respect to the parameters, and then use the estimates and the data to draw the optimal statistical inference about the unobserved regimes. The new technique is applied to postwar US quarterly data on real GNP. The best empirical fit to the data is given when the growth states are associated with the business cycle. Detailed results are found in the application section.

In the following, researchers developed further the new framework. Hamilton[24] introduces the EM algorithm to obtain maximum likelihood estimates of the parameters for time series processes subject to discrete regime shifts. In Lam[32] the Hamilton model is extended to the case where the autoregressive part has no unit root. The important issue of testing N versus $N-1$ regimes is addressed in Hansen[27], who derives a bound for the likelihood ratio test. Garcia[16] derives analytically the asymptotic null distribution of the likelihood ratio test and the related covariance functions for various Markov switching models. Finally, Kim[29] casts the Markov-switching model into a state space form which allows a much broader class of models to be estimated than before within that framework. Moreover, his algorithm proves to be much more efficient than the previous ones.

Why has this framework become so attractive for economic research? One advantage is certainly that several facts arising in time series can be modelled in a simultaneous way. Often we observe that time series undergo several breaks in the trend, that some periods without shifts last longer than others, and that periods with high volatility follow periods with low volatility. All these facts may arise in macroeconomic time series if the business cycle is asymmetric. Previous research tried to assess for asymmetries. Already in Neftçi[35] a finite state Markov process is used to test if the unemployment rate is

asymmetric over the business cycle. He finds considerable evidence that time series go through two different processes over the business cycle. However, in Falk[12] the evidence for asymmetry in U.S. GNP, investment and productivity, and in industrial production for five other countries is weaker than in the unemployment evidence. Westlund and Öhlén[48] find also only weak evidence in Swedish data. They cannot reject the null of symmetry in deterministically detrended data and in data detrended by Beveridge-Nelson. Nevertheless, the Markov regime switching framework of Hamilton has revived the discussion about asymmetric business cycles. The evidence found in data analysis would enforce the relevance of theoretical work presented recently. The statistical properties of these models are well described py processes following a Markov process. As we will concentrate on the estimation and testing procedures, the interested reader can find a brief overview of the macroeconomic literature in Diebold and Rudebusch[6] and a survey on financial econometrics in Pagan[36].

The remaining part of the paper is organized as following. Section 2 and 3 describe the EM algorithm of Hamilton. Notation and description therein are closely following Hamilton[25, chap. 22]. Hansen's approach to test the Markov switching model is reproduced in section 4. Finally, section 5 and 6 give a selective survey on Markov switching models in macroeconomics and in finance, respectively.

Part I

Modelling, estimating and testing

2 Basic concepts

2.1 An example

In figure 1 we see the first differences of the logarithm of US real GNP. We can observe that longer periods of positive growth rates are followed by shorter periods of negative growth rates. Moreover, the duration of these periods is changing over time, so that shifts from positive to negative (or from negative to positive) growth rates are often difficult to forecast and not observable. A simple model should then describe with what probability a shift from positive to negative growth rate can occur. Modelling the time series as a first order autoregressive process with changes in mean, we can write:

$$y_t - \mu_{s_t^*} = \varphi(y_{t-1} - \mu_{s_{t-1}^*}) + \varepsilon_t, \tag{1}$$

where $\mu_{s_t^*}$ is $\mu_1 \leq 0$ when $s_t^* = 1$, and is $\mu_2 > 0$ when $s_t^* = 2$.

s_t^* is a discrete variable, that takes on the values 1 or 2. One possibility of modelling such a variable is to let her follow a Markov process. To simplify notation we further assume $\varphi = 0$.

Then we have

$$
\begin{aligned}
y_t &= c_{s_t} + \varepsilon_t \qquad \varepsilon_t \sim i.i.d. \ \mathcal{N}(0, \sigma^2) \\
c_{s_t} &= \mu_1, \ \text{if } s_t = 1, \\
c_{s_t} &= \mu_2, \ \text{if } s_t = 2,
\end{aligned}
\tag{2}
$$

with the matrix of transition probabilities:

$$
P = \begin{bmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{bmatrix} = \begin{bmatrix} p_{11} & 1 - p_{22} \\ 1 - p_{11} & p_{22} \end{bmatrix}
$$

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

$$
\sum_{j=1}^{2} p_{ij} = 1 \qquad i = 1, 2.
\tag{3}
$$

Assuming a Markov process of order 1^1 for s_t is not restrictive, as a N-state Markov process of order p can always be rewritten as a N^p-state Markov process of order 1 (see example II on page 6).

Before we turn to the estimation of the parameters governing this process, we introduce two useful concepts in the next subsections.

2.2   The Markov process as a vector autoregression

As in (2) let us assume that the process for the unobservable state variable s_t is described by a first-order, 2-state Markov process. We get a useful representation when we define a 2×1 vector ξ_t, whose jth element is 1 when $s_t = j$ and 0 otherwise.

$$
\xi_t = \begin{cases} (1, 0)' & s_t = 1 \\ (0, 1)' & s_t = 2 \end{cases}
\tag{4}
$$

If $s_t = 1$, then the first element of ξ_{t+1} is a stochastic variable, that takes on the value 1 with probability p_{11} and the second element of ξ_{t+1} takes on value 1 with probability $1 - p_{11}$. Thus, the conditional expectation of ξ_{t+1} given $s_t = i$ is

$$
E(\xi_{t+1} | s_t = i) = \begin{bmatrix} p_{i1} \\ p_{i2} \end{bmatrix}
\tag{5}
$$

This vector is the ith column of P. If $s_t = i$, then the vector ξ_t is the ith column of I_2 (the 2×2 identity matrix), so that we can write the previous equation (5) as

$$
E(\xi_{t+1} | \xi_t) = P \xi_t.
\tag{6}
$$

---

[1]This meens that the probability of $s_t = j$ depends on the past only through the lagged value s_{t-1}:

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

Having assumed a Markov process of order one we can further write:

$$E\left(\xi_{t+1}|\xi_t,\xi_{t-1},\ldots\right) = P\xi_t. \tag{7}$$

This implies that we can express a Markov process as:

$$\xi_{t+1} = P\xi_t + \mathbf{v}_{t+1} \tag{8}$$

where

$$\mathbf{v}_{t+1} \equiv \xi_{t+1} - E\left(\xi_{t+1}|\xi_t,\xi_{t-1},\ldots\right).$$

Expression (8) is a first order vector autoregression. $\mathbf{v}_t$ is a martingale difference sequence: Although it can take on a finite set of values, on average it is zero and moreover, its value is impossible to forecast on the basis of previous states of the process.

2.3 Ergodic probabilities, unconditional probabilities

From the summation restriction (3) we have

$$P'\mathbf{1} = \mathbf{1}, \tag{9}$$

where here $\mathbf{1}$ is a 2×1 vector of ones. This equation means that $\mathbf{1}$ is an eigenvalue of P' and that $\mathbf{1}$ is the corresponding eigenvector. As a matrix and its transpose have the same eigenvalues, $\mathbf{1}$ is an eigenvalue of P for every Markov process. If a non-reducible Markov process[2] with transition probabilities matrix P has an eigenvalue equal to one and all other eigenvalues are smaller than 1, then this Markov process is said to be *ergodic*. The vector of *ergodic probabilities* is represented by π. This vector is the eigenvector corresponding to the eigenvalue 1. Thus the vector of ergodic probabilities π satisfies:

$$P\pi = \pi. \tag{10}$$

The elements of π are normalized to sum to unity. It can be shown (see Hamilton[25, chap. 22.2]) that if P is the matrix of transition probabilities of an ergodic Markov process, then

$$\lim_{m\to\infty} P^m = \pi \cdot \mathbf{1}'. \tag{11}$$

This result implies that the long run forecast of an ergodic Markov chain is independent of the current state:

$$E\left(\xi_{t+m}|\xi_t,\xi_{t-1},\ldots\right) = P^m\xi_t \xrightarrow{p} \pi \cdot \mathbf{1}'\xi_t = \pi. \tag{12}$$

The vector of ergodic probabilities can also be interpreted as the *vector of the unconditional probabilities* of being in state j: $P\left(s_t = j\right)$, $j = 1, 2$. The vector π is then the unconditional expectation of ξ_t:

$$\pi = E\left(\xi_t\right) \tag{13}$$

---

[2]Suppose that p_{11} is 1, so that P is upper triangular. This means that once the process enters state 1, it will never return to state 2. In this case we name state 1 an absorbing state and the Markov chain reducible. If a Markov chain is not reducible it is called irreducible or non-reducible.

We easily derive this by taking unconditional expectations of (8),

$$E\left(\xi_{t+1}\right) = P \cdot E\left(\xi_t\right).\tag{14}$$

Assuming stationarity and using (13) this becomes

$$\pi = P \cdot \pi.\tag{15}$$

But π is the eigenvector corresponding to the unit eigenvalue of P. For an ergodic Markov chain, this eigenvector is unique. Thus the vector of ergodic probabilities π can be interpreted as the vector of unconditional probabilities.

For a Markov process with two states the elements of this vector are:

$$\pi = \left[\begin{array}{c} (1 - p_{22}) / (2 - p_{11} - p_{22}) \\ (1 - p_{11}) / (2 - p_{11} - p_{22}) \end{array} \right]$$

3 Estimating Markov-switching models

3.1 Notation

We introduce some more notation that we need to estimate the model.

Let y_t be a $n \times 1$ vector of observable endogenous variables, x_t a $k \times 1$ vector of exogenous variables and $\mathcal{Y}_t = \left(y'_t, \ldots, y'_{-m}, x'_t, \ldots, x'_{-m} \right)'$ be a vector containing all observations through date t. Given the regime $s_t = j$ at date t, the conditional density of y_t is assumed to be:

$$f\left(y_t | s_t = j, x_t, \mathcal{Y}_{t-1}; \theta\right),\tag{16}$$

where θ is a vector of parameters that includes the parameters characterizing the conditional density. If there are N different regimes, then there are N conditional densities given by 16. These densities are collected in a $N \times 1$ vector η_t:

$$\eta_t = \left[\begin{array}{c} f\left(y_t | s_t = 1, x_t, \mathcal{Y}_{t-1}; \theta\right) \\ \vdots \\ f\left(y_t | s_t = N, x_t, \mathcal{Y}_{t-1}; \theta\right) \end{array} \right].\tag{17}$$

It is assumed that the unobserved stochastic state variable s_t evolves according to a Markov chain that is independent of past observations on y_t or current or past x_t:

$$P\left(s_t = j | s_{t-1} = i, s_{t-2} = k, \ldots, x_t, \mathcal{Y}_{\sqcup-\infty}; \theta\right) =$$
$$P\left(s_t = j | s_{t-1} = i; \theta\right) = p_{ij} \qquad \text{for } i, j = 1, 2, \ldots, N.\tag{18}$$

The probabilities p_{ij}, $i, j = 1, \ldots, N$ are also included in the parameter vector θ.

3.2 Examples

I. Let y_t be a scalar variable normally distributed with mean μ_1 and variance σ^2 when the first regime is prevailing at date t and normally distributed with mean μ_2 and variance σ^2 if the second regime is prevailing at date t. The model for this variable is simply:

$$y_t = c_{s_t} + \varepsilon_t \qquad \varepsilon_t \sim i.i.d. \; \mathcal{N}(0, \sigma^2)$$
$$c_{s_t} = \mu_1, \; \text{if } s_t = 1$$
$$c_{s_t} = \mu_2, \; \text{if } s_t = 2.$$

The vector η_t of conditional densities then becomes:

$$\eta_t = \begin{bmatrix} f(y_t | s_t = 1; \theta) \\ f(y_t | s_t = 2; \theta) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(y_t - \mu_1)^2}{2\sigma^2} \right\} \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(y_t - \mu_2)^2}{2\sigma^2} \right\} \end{bmatrix}$$

where $\theta = (\mu_1, \mu_2, \sigma^2, p_{11}, p_{22})$.

II. Alternatively, assume that y_t followed an autoregressive process with changes in mean:

$$y_t - c_{s_t}^* = \phi(y_{t-1} - c_{s_{t-1}}^*) + \varepsilon_t \qquad \varepsilon_t \sim i.i.d. \; \mathcal{N}(0, \sigma^2)$$
$$c_{s_t}^* = \mu_1, \; \text{if } s_t^* = 1$$
$$c_{s_t}^* = \mu_2, \; \text{if } s_t^* = 2$$

To make y_t independent of the lagged state we expand the number of states to four and define them in the following way:

$$s_t = 1 \quad \text{if} \quad s_t^* = 1, \quad s_{t-1}^* = 1 \quad \text{with prob.} \qquad p_{11}^*$$
$$s_t = 2 \quad \text{if} \quad s_t^* = 2, \quad s_{t-1}^* = 1 \quad \text{with prob.} \quad p_{12}^* = 1 - p_{11}^*$$
$$s_t = 3 \quad \text{if} \quad s_t^* = 1, \quad s_{t-1}^* = 2 \quad \text{with prob.} \quad p_{21}^* = 1 - p_{22}^*$$
$$s_t = 4 \quad \text{if} \quad s_t^* = 2, \quad s_{t-1}^* = 2 \quad \text{with prob.} \qquad p_{22}^*$$

The matrix of transition probabilities for s_t is then:

$$P = \begin{bmatrix} p_{11}^* & 0 & p_{11}^* & 0 \\ p_{12}^* & 0 & p_{12}^* & 0 \\ 0 & p_{21}^* & 0 & p_{21}^* \\ 0 & p_{22}^* & 0 & p_{22}^* \end{bmatrix}.$$

Moreover, we get the vector of conditional densities:

$$\eta_t = \begin{bmatrix} f(y_t | y_{t-1}, s_t = 1; \theta) \\ f(y_t | y_{t-1}, s_t = 2; \theta) \\ f(y_t | y_{t-1}, s_t = 3; \theta) \\ f(y_t | y_{t-1}, s_t = 4; \theta) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(y_t - \mu_1 - \phi(y_{t-1} - \mu_1))^2}{2\sigma^2} \right\} \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(y_t - \mu_2 - \phi(y_{t-1} - \mu_1))^2}{2\sigma^2} \right\} \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(y_t - \mu_1 - \phi(y_{t-1} - \mu_2))^2}{2\sigma^2} \right\} \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(y_t - \mu_2 - \phi(y_{t-1} - \mu_2))^2}{2\sigma^2} \right\} \end{bmatrix}$$

where $\theta = (\mu_1, \mu_2, \phi, \sigma^2, p_{11}^*, p_{22}^*)$.

3.3 Estimation

The issue will be to estimate the parameters in θ based on observations of $\mathcal{Y}_T$. The problem that arises now is that even if θ were known, because s_t is not observable we don't know which regime was prevailing at every point in time. Let $P\left(s_t = j|\mathcal{Y}_t; \theta\right)$ be the statistical inference about the value of s_t based on data through observation t and based on knowledge of the population parameters θ. This inference is a conditional probability of the observation t being generated by regime j. These conditional probabilities $P\left(s_t = j|\mathcal{Y}_t; \theta\right), j = 1, \ldots, N$ are gathered in a $N \times 1$ vector $\widehat{\xi}_{t|t}$. Forecasts about the probability of being in regime j at date $t + 1$ given information through period t are collected in a vector $\widehat{\xi}_{t+1|t}$, the jth element of which is $P\left(s_{t+1} = j|\mathcal{Y}_t; \theta\right)$. The optimal forecast and inference for each period t is made by iterating through the following equations:

$$\widehat{\xi}_{t|t} = \frac{\left(\widehat{\xi}_{t|t-1} \odot \eta_t\right)}{\mathbf{1}'\left(\widehat{\xi}_{t|t-1} \odot \eta_t\right)} \tag{19}$$

$$\widehat{\xi}_{t+1|t} = P \cdot \widehat{\xi}_{t|t}. \tag{20}$$

η_t is the $N \times 1$ vector of the conditional densities of y_t, P is the matrix of transition probabilities, $\mathbf{1}$ is a $N \times 1$ vector of ones and the symbol $\odot$ means the Hadamard vector product. Given a starting value $\widehat{\xi}_{1|0}$ and given a first guess of the population parameters in θ we can iterate through both equations and compute values for $\widehat{\xi}_{t|t}$ and for $\widehat{\xi}_{t+1|t}$. As a by-product we can evaluate the likelihood function $L\left(\theta\right)$ for the observations $\mathcal{Y}_T$ at the value fixed for θ that was used to iterate over (19) and (20).

$$L\left(\theta\right) = \sum_{t=1}^{T} \log f\left(y_t|x_t, \mathcal{Y}_{t-1}; \theta\right), \tag{21}$$

where

$$f\left(y_t|x_t, \mathcal{Y}_{t-1}; \theta\right) = \mathbf{1}'\left(\widehat{\xi}_{t|t-1} \odot \eta_t\right).$$

3.4 Interpretation of filter equations

We defined

$$\widehat{\xi}_{t|t} = \begin{bmatrix} P\left(s_t = 1|\mathcal{Y}_t; \theta\right) \\ \vdots \\ P\left(s_t = N|\mathcal{Y}_t; \theta\right) \end{bmatrix}, \tag{22}$$

each element of which is the conditional probability $s_t = j$ given $\mathcal{Y}_t$, $j = 1, \ldots N$. The forecasts given information through period t are summed up in

$$\widehat{\xi}_{t+1|t} = \begin{bmatrix} P\left(s_{t+1} = 1|\mathcal{Y}_t; \theta\right) \\ \vdots \\ P\left(s_{t+1} = N|\mathcal{Y}_t; \theta\right) \end{bmatrix}. \tag{23}$$

The numerator of (19) is given by:

$$\widehat{\xi}_{t|t-1} \odot \eta_t = \begin{bmatrix} P(s_t = 1 | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta) \times f(\mathsf{y}_t | s_t = 1, \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta) \\ \vdots \\ P(s_t = N | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta) \times f(\mathsf{y}_t | s_t = N, \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta) \end{bmatrix}$$

$$= \begin{bmatrix} p(\mathsf{y}_t, s_t = 1 | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta) \\ \vdots \\ p(\mathsf{y}_t, s_t = N | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta) \end{bmatrix}. \tag{24}$$

This vector can be interpreted as the joint conditional density of y_t and s_t given information through date t. Note that x_t is exogenous and contains no information about s_t so the jth element of $\widehat{\xi}_{t|t-1}$ can also be written as $P(s_t = j | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta)$.

Let's turn to the denominator of (19). Summing up over the N values of (24) gives us the conditional density of y_t given past observations. This can be written as

$$1'\left(\widehat{\xi}_{t|t-1} \odot \eta_t\right) = f(\mathsf{y}_t | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta) \tag{25}$$

Forming the ratio of (24) and (25) leads to

$$\begin{bmatrix} \frac{p(\mathsf{y}_t, s_t = 1 | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta)}{f(\mathsf{y}_t | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta)} \\ \vdots \\ \frac{p(\mathsf{y}_t, s_t = N | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta)}{f(\mathsf{y}_t | \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta)} \end{bmatrix} = \begin{bmatrix} P(s_t = 1 | \mathsf{y}_t, \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta) \\ \vdots \\ P(s_t = N | \mathsf{y}_t, \mathsf{x}_t, \mathcal{Y}_{t-1}; \theta) \end{bmatrix}$$

$$= \begin{bmatrix} P(s_t = 1 | \mathcal{Y}_t; \theta) \\ \vdots \\ P(s_t = N | \mathcal{Y}_t; \theta) \end{bmatrix} \tag{26}$$

$$= \widehat{\xi}_{t|t}.$$

Dividing the joint conditional density of y_t and s_t by the conditional density of y_t gives the conditional probability of s_t given the observations through t. This way, we obtain a statistical inference about the probability with which s_t was in state j at date t

Taking expectations of (8) yields

$$E(\xi_{t+1} | \mathcal{Y}_t; \theta) = P \cdot E(\xi_t | \mathcal{Y}_t; \theta) + \overbrace{E(v_{t+1} | \mathcal{Y}_t; \theta)}^{=0}$$

$$\begin{bmatrix} P(s_{t+1} = 1 | \mathcal{Y}_t; \theta) \\ \vdots \\ P(s_{t+1} = N | \mathcal{Y}_t; \theta) \end{bmatrix} = P \cdot \begin{bmatrix} P(s_t = 1 | \mathcal{Y}_t; \theta) \\ \vdots \\ P(s_t = N | \mathcal{Y}_t; \theta) \end{bmatrix} \tag{27}$$

$$\widehat{\xi}_{t+1|t} = P \cdot \widehat{\xi}_{t|t},$$

which is the forecast of $\widehat{\xi}_t$ given information up to period t.

In appendix C we show how the EM algorithm is implemented for example I and II on page 6.

3.5   Smoothing

The result of the smoothing algorithm is an inference about the probability of being in state j at date t given the information in the whole sample.

$$\widehat{\xi}_{t|T} = \widehat{\xi}_{t|t} \odot P' \cdot \left[\widehat{\xi}_{t+1|T}(\div)\widehat{\xi}_{t+1|t} \right] \tag{28}$$

$(\div)$ is similar to $\odot$ and means elementwise division.

3.6   Maximizing

The second step consists in maximizing the log likelihood with respect to θ. If the transition probabilities are restricted only by $p_{ij} \geq 0$ and $(p_{i1} + \ldots + p_{iN}) = 1$ for all i, and if the initial probability $\widehat{\xi}_{1|0}$ is assumed to be a fixed value ρ, then the maximum likelihood estimates of the transition probabilities are

$$\widehat{p}_{ij} = \frac{\sum_{t=2}^{T} P\left(s_t = j, s_{t-1} = i | \mathcal{Y}_T; \widehat{\theta} \right)}{\sum_{t=2}^{T} P\left(s_{t-1} = i | \mathcal{Y}_T; \widehat{\theta} \right)} \tag{29}$$

where $\widehat{\theta}$ is the maximum likelihood estimate of θ. $\widehat{p}_{ij}$ is the number of times that state i seemed to be followed by state j divided by the number of times the process was in state i. This inference is made on the basis of the smoothed probabilities in (28) obtained through the filter equations.

The maximum likelihood estimates of the other parameters governing the conditional density (16) are given by

$$\sum_{t=1}^{T} \left(\frac{\partial \log \eta_t}{\partial \alpha'} \right)' \widehat{\xi}_{t|T} = 0 \tag{30}$$

where α is equal to θ without the transition probabilities.

If for example we have

$$y_t = z_t' \beta_{s_t} + \varepsilon_t, \qquad \varepsilon_t \sim i.i.d. \ \mathcal{N}(0, \sigma^2)$$
$$\beta_{s_t} = \beta_1 \text{ if } s_t = 1$$
$$\vdots \tag{31}$$
$$\beta_{s_t} = \beta_N \text{ if } s_t = N$$
$$\text{with } \alpha = (\beta_1', \ldots, \beta_N', \sigma^2).$$

Here z_t is a vector of explanatory variables that could include lagged values of y_t. The coefficients take on the value β_j if state j is prevailing at date t. The vector of conditional densities is

$$\eta_t = \begin{bmatrix} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(y_t - z_t'\beta_1)^2}{2\sigma^2} \right\} \\ \vdots \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(y_t - z_t'\beta_N)^2}{2\sigma^2} \right\} \end{bmatrix}. \tag{32}$$

Condition (30) yields for $\hat{\beta}_j$:

$$\hat{\beta}_j = \frac{\sum_{t=1}^{T} z_t y_t \cdot P\left(s_t = j | \mathcal{Y}_T; \theta\right)}{\sum_{t=1}^{T} z_t z_t' \cdot P\left(s_t = j | \mathcal{Y}_T; \theta\right)} \tag{33}$$

and for $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{N} \left(y_t - z_t' \hat{\beta}_j\right)^2 \cdot P\left(s_t = j | \mathcal{Y}_T; \theta\right) \tag{34}$$

In this case, we see that β_j can be estimated by weighted least squares, where observations for y_t and z_t are weighted by the probability that they came from regime j. $\hat{\beta}_j$ is then estimated by an OLS regression from $y_t^*(j)$ on $z_t^*(j)$, where

$$y_t^*(j) = y_t \cdot \sqrt{P(s_t = j | \mathcal{Y}_t; \hat{\theta})}$$
$$z_t^*(j) = z_t \cdot \sqrt{P(s_t = j | \mathcal{Y}_t; \hat{\theta})}$$

The estimate of σ^2 is then $\frac{1}{T}$ times the combined sum of squared residuals from these N regressions. A derivation of the estimators (33) and (34) can be found in appendix D. Maximum likelihood estimators for further model specifications are found in Hamilton[23, 28].

3.7 Extensions of estimation procedure

The EM-algorithm of Hamilton is easily implemented, but has proven to be time-consuming (especially when estimating multivariate models). Moreover, when the models become more sophisticated the variable transformation and the derivation of the maximum likelihood estimates are not straightforward any more. Kim[29] has extended the Hamilton estimation procedure. He casts the regime switching model into a state space model and presents a new filtering and smoothing algorithm. The approximation needed to make the Kalman filter operable is not restrictive. The major improvements are that his algorithm allows a broader class of models to be estimated than before and the extremely increased efficiency in computing smoothed probabilities.

Moreover, Hamilton's model is not restricted to constant transition probabilities. The extension to time-varying transition probabilities is presented in Diebold, Lee and Weinbach[5]. They develop the EM algorithm for this specification and present a simulation example. They suggest that the extended model may be applied to exchange-rate dynamics, as transition probabilities may vary with fundamentals, or to output dynamics to model duration dependence. In the latter case varying transition probabilities reflect the fact that the longer a contraction or recovery period persists the more likely they are to end.

4 Testing Markov switching models

One question we would like to answer is "How many states are governing the process of s_t?". Unfortunately, the well-known likelihood ratio (LR) test cannot be performed. If we estimate a model with N states and test the null of $N-1$ states, the parameter for the Nth state are not identified under the null. The issue was first addressed in Hansen[27]. Recently, Garcia[16] derived the asymptotic null distributions of the LR statistic for several specific Markov switching models commonly used in the literature.

We begin with a simple model to illustrate the testing purpose. Assume that y_t follows a first-order two-state Markov process with an autoregressive component.

$$y_t = \mu + \mu_d s_t + u_t, \qquad \phi(L)u_t = e_t,$$
$$P(s_t = 1|s_{t-1} = 1) = p \tag{35}$$
$$P(s_t = 0|s_{t-1} = 0) = q.$$

The hypothesis we want to test is one against two states, that is

$$H_0 : \ \mu_d = 0. \tag{36}$$

Two problems arise when we want to perform this test. First, under the null, the parameter p and q are not identified, that is, under the null we get the same value for the likelihood for every value of p and q between 0 and 1. Second, under the null of a linear model with $p = 0$ or $p = 1$ the information matrix becomes singular, because the scores of p, q and μ_d are zero. In this case, asymptotic distributional theory cannot be applied to derive the distribution of the LR statistic under the null.

Instead, Hansen views the likelihood function as an empirical process of the unknown parameters and uses empirical process theory to derive a bound for the asymptotic distribution of a standardized LR statistic. This distribution depends upon the covariance function of the empirical process associated with the likelihood surface, but nevertheless can be obtained by simulation[3]. In this paper, we will briefly reproduce Hansen's intuitive description of the approach and sum up the final results[4].

Assume that the log likelihood of a likelihood function that depends on unknown parameters can be written as

$$L_n(\alpha) = \sum_{i=1}^{n} l_i(\alpha). \tag{37}$$

The hypothesis we want to test is

$$H_0 : \alpha = \alpha_0 \qquad H_1 : \alpha \neq \alpha_0. \tag{38}$$

---

[3]The interested reader may refer to Gaenssler and Stute[15, 14] and to Dold and Eckmann[7] for further readings on empirical process theory.

[4]In the following, $\Rightarrow$ denotes weak convergence of probability measures with respect to the uniform metric.

The corresponding LR function is

$$LR_n(\alpha) = L_n(\alpha) - L_n(\alpha_0) = \sum_{i=1}^{n} [l_i(\alpha) - l_i(\alpha_0)] . \tag{39}$$

Because the LR function is a linear transformation of the log likelihood function, the maximum likelihood estimate of α is given by the parameters that maximize the LR function. Moreover, the LR statistic for H_0 versus H_1 is given by the supremum of the LR function:

$$LR_n = \sup_{\alpha \in A} LR_n(\alpha). \tag{40}$$

Further, the LR function can be decomposed in its mean plus the deviation from the mean

$$LR_n(\alpha) = R_n(\alpha) + Q_n(\alpha), \tag{41}$$

where

$$R_n(\alpha) = E[LR_n(\alpha)]$$

is the mean,

$$Q_n(\alpha) = \sum_{1}^{n} q_i(\alpha)$$

is the deviation from the mean and

$$q_i(\alpha) = [l_i(\alpha) - l_i(\alpha_0)] - E[l_i(\alpha) - l_i(\alpha_0)] .$$

Under standard regularity conditions $n^{-1}R_n(\alpha) \to_p R(\alpha)$ for all α, where $R(\alpha) = E[l_i(\alpha) - l_i(\alpha_0)]$. The function $R_n(\alpha)$ is maximized at the true value for α. Under the null, then, $R_n(\alpha)$ is not positive, strictly negative for $\alpha \neq \alpha_0$. If we could observe $R_n(\alpha)$ there would be no uncertainty. Instead we observe $LR_n(\alpha)$, which contains the influence of the random function $Q_n(\alpha)$. Some insights may then be gained by studying the stochastic process $Q_n(\alpha)$. When standardized, we find

$$\frac{1}{\sqrt{n}}Q_n(\alpha) = \frac{1}{\sqrt{n}}\sum_{1}^{n} q_i(\alpha) \Rightarrow Q(\alpha), \tag{42}$$

where $Q(\alpha)$ is a Gauss process with mean 0 and covariance function[5]

$$K(\alpha_1, \alpha_2) = \sum_{k=-\infty}^{\infty} E(q_i(\alpha_1)q_{i+k}(\alpha_2)) .$$

For every value of α, $Q(\alpha)$ is a normal random variable with mean 0 and variance $K(\alpha, \alpha)$. The function $K(\cdot, \cdot)$ describes the covariances between $Q(\alpha)$ at different values for α.

---

[5]This covariance function and the sample analogue in 56 as well as equation 57 are taken from Hansen's own erratum to his paper that he kindly sent us upon request.

The decomposition (41) can be written as an asymptotic approximation:

$$
\begin{aligned}
\frac{1}{\sqrt{n}} LR_n(\alpha) &= \frac{1}{\sqrt{n}} R_n(\alpha) + \frac{1}{\sqrt{n}} Q_n(\alpha) \\
&= \frac{1}{\sqrt{n}} R_n(\alpha) + Q(\alpha) + o_p(1)
\end{aligned}
\tag{43}
$$

where $o_p(1)$ holds uniformly in α. This equality states that the LR function is equal to the mean function plus a Gaussian process. The Gaussian process $Q(\alpha)$ is completely determined by its covariance function, which can be estimated from the data. The mean function $R_n(\alpha)$ is unknown. But it is possible to derive a bound for the LR test statistic from the fact that $R_n(\alpha) \leq 0$ for all α, if the null is true. We get then

$$
\frac{1}{\sqrt{n}} LR_n(\alpha) \leq \frac{1}{\sqrt{n}} Q_n(\alpha) \Rightarrow Q(\alpha).
\tag{44}
$$

Using this equation we can find a bound for the asymptotic distribution of the LR test of H_0 against H_1. Since $LR_n = \sup_\alpha LR_n(\alpha)$, we have as $n \to \infty$,

$$
P\left\{ \frac{1}{\sqrt{n}} LR_n \geq x \right\} \leq P\left\{ \sup_\alpha \frac{1}{\sqrt{n}} Q_n(\alpha) \geq x \right\} \to P\left\{ \sup_\alpha Q(\alpha) \geq x \right\}
\tag{45}
$$

However, the process $Q(\alpha)$ is not standardized. It can be shown that this test has true size that converges to zero as the sample diverges. In this form, the test is overconservative. Alternatively, we can standardize the LR function, so that all values of α yield the same variance. Moreover, we must account for nuisance parameters as they are present in most problems. Suppose the log likelihood is

$$
L_n(\beta, \gamma, \theta) = \sum_{i=1}^{n} l_i(\beta, \gamma, \theta).
\tag{46}
$$

The parameter γ and θ are nuisance parameters. θ is identified, γ is not identified under the null. This means that $L_n(0, \gamma, \theta)$ does not depend on γ. To apply the test, first θ must be eliminated. It is concentrated out of the likelihood function. Let $\alpha = (\beta', \gamma')$. Define the sequence of parameter estimates

$$
\widehat{\theta}(\alpha) = \max_{\theta \in \Theta} L_n(\alpha, \theta)
\tag{47}
$$

which represents the maximum likelihood estimate of θ for fixed values of α. The concentrated log likelihood function is then

$$
\widehat{L}_n(\alpha) = L_n(\alpha, \widehat{\theta}(\alpha)).
\tag{48}
$$

With analogous arguments as before a bound for the LR test statistic and for its distribution can be derived. To standardize the LR process we can use the sample variance:

$$
V_n(\alpha, \widehat{\theta}(\alpha)) = \sum_{i=1}^{n} q_i(\alpha, \widehat{\theta}(\alpha))^2
\tag{49}
$$

where

$$q_i(\alpha, \widehat{\theta}(\alpha)) = l_i(\alpha, \widehat{\theta}(\alpha)) - l_i(0, \gamma, \widehat{\theta}(0, \gamma)) - \frac{1}{n}\widehat{LR}_n(\alpha)$$

with LR function

$$\widehat{LR}_n(\alpha) = \widehat{L}_n(\alpha) - \widehat{L}_n(0, \gamma).$$

The standardized LR function is defined as

$$\widehat{LR}_n^*(\alpha) = \frac{\widehat{LR}_n(\alpha)}{V_n(\alpha)^{1/2}}, \tag{50}$$

and the standardized LR test statistic is

$$\widehat{LR}_n^* = \sup_{\alpha \in A} \widehat{LR}_n^*(\alpha). \tag{51}$$

Next define the centered stochastic process and his "large sample" counterpart as

$$\widehat{Q}_n^*(\alpha) = \frac{\widehat{Q}_n(\alpha)}{V_n(\alpha)^{1/2}}, \quad Q_n^*(\alpha) = \frac{Q_n(\alpha)}{V_n(\alpha)^{1/2}}, \tag{52}$$

assume additionally that $Q_n^*(\alpha)$ satisfies an empirical process law

$$Q_n^*(\alpha) \Rightarrow Q^*(\alpha) \tag{53}$$

where $Q^*(\alpha) = Q(\alpha)/V(\alpha)^{1/2}$ is a Gaussian process with covariance function

$$K^*(\alpha_1; \alpha_2) = \frac{K(\alpha_1; \alpha_2)}{V(\alpha_1)^{1/2}V(\alpha_2)^{1/2}}.$$

As before, we obtain

$$\begin{aligned}
\widehat{LR}_n^* &\leq \sup_{\alpha \in A} \widehat{Q}_n^*(\alpha) \\
&\leq \sup_{\alpha \in A} Q_n^*(\alpha) + o_p(1) \Rightarrow \sup_{\alpha \in A} Q^*(\alpha) \equiv \text{Sup } Q^*.
\end{aligned} \tag{54}$$

The following result then holds:

$$P\left\{\widehat{LR}_n^* \geq x\right\} \leq P\left\{\sup_{\alpha \in A} \widehat{Q}_n^*(\alpha) \geq x\right\} \to P\left\{\text{Sup } Q^* \geq x\right\}. \tag{55}$$

This gives a bound for the standardized LR test statistic, which is characterized by the distribution of the random variable Sup Q^*. This random variable is the supremum of the empirical process $Q^*(\alpha)$, which is characterized by its covariance function $K^*(\cdot)$. $K^*(\cdot)$ is unknown but we have the sample analogue

$$K_n^*(\alpha_1\alpha_2) = \frac{\widehat{K}_n(\alpha_1\alpha_2)}{V(\alpha_1)^{1/2}V(\alpha_2)^{1/2}}, \tag{56}$$

where

$$\begin{aligned}
\widehat{K}_n(\alpha_1\alpha_2) &= \sum_{i=1}^{n} \widehat{q}_i(\alpha_1)\widehat{q}_i(\alpha_2) + \sum_{k=1}^{M} w_{kM}\left[\sum_{1 \leq i \leq n-k} \widehat{q}_i(\alpha_1)\widehat{q}_{i+k}(\alpha_2) + \sum_{1+k \leq i \leq n} \widehat{q}_i(\alpha_1)\widehat{q}_{i-k}(\alpha_2)\right], \\
\widehat{q}_i(\alpha) &= q_i(\alpha, \widehat{\theta}(\alpha)).
\end{aligned}$$

$w_{kM} = 1 - |k|/(M+1)$ is the Bartlett kernel and M is a bandwidth number. If we can generate processes with covariance function $K^*(\cdot)$, then the supremum of each of these processes has approximately the distribution Sup Q^*.

Sample draws with covariance function $K^*(\cdot)$ can be generated by the construction of

$$\widetilde{LR}^*(\alpha) = \frac{\sum_{k=0}^{M} \sum_{i=1}^{n} q_i(\alpha, \widehat{\theta}(\alpha)) u_{i+k}}{\sqrt{1 + M V_n(\alpha)^{1/2}}}, \tag{57}$$

where $\{u_i\}_{i=1}^{n+M}$ is a draw of $\mathcal{N}(0,1)$ random variables.

To implement this test the following steps have to be done:

- Fix a grid for α (the parameters that become zero under the null and for p and q.)

- Maximize the likelihood function for the nuisance parameters.

- Compute the values $q_i(\alpha, \widehat{\theta}(\alpha))$.

- Compute the test statistic and its distribution.

Despite the implementation being rather easy, the computational burden becomes heavy when the number of parameters that depend on the regime increases. Moreover, this method provides a bound and not a critical value for the LR statistic. Using previous work of Hansen, Garcia[16] derives the asymptotic null distribution of the LR test and the related covariance functions for various Markov switching models. He defines

$$LR_n = 2n \left[Q_n(\widehat{\theta}, \widehat{\gamma}) - Q_n(\widetilde{\theta}) \right] \tag{58}$$

$$LR_n(\gamma) = 2n \left[Q_n(\widehat{\theta}(\gamma), \gamma) - Q_n(\widetilde{\theta}) \right] \tag{59}$$

where Q_n is the average log likelihood function of a sample of n observations:

$$Q_n(\theta, (\gamma) = \frac{1}{n} \log p(y_n, \dots, y_1; \theta, \gamma). \tag{60}$$

$\gamma = (p, q)$ and θ is the vector of parameters characterizing the Markov switching process. The first statistic represents the difference between the estimated unconstrained and constrained model. For the second the maximizing value of θ under the alternative is obtained for a given value of γ. Both are related by

$$LR_n = \sup_{\gamma \in \Gamma} LR_n(\gamma), \tag{61}$$

where Γ is a metric space from which the values 0 and 1 have to be excluded to keep the information matrix positive definite.

With this result, to test for the null of one against two regimes in Markov switching models has further been simplified. All we need here is to estimate the constrained and the unconstrained model for given values of γ, to compute the test statistic in (61). The critical values for the corresponding model can then be looked up in Garcia.

Part II

A selective survey

5 Markov switching models in macroeconomics

Hamilton[23] applies the introduced method to US real GNP. He uses quarterly data running from 51:2 to 84:4 and fits an AR(4)-process with potentially switching mean to 100 times the change in log real GNP:

$$(y_t - \mu_{s_t}) = \sum_{i=1}^{4} \phi_i(y_{t-i} - \mu_{s_{t-i}}) + \varepsilon_t.$$

It turns out that the estimated two means can be associated with the dynamics of the business cycle, one being negative (-0.4%) during state 0 and the other being positive (+1.2%) during state 1. Moreover, using a rule based on the full sample smoothed probabilities he can date turning points of business cycles that are quite in line with the chronology of the NBER. He even demonstrates that where the dating differs his is more in line with political events. Also, the expected duration of a recession is 4.1 quarter and that of an expansion 10.5 quarters, which are 4.7 and 14.3 quarters, respectively, according to the NBER. An interesting investigation follows the presentation of the results. He addresses the question why linear model seem to fit the data well even if the true process is the nonlinear Markov switching one. An AR(4) specification is fitted to Monte Carlo draws of a Markov switching series. It turns out that the average residual autocorrelation function would provide only negligible evidence against a linear AR(4) representation, even though the true model is in fact the nonlinear one. The forecasts of growth rates restricted to linear functions of past values will be suboptimal, however, because they will differ when we know wether the economy was in expansion rather than in recession last period. Finally, he derives the permanent effect of business cycle shocks on the long run level of output and finds that if in t the economy is in the recession state, the long run effect of this shock will be a 3% decline in GNP.

The problem that is not addressed is his work and in all the following ones is to find appropriate starting values for the parameters. Rather informally, the iterations are initialized with many combination of starting values. Those are chosen, that maximize the likelihood.

A bivariate application is found in Phillips[38] who evaluates the transmission of business cycles from one country to another. The results give evidence that worldwide shocks dominate the transmission of business cycles. For the research he uses quarterly time series in industrial production for the US, Canada, Germany and Great Britain. The data cover the period 62:2 to 83:3. He investigates the business cycles relationships between the US and each other country within a bivariate framework. Let y_t denote the

growth of output in period t:

$$y_t = n_t + \varepsilon_t$$
$$n_t = \mu_1 s_{1t} + \mu_2 s_{2t} + \mu_3 s_{3t} + \mu_4 s_{4t}$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad u_t \sim \mathcal{N}(0, \Sigma)$$

The last equation allows the error term to be vector-autocorrelated. For the two-country model, y_t, n_t, ε_t, u_t and each of the μ's are 2×1 vectors. As there can be a high growth or a low growth state in each country, the combination of these will define the four states of the Markov process. The four values for μ will then be:

$$\mu_1 = \begin{bmatrix} \mu_1^h \\ \mu_1^f \end{bmatrix}, \; \mu_2 = \begin{bmatrix} \mu_2^h \\ \mu_1^f \end{bmatrix}, \; \mu_3 = \begin{bmatrix} \mu_1^h \\ \mu_2^f \end{bmatrix}, \; \mu_4 = \begin{bmatrix} \mu_2^h \\ \mu_2^f \end{bmatrix},$$

with $\mu_1^c > \mu_2^c$, $c = h, f$. The superscripts h and f refer to home and foreign country, respectively. This set up is rather elegant as several assumptions about the business cycles correlation between the two countries are easily implemented by restricting appropriately the matrix of transition probabilities. The business cycles correlations are reflected in the Markov processes governing the state variable in each country. If the business cycles of two countries are independent, then the two Markov processes will be independent. In this case the matrix of transition probabilities for the bivariate model is not restricted. If the two Markov processes are perfectly correlated, the four-by-four matrix of transition probabilities reduces to a two-by-two matrix, because state 2 and 3 defined above will in fact never occur. It is also possible to account for the fact that one country may lead the other or to incorporate leads that last longer than one period. Each specification may be tested against the alternative of independence using a LR statistic.

The results give only weak evidence for the transmission of business cycles from one country to the other. Rather, recessions and recoveries seem to occur at the same time in all countries. Moreover, there is no evidence for the US leading or following into or out of recessions. Of course, as Phillips mentions himself, the results may change when considering the relationships between large industrialized and developing countries or between large and small country pairs. To get his results, he first estimates the model with a bunch of starting values and a weak convergence criterion. In a second round he takes the parameters that maximized the likelihood as starting values and estimates the model again with a strong convergence criterion.

Goodwin[20] estimates the Hamilton model for eight countries (G-7 and Switzerland) using quarterly data. To get the parameter estimates he first defines a grid for the state-dependent mean and the transition probabilities. Next the likelihood is maximized with respect to the nuisance parameters for each fixed gridpoint. The parameters yielding the maximum of the likelihood are then used as starting values to find the optimum of all the

parameters jointly. To assess the improvement a Markov switching model may bring over a linear AR(4) model a bunch of specification and forecasting tests are performed. The most striking result is that for all countries the null of one regime against the alternative of two cannot be rejected using the Hansen standardized LR test. Moreover, although the Markov switching model has more explanatory power than the linear model, it cannot explain all the nonlinearities in the data. Goodwin concludes that the Hamilton model brings only a marginal improvement over linear models for output series. He uses smoothed probabilities to date turning points of economic activity. The chronologies, except for Italy, are quite in line with that of the NBER for the US and with the dating of previous studies for other countries. For Italy, some outliers are recognized as the high-growth state relegating all other observations to the low-growth state. As dating is not very sensible in this case, it is left apart for this analysis.

Ghysels[17] presents a Markov switching framework to assess for the periodic structure of the business cycle. To test wether recoveries can begin in each month of the year with equal probability, he relies on LR, Lagrange multiplier (LM) and Wald tests. Using data for the US he finds evidence that spring months and December are most favourable for recoveries to begin and that the lengths of recessions or booms are different depending on which period in the year the turning point occurred. According to his results the business cycle then displays a seasonal pattern. In the model the matrix of transition probabilities is quarter or month dependent:

$$
\begin{array}{c}
t+1 \\
\begin{array}{ccc}
 & \text{Expansion} & \text{Recession} \\
\text{Expansion} & \sum_{i=1}^{s} d_{it}p_i & 1-\sum_{i=1}^{s} d_{it}p_i \\
\\
\text{Recession} & 1-\sum_{i=1}^{s} d_{it}q_i & \sum_{i=1}^{s} d_{it}q_i,
\end{array}
\end{array}
$$

where d_{it} is a seasonal dummy process. However, this process can be brought into a time-independent form when expanding appropriately the number of states. For the analysis, Ghysels derives the likelihood function of the transition probabilities. Their maximum likelihood estimates are based on two chronologies of business cycle turning points, one dated by the NBER and the other by Romer[40]. The null of equal transition probabilities across seasons and equal transition probabilities for either expansions or recessions are rejected, giving evidence for a periodic structure of the business cycle.

Observing that in the long run there are equilibrium relationhips between economic variables, but that these relationships may change as the economic environment changes, leads Hall, Psaradakis and Sola[22] to allow for time-varying cointegration. Thereby, the long run parameters are allowed to switch stochastically between two different cointegration regimes. They illustrate their approach using quarterly, seasonally not adjusted Japanese data for real total consumption and real disposable income from 61:1 to 87:4. They find considerable evidence in favour of a time-varying cointegration model, in which the parameters of a cointegration regression depend on two regimes. Let c_t and

y_t represent consumption and disposable income in t, respectively, define D_t^j a dummy variable for season j, then the model fitted to the data is:

$$c_t = (\beta_0 + \beta_1 s_t)y_t + (\mu_0 + \mu_1 s_t) + \sum_{j=1}^{3}(\alpha_j^0 + \alpha_j^1 s_t)D_t^j + [\omega_0(1 - s_t) + \omega_1 s_t]u_t,$$

where $\{u_t\}$ is a stochastic sequence with zero mean and variance of unity. In their analysis the residual variance is also dependent on the regime. Their specification tests are based on the standardized residuals of the Markov switching cointegration regressions. Using the augmented Dickey-Fuller test, the test of Phillips and Ouliaris and the Durbin-Hausmann DHS test of Choi, the null of no cointegration can firmly be rejected in all cases.

The ability of the Markov switching model to account for the changing pattern of economic variables over the business cycle is combined in Diebold and Rudebusch[6] with a dynamic factor model. The latter models can additionally describe the correlation of macroeconomic variables over the business cycle. The number of driving forces in the economy is often less than the number of variables affected by these forces, implying a factor structure for the variables. Collect n variables in a $n \times 1$ vector X_t. The factor regime switching model can then be set up:

$$\begin{aligned} \Delta X_t &= \beta + \lambda f_t + u_t \\ D(L)u_t &= \varepsilon_t \\ \phi(L)(f_t - \mu_{s_t}) &= \eta_t \end{aligned}$$

Movements in n macroeconomic variables are described by the one-dimensional non-observable common factor f_t and by the n-dimensional idiosyncratic component u_t. The common factor potentially switches regime depending on the state of a latent variable s_t following a two-state, first-order Markov process. Collect the relevant past for the factor in $z_t = (s_t, s_{t-1}, \ldots, s_{t-p}, f_{t-1}, \ldots, f_{t-p})'$. Then, the conditional density of f_t given z_t is

$$p(f_t|z_t; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-\left((f_t - \mu_{s_t}) - \sum_{i=1}^{p} \phi_i(f_{t-i} - \mu_{s_{t-1}})\right)^2}{2\sigma^2}.$$

The common factor follows then an autoregressive process of order p with a potentially switching mean. The results presented by Diebold and Rudebusch are based on four quarterly time series of coincident indicators, namely personal disposable income minus transfer payments, index of industrial production, manufacturing and trade sales, number of persons on non-agricultural payrolls. They use the composite index of coincident indicators computed by the Commerce department as common factor. The time series run from 52:1 to 93:1. They estimate a first-order autoregressive process with switching mean for each series. A full factor regime switching model is not estimated. However, the similarity of the results across the series makes it very plausible. The two means capture business cycle patterns for all series. μ_0 is negative and μ_1 is positive, representing recession or recovery periods, respectively. Moreover, using critical values from

Garcia[16], the null of one against two regimes can be rejected significantly in each case. The estimated transition probabilities suggest that there is a persistence to remain in the prevailing state and that expansions last longer than recessions. The evidence of these preliminary results is reinforced in Kim and Nelson[30], who estimate the full factor model for four monthly coincident indicators by Gibbs sampling. They find positive duration dependence for recessions and recoveries, and further that the effect is concentrated in the months before the switches in regime occur.

We did the same investigation as Diebold and Rudebusch for three quarterly, seasonally adjusted time series of macroeconomic data for the G-7 countries plus Austria and Switzerland. The sample runs from 70:1 to 94:4. We took the series for real GNP or GDP, real consumption and real investment as economic theory suggests that these three are driven by a common stochastic trend. This common I(1)-factor (depicted in figure 2) is estimated by the principal component that explains most of the variation in the variables (see Stock and Watson[44] who use this approach to test for common trends). Taking these series we will also be able to partly compare our results with those of Goodwin[20]. To restrict the analysis we estimated a Markov switching model only for the common factor. Table 1 displays the results. Except for Austria no AR(1) term was included. For Switzerland, Italy, Japan and the US the estimation of an autoregressive first-order model with switching mean yields no plausible results, for Canada, Germany, France and Great Britain the inclusion of an AR(1) term brings no significant improvement in the inference. Worth mentioning is also that the data for Austria need special treatment to get sensible results. In 1972 the general sales tax system was replaced by the value added tax system. Later on, in 1977 the value added tax rate for expensive (mostly imported) goods was increased above the normal rate (*Luxussteuer*). Correspondingly, we observe outliers in the changes of the principal component in the third and fourth quarters of these years. We account for this with dummy-variables. It is only after this transformation that we can obtain usable results for Austria. Moreover, an AR(1)-term must be included in the analysis as no maximum is found when omitting it. For all countries but France the two means capture business cycles features. There is a considerable persistence to remain in the prevailing state, the persistence being higher for the positive (France: for the high) growth state. Moreover, using the critical values in Garcia[16] we can reject the null of one against two regimes at the 1% significance level for all countries except Austria and Great Britain. We further dated business cycles turning points as Goodwin[20] did in his paper. We designate date t a peak if $P(s_t = 2|\mathcal{Y}_T) > 0.5$ and $P(s_{t+1} = 2|\mathcal{Y}_T) < 0.5$. Similarly, a date t is designated a trough if $P(s_t = 2|\mathcal{Y}_T) < 0.5$ and $P(s_{t+1} = 2|\mathcal{Y}_T) > 0.5$. The dating is based on the smoothed probabilities estimated by the filter. The probabilities of being in state 1 at date t are shown in figure 3 and 4. The chronologies are shown in table 2. The column labeled FS refers to our dating using the common factor of real GNP or GDP, real consumption and real investment, the others are reproduced from Goodwin. Our dating closely follows previous ones. Using the common factor rather than GDP even improves the dating

for two countries. In the case of Switzerland, the inference of Goodwin reproduced only the recession in the middle of the 70ies. The recession at the beginning of the 80ies can be reproduced when using the common factor. Regarding Italy, a fourth-order autoregressive Markov-switching model for GDP identified two outliers as the high growth state. A dating is not very sensible in this case and is left out in Goodwin. Again, using the common factor yields plausible results and we can also date business cycle turning points in this case. Finally, Austria seems to have undergone an additional business cycle in the middle of the 80ies. Again, this may be a reflection of changes in tax policy, namely a change in the value added tax rates.

The conclusions we draw from these preliminary results are mixed. We can make no general assessment of the appropriateness of the Markov switching model for the principal component of real GDP (GNP), consumption and investment. For most countries investigated the null of one against two regimes is rejected, the exceptions being Austria and Great Britain. To improve the evidence given in this paper, in future research the alternatives under test must be generalized (e.g. include more autoregressive terms), so that the null is able to capture more of the dynamic patterns displayed in the data. Moreover, it remains to assess the usefulness of Markov switching models in judging the economic situation prevailing in a country. Our results suggest that the inferred smoothed probabilities can be used to date business cycles turning points. The dating closely follows previous ones and is more in line with observed economic evolution for Italy and Switzerland when compared to the single time series analysis of Goodwin[20].

6 Markov switching models in finance

One of the areas where Markov switching (MS) has been applied successfully is in financial econometrics. The main motivation for this lies in the stylized facts of financial time series. So before outlining the specifications of MS used in finance we will briefly summarize the most important stylized facts. We use the returns of the \$/DM exchange rate to illustrate the stylized facts. Our sample of daily observations starts in January 1980 and ends in December 1992. This series has already been extensively studied, e.g. in Hsieh (1989) but here it serves as a benchmark.

6.1 Stylized facts for financial time series

In figure 5 we plot the (compound) returns. The first question to ask is whether there is significant dependence in the levels of the series. The idea that future returns cannot be predicted by information available today is the main content of the Efficient Markets Hypothesis (EMH). The benchmark model for the EMH is the random walk (here without drift):

$$\log P_t = \log P_{t-1} + u_t, \quad u_t \sim i.i.d. \ \mathcal{N}(0, \sigma^2) \tag{62}$$

A weaker formulation is obtained under the martingale assumption which leads to the return process (see Spanos[43]):

$$E(\log P_t | I_{t-1}) = \log P_{t-1}. \tag{63}$$

The difference between these two specifications lies in the restrictions on higher moments: The random walk requires independence whereas the martingale imposes the much weaker condition of no autocorrelation in the first moment. This means that e.g., a process which is uncorrelated but has a time-changing variance (heteroscedasticity) is rejected under the random walk model ot the EMH and is not rejected under the martingale model. For our daily series the autocorrelation for 100 lags is plotted in figure 6. There is only little autocorrelation (autocorrelation at lag one is 0.025 with a p-value of .21) and so the linear dependence is not very pronounced. Linear dependence is the simplest form of return predictability, namely by using past returns. Later on we will study nonlinear dependence, which is very important for MS.

Besides the EMH a central topic in work on the behavior of financial time series is the question of the distribution of returns and how to measure the volatility of returns. This is important for the specification of any model of returns and it has wide-ranging consequences in empirical work. Examples reach from event studies in corporate finance to risk management or the pricing of complex derivative products. The benchmark model is the time-independent Gaussian as given below:

$$r_t \sim \mathcal{N}(\mu, \sigma^2) \tag{64}$$

In 64 the variance or the standard deviation can be interpreted as the "volatility" of returns. However the literature offers a variety of definitions for the term volatility. Here we only quote the description of Engle[11, p. 73]: "Volatility measures the variability of returns." Depending on the information set we differentiate between conditional volatility ($= h_t$) and unconditional volatility ($= \sigma^2$), defined again from Engle[11]:

$$h_t = E\left[(r_t - \mu_t)^2 | I_{t-1}\right] \tag{65}$$

$$\sigma^2 = E[r_t - \mu_t]^2 \tag{66}$$

$$\text{where } \mu_t = E[r_t | I_{t-1}]$$

Here μ_t is the conditional mean, which can also contain exogenous variables and I_{t-1} is the information set known at time $t - 1$ which is used to compute the conditional expectations. Most of the time it contains past returns. The conditional model always nests the unconditional one, but not vice versa. As we will see this difference is very important in the context of MS.

Since the 1960's there has been growing doubt on the validity of the time-independent Gaussian model. In particular the finding of *fat tails* and *volatility clustering*, both discussed below are contrary evidence.

Mandelbrot[33, p. 418] finds that: "Large changes tend to be followed by large changes – of either sign – and small changes tend to be followed by small changes". This finding is evident for the \$/DM when turning again to figure 5. Here we see that high and low volatilities, i.e. times with variable returns "cluster" together. Once there are large price movements in the market these "high-volatility regimes" are persistent for some time. Volatility clustering, also known as nonlinear dependence implies that the variance is not constant over time, i.e. there is heteroscedasticity. In the simplest case we could imagine that there are two periods: Times with high and low variance respectively. A consequence of volatility clustering is that the variance can be predicted reasonably well. The reason is that the variance remains high for some time once some extreme price moves have taken place. This hypothesis can be analyzed by estimating the autocorrelations in the squares: here any predictability in the squares should show up as high autocorrelations. For the \$/DM we illustrate these findings in figure 6: There we plot the autocorrelation of the squared time series. The values are all very high and they remain significant at least until lag ten. On this basis a simple model to forecast volatility would use an autoregression on squared returns. However as we will demonstrate MS offers a superior approach. Figure 6 also illustrates that the autocorrelation in the squared returns is higher than in the levels. This means that compared to the pronounced heteroscedasticity there is relatively little structure in the first moments of financial time series. The consequence is that in the process of modelling returns more weight is put on finding an appropriate specification for the second moment (conditional variance and covariance between assets) instead of the first moment (conditional mean).The dependence in the second moments implies that the random walk model is rejected, but under the martingale model there is no contradiction to the EMH.

The finding of volatility clustering is strongly related to the phenomenon of fat tails. Fama[13] studies daily time series of the thirty stocks in the Dow Jones Industrial Average index. He finds that (p. 48): "In every case the empirical distributions are more peaked than the Normal in the centre and have longer tails than the normal distribution." This means that the empirical distribution allocates the probability mass differently from the Normal. We see this result for the \$/DM series in figure 7: It plots a nonparametric estimate of the normalized unconditional distribution of the returns (FHAT). For a comparison the figure also plots the $\mathcal{N}(0,1)$ distribution. The probability of extreme positive or negative price movements is much larger than in the Normal model, thus the tails are "fatter".

It is important to recognize that volatility clustering and fat tails are two related phenomena. The reason is that a series where the variance changes over time can not be drawn from a stationary Normal and so fat tails are the static representation of the dynamic phenomenon of time-dependence in second moments.

6.2 The MS model in finance

Given these stylized facts the literature offers two types of models to explain them: MS and GARCH[6]. MS assumes there are two "regimes" or states. State one has a high mean and low volatility and state two has a low (negative) mean and high volatility, the latter one being less likely, i.e. it has a lower unconditional probability of being observed. This specification can be interpreted as a mixture distribution with dynamics generated by a Markov chain. In contrast to the mixture distribution studied in e.g. Kon[31] the returns generated by the MS model are not independent. So it is a considerable improvement over the restrictive static mixture-of-normals model as given in 67.

$$
\begin{aligned}
r_t &\sim N(\mu_1, \sigma_1^2) \quad \text{with probability } p_j \\
r_t &\sim N(\mu_2, \sigma_2^2) \quad \text{with probability } p_i = 1 - p_j
\end{aligned}
\tag{67}
$$

The probability density function for two regimes is:

$$
f(r_t) = \sum_{j=1}^{2} p_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2\sigma_j^2} \left(r_t - \mu_j \right)^2 \right]
\tag{68}
$$

In this model the Gaussian distribution is generalized by introducing two regimes with different moments. Leptocurtosis is obtained here because the variances in the two regimes differ. In Tucker[46] generalizations to more than two regimes are studied. However the result there is that out of 200 US stocks 172 have 2 or 3 regimes only. This model is estimated by maximum likelihood techniques.

MS can capture all the stylized facts of section 6.1: The mixture distribution generates the leptocurtosis and the Markov chain is responsible for the nonlinear dynamics. Engel and Hamilton[9] are the first to apply MS to financial time series. They study a sample of quarterly returns, finding that MS gives a good fit. They also apply a variety of tests to analyze the performance of their model. Applications of MS to stock market returns started with Pagan and Schwert[37]. There MS is compared with GARCH and several other volatility models. The best forcasting performance is achieved by nonparametric models. McQueen and Thorley[34] specify a MS model where the two states differ only in their means but not in their variances. This model is used for tests of the Efficient Markets Hypothesis represented as a random walk. Town[45] applies MS to data from corporate finance: He studies evidence for merger waves in the US. Schmitt[41] repeats the exercise of Pagan and Schwert for several return series from the German stock market. Results are mixed with some evidence in favor of GARCH. Rockinger[39] estimates a switching regression model for the main French stock index. He introduces innovations in macroeconomic time series as explanatory variables. Sola and Timmerman[42] estimate GARCH, EGARCH and MS for daily stock returns from the London Stock Exchange. In an in-sample comparison MS beats GARCH because it can generate skewness in addition to volatility clustering and fat tails which both models share. With a

---

[6]Recently stochastic volatility models have also become popular, see the survey by Ghysels et al.[18].

simulation and application to daily data Sola and Timmerman also show that GARCH and MS are "observationally equivalent" for generating volatility clustering. Van Norden and Schaller[47] also apply the MS model to stock returns. They study possible causes for stock market crashes and test the hypothesis of speculative behavior against the hypothesis of changes in fundamentals. Their result is that the two explanations are complements rather than substitutes. Engel[8] studies the forecasting performance of MS for FX markets, obtaining mixed results. The predictive capabilities of MS are also studied in Dacco and Satchel[4]. Recently Bekeart and Harvey[1] have introduced a model which manages to combine MS with multivariate ARCH. They study evidence for time-dependence in the integration of emerging markets into the "world market".

The MS model as it is specified for financial econometrics is defined as in 69 or 70, with S_t indicating the state characterized by the value 0 or 1:

$$\text{State } S_t = i: \ r_t|S_t \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right), \ i = 0, 1; \tag{69}$$

$$r_t = \alpha_0 + \alpha_1 S_t + [w_o + w_1 S_t]\, \varepsilon_t, \ \text{with } \varepsilon_t \sim i.i.d. \ \mathcal{N}(0,1). \tag{70}$$

The conditional mean and variance are then defined as:

$$\mu_t = \alpha_0 + \alpha_1 S_t \quad \sigma_t = w_0 + w_1 S_t. \tag{71}$$

The transition (**conditional**) probabilities are:

$$\begin{aligned} Prob(S_t = 1|S_{t-1} = 1) = p \qquad Prob(S_t = 0|S_{t-1} = 1) = 1 - p \\ Prob(S_t = 1|S_{t-1} = 0) = 1 - q \quad Prob(S_t = 0|S_{t-1} = 0) = q. \end{aligned} \tag{72}$$

The Markov behavior of the probabilities allows the following representation of S_t:

$$S_t = (1 - q) + (p + q - 1)S_{t-1} + v_t \tag{73}$$

where conditional on $S_{t-1} = 1 : v_t = 1 - p$ with probability p and $v_t = -p$ with probability $1 - p$ and conditional on $S_{t-1} = 0 : v_t = -(1 - q)$ with probability q and $v_t = q$ with probability $1 - q$.

From 73 an AR(1) representation of the conditional variance σ_t^2 can be derived:

$$\sigma_t^2 = (p + q - 1)\,\sigma_{t-1}^2 + (1 - p + 1 - q)\,w_0 + (1 - q)\,w_1 + v_t \tag{74}$$

In this context the MS model can also be interpreted as a stochastic volatility model. The main difference between the two classes of models is that the innovations in the variance equation, v_t, are a discrete random variable with four states as defined in equation 73. The ergodic (**unconditional**) probabilities are:

$$\begin{aligned} Prob\,(S_t = 1) \ &= \ \pi = \frac{1 - q}{(1 - p) + (1 - q)} \tag{75} \\ Prob(S_t = 0) \ &= \ 1 - \pi. \tag{76} \end{aligned}$$

The filter probabilities are: $p_t^F = p\left(S_t | r_t, ..., r_0\right)$.

The parameter vector $\theta = \left(\mu_1, \sigma_1^2, \mu_0, \sigma_0^2, p, q\right)$ is computed by numerical optimization of the log likelihood function[7]. This is a by-product of the filter iteration which estimates the filter probabilities.

As in the GARCH case we need the standardized residuals for the diagnostic tests. Their computation is performed in two steps[8]: First, the one-period forecast based on the parameter vector θ is:

$$E\left[r_{t+1} | r_{t...}, \theta\right] = \mu_0 + \left\{\pi + (-1 + p + q)\left[p_t^F - \pi\right]\right\}\left\{\mu_1 - \mu_0\right\}. \tag{77}$$

Then the conditional variance is computed as:

$$\sigma_{t+1|t}^2 = (\mu_1^2 + \sigma_1^2)p_{t+1|t} + (\mu_0^2 + \sigma_0^2)(1 - p_{t+1|t}) - \left[\mu_1 p_{t+1|t} + \mu_0(1 - p_{t+1|t})\right]^2. \tag{78}$$

with $p_{t+1|t} = (1 - q) + (-1 + p + q)\, p_t^F$.

Finally the standardized residuals are:

$$u_{t+1} = \frac{r_{t+1} - E\left(r_{t+1} | r_{t...}, \theta\right)}{\sigma_{t+1|t}} \tag{79}$$

The results are displayed in table 3. The 2-state MS model generates one regime with high volatility and negative mean (less likely) and one with low volatility and positive mean. The means are not significant. Further information on the performance of this model is gained from a plot of the filter probabilities of being in state 1 (see figure 8). There are frequent changes between the two states. This shows that MS can differentiate between the two states. State one can be interpreted as the "leverage effect", i.e. an asymmetry in variance. We will analyze the results of the test statistics by contrasting them with the results for a GARCH model. As we mentioned above, GARCH is the main competitor to MS.

6.3 Comparison of MS to GARCH

The basic ARCH model was introduced by Engle[10]. The literature is very comprehensive, for a survey see e.g. Bollerslev et al.[3]. As MS GARCH can also explain the stylized facts for financial time series. The parameterization for a standard GARCH (1,1) is given below. The residuals are drawn from a conditional Student's t distribution. The reason is that in many cases the volatility clustering does not cause all the nonnormality found in the data. This extension has been introduced by Bollerslev[2].

$$r_t = c + u_t \tag{80}$$

$$u_t = \varepsilon_t \sqrt{h_t} \qquad \varepsilon_t \sim t_N \tag{81}$$

$$h_t = a_0 + a_1 u_{t-1}^2 + a_2 h_{t-1} \tag{82}$$

---

[7]The code for estimating the model was kindly provided by Prof. Hamilton.
[8]See Engel and Hamilton[9]

The MS model gives the highest information criterion (See table 4). Its standardized residuals are much closer to the normal distribution than those from GARCH models. However the GARCH model gives a better fit for the volatility clustering: Its squared standardized residuals show no autocorrelation whereas there seem to be some omitted dynamics in MS residuals. Thus for this sample the question whether MS is superior to GARCH can not be clearly answered, as the choice depends on the performance criterion. When using out-of-sample forecasting experiments MS is frequently dominated by GARCH, see Pagan and Schwert[37]. As we saw here, for in-sample studies MS may be equal to GARCH depending on the selection criterion.

6.4 Extensions of MS models

Hamilton and Susmel[26] show that GARCH models are very persistent, i.e. the coefficients are close to the nonstationary (integrated GARCH) case. Despite this behavior the GARCH models do not give a satisfying performance for volatility forecasts. They propose the following specification to achieve improved forecasts:

$$r_t = c_0 + c_1 r_{t-1} + u_t \qquad S_t = 1 \text{ or } 2 \tag{83}$$

$$u_t = \sqrt{g(S_t)}\widetilde{u}_t \tag{84}$$

Here $g(S_t)$ is the state-dependent multiplier with $g(S_t = 1) = 1$.

$$\widetilde{u}_t = h_t v_t \tag{85}$$

$$v_t \sim t_N(0,1) \tag{86}$$

$$h_t^2 = a_0 + a_1 \widetilde{u}_{t-1}^2 \tag{87}$$

Matrix of transition probabilities between states:

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \tag{88}$$

This model, known as switching ARCH (SWARCH) is a useful extension to the ARCH model. The main addition here is that the conditional variance is multiplied by a state-dependent factor. As in the simple MS model, the state probabilities are driven by a Markov chain. Additionally this model is specified conditional on a Student's t distribution with N degrees of freedom instead of the Normal. The estimation of this model is performed with the EM-Algorithm. Hamilton and Susmel[26] find that the model is superior to GARCH when it comes to forecasting performance. However, when we tried to estimate this model for the $/DM rate, we could not achieve a regular convergence for any set of starting values which we tried. The reason is that the starting values for the transition probabilities and the state-dependent factor have to be specified very close to the final values for the algorithm to converge. However we found no method to get these close estimates as also various search procedures proved to be of no help.

The SWARCH model has been generalized by Gray[21]. This model allows for a GARCH component in the regime-dependent conditional variance and also for state-dependent transition probabilities.

References

[1] Geert. Bekaert and Campbell R. Harvey. Time-varying world market integration. *Journal of Finance*, pages 403–444, 1995.

[2] Tim Bollerslev. A conditional heteroscedastic time series model for speculative prices and rates of regurn. *Review of Economics and Statistics*, 69:542–547, 1987.

[3] Tim Bollerslev, Ray Chou, and Kenneth Kroner. Arch modelling in finance. *Journal of Econometrics*, pages 5–59, 1992.

[4] Roberto Dacco and Stephen Satchell. Why do regime switching models forecast so badly? Working Paper FE-7, Birkbeck College, University of London, 1995.

[5] Francis X. Diebold, Joon-Haeng Lee, and Gretchen C. Weinbach. Regime switching with time-varying transition probabilities. In Colin P. Hargreaves, editor, *Nonstationary Time Series Analysis and Cointegration*, pages 283–302. Oxford University Press, 1994.

[6] Francis X. Diebold and Glenn D. Rudebusch. Measuring business cycles: A modern perspective. Working Paper 4643, NBER, 1994.

[7] A. Dold and B. Eckmann. *Empirical Distributions and Processes*. Springer-Verlag, 1976.

[8] Charles Engel. Can the markov switching model forecast exchange rates? *Journal of International Economics*, pages 151–165, 1994.

[9] Charles Engel and James D. Hamilton. Long swings in the Dollar: Are they in the data and do markets know it? *The American Economic Review*, 80:689–713, 1990.

[10] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50:987–1007, 1982.

[11] Robert F. Engle. Statistical models for financial volatility. *Financial Analysts Journal 1993*, pages 72–78, 1993.

[12] Barry Falk. Further evidence on the asymmetric behavior of economic time series over the business cycle. *Journal of Political Economy*, 94:1096–1109, 1986.

[13] Eugene Fama. The behavior of stock market prices. *Journal of Business*, pages 34–105, 1965.

[14] Peter Gaenssler and Winfried Stute. Empirical processes: A survey of results for independent and identically distributed random variables. *The Annals of Probability*, 7:193–243, 1979.

[15] Peter Gaenssler and Stute Winfried. *Seminar on Empirical Processes*. Birkhäuser Verlag, 1987.

[16] René Garcia. Asymptotic null distribution of the likelihood ratio test in Markov switching models. Cahier 1095, Université de Montréal, Centre de Recherche et développement en économique, 1995.

[17] Eric Ghysels. On the periodic structure of the business cycle. *Journal of Business & Economic Statistics*, 12:289–298, 1994.

[18] Eric Ghysels, Andrew C. Harvey, and Eric Renault. Stochastic volatility. mimeo, 1995.

[19] Stephen M. Goldfeld and Richard E. Quandt. A Markov model for switching regressions. *Journal of Econometrics*, 1:3–16, 1973.

[20] Thomas H. Goodwin. Business-cycle analysis with a Markov-switching model. *Journal of Business & Economic Statistics*, 11:331–339, 1993.

[21] Steven Gray. Modelling the conditional distribution of interest rates as a regime-switching process. mimeo, Duke University, 1995.

[22] Stephen Hall, Zacharias Psaradakis, and Martin Sola. Changes in regime, cointegration, seasonality, and the Japanese consumption function. Discussion Paper in Economics 21/94, Birkbeck College, University of London, 1994.

[23] James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 1989.

[24] James D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45:39–70, 1990.

[25] James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[26] James D. Hamilton and Raul Susmel. Arch and changes in regime. *Journal of Econometrics*, pages 307–333, 1994.

[27] Bruce E. Hansen. The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. *Journal of Applied Econometrics*, 7:S61–S82, 1992.

[28] Andrew C. Harvey. *Time Series Models*. Philip Allan, 1981.

[29] Chang-Jin Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60:1–22, 1994.

[30] Chang-Jin Kim and Charles R. Nelson. Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime-switching. mimeo, Dept. of Economics, University of Washington, Box Seatlle, WA, 98195,USA, October 1995.

[31] Stanley Kon. Models of stock returns. *Journal of Finance*, pages 147–165, 1984.

[32] Pok-sang Lam. The Hamilton model with a general autoregressive component. *Journal of Monetary Economics*, 26:409–432, 1990.

[33] Benoit Mandelbrot. The variation of certain speculative prices. *Journal of Business*, pages 394–419, 1963.

[34] Grant McQueen and Steven Thorely. Are stock returns predictable: A test using markov chains. *Journal of Finance*, pages 239–263, 1991.

[35] Salih N. Neftçi. Are economic time series asymmetric over the business cycle? *Journal of Political Economy*, 92:307–328, 1984.

[36] Adrian Pagan. The econometrics of financial markets. *Journal of Empirical Finance*, pages 15–102, 1996.

[37] Adrian Pagan and William Schwert. Alternative models for conditional stock volatility. *Journal of Econometrics*, pages 267–290, 1990.

[38] Kerk L. Phillips. A two-country model of stochastic output with changes in regime. *Journal of International Economics*, 31:121–142, 1991.

[39] Michael Rockinger. Switching regressions of unexpected macroeconomic events explaining the french stock index. mimeo, HEC Paris, 1994.

[40] Christina D. Romer. Remeasuring business cycles: A critique of the prewar NBER dates. Discussion paper, National Bureau of Economic Research: Cambridge, MA, 1991.

[41] Christian Schmitt. Volatilit tsprognosen. Working Paper 7, ZEW, 1994.

[42] M. Sola and A. Timmerman. Fitting the moments: A comparison of arch and regime switching models for daily stock returns. mimeo, 1994.

[43] Aris Spanos. *Statistical foundations of econometric modelling.* Cambridge University Press, 1987.

[44] James H. Stock and Mark W. Watson. Testing for common trends. *Journal of the American Statistical Association*, 83:1097–1107, 1988.

[45] Robert J. Town. Merger waves and the structure of m& a time series. In Pesaran M.H. and Potter S., editors, *Nonlinear Dynamics, Chaos and Econometrics.* John Wiley, 1993.

[46] Alan Tucker. A reexamination of finite- and infinite variance distributions as models of daily stock returns. *Journal of Business & Economic Statistics*, pages 73–81, 1992.

[47] Simon van Norden and Huntley Schaller. Regime switching in stock market return. mimeo, 1993.

[48] Anders H. Westlund and Sven Oehlén. On testing for symmetry in business cycles. *Empirical Economics*, 16:479–502, 1991.

A Tables

Table 1: Estimated Markov switching models for principal components

	Start	AT	CA	CH	DE	FR	GB	IT	JP	US
μ_1	0.0	-0.12	-3.19	-1.77	-0.66	0.06	-0.56	-0.70	-1.24	-1.66
μ_2	0.5	1.88	2.14	1.46	1.58	1.70	1.88	1.49	2.15	2.05
ϕ_1	0.0	-0.27								
σ^2	1.0	4.36	4.07	2.55	1.53	1.06	4.40	1.23	3.16	3.00
p_{11}	0.69	0.80	0.71	0.86	0.85	0.89	0.88	0.81	0.88	0.78
p_{22}	0.69	0.92	0.97	0.96	0.95	0.93	0.92	0.95	0.97	0.94
LR		2.28	21.99*	29.62*	19.83*	18.28*	7.23	23.39*	19.71*	24.57*

LR denotes the likelihood ratio test statistic for the null hypothesis of one against two regimes. * denotes significance at the 1% level using the critical values from Garcia[16]

Table 2: Peaks and troughs

	US MS(4)	US NBER	US FS	UK MS(4)	UK FS	DE MS(4)	DE FS	JP MS(4)	JP FS	CA MS(4)	CA FS	CH MS(4)	CH FS	FR MS(4)	FR FS	AT FS	IT FS
P	69:3	69:4															
T	70:4	70:4															
P																	
T																	
P	73:4	73:4	73:2	73:4	73:1	73:4	73:2	73:2	73:4				73:2				
T								74:1					73:4				
P												74:2	74:2	74:3	74:1	74:1	74:1
T	75:1	75:1	75:1	75:2		75:2	75:1		75:1			76:1	76:2	75:1	75:1	75:2	75:2
P																77:4	
T					77:1											78:1	
P	80:1	80:1	79:3	79:2	79:2	80:1	80:1								80:1	80:1	80:2
T	80:3	80:3	80:2	81:2	81:4												
P	81:1	81:3	81:2							81:2	81:2		81:3				
T	82:4	82:4	82:3			82:4	82:3			82:3	82:3		82:2		84:4	82:4	82:4
P																83:4	
T																84:2	
P			90:1		89:2		92:1		91:3		90:1		90:3		90:1	92:1	92:1
T			91:2		92:4		93:4				90:4		93:1		94:1	93:1	93:3

Table 3: Estimation results of Markov switching model

	State 0	State 1
Means	0.0118 (.92)	-0.0228 (-.76)
Variances	0.2549 (16.54)	0.9211 (16.67)
Matrix of Markov transition probabilities	0.9708	0.0450
	0.0291	0.9549
Ergodic probs	0.6073	0.3926
SIC	-419.79	
(24)	39.51 [0.020]	
QQ(24)	60.72 [0.000]	
Skewness	-0.079 [0.060]	
Kurtosis	0.3378 [0.000]	
(t-statistics) [p-values]		

Table 4: Estimation results of GARCH model

Variable	Estimate	T-Stat [p-value]
1. c	0.0059	0.55
2. a_0	0.0164	3.74
3. a_1	0.077	6.36
4. a_2	0.895	55.21
6. N	5.85	8.65
SIC	-1563.81	
Q(24)	42.95	[0.001]
QQ(24)	25.66	[0.141]
Skewness	-0.0425	[0.311]
Kurtosis	1.5018	[0.000]

B   Figures

Figure 1:

UNITED STATES
first differences of log real GDP
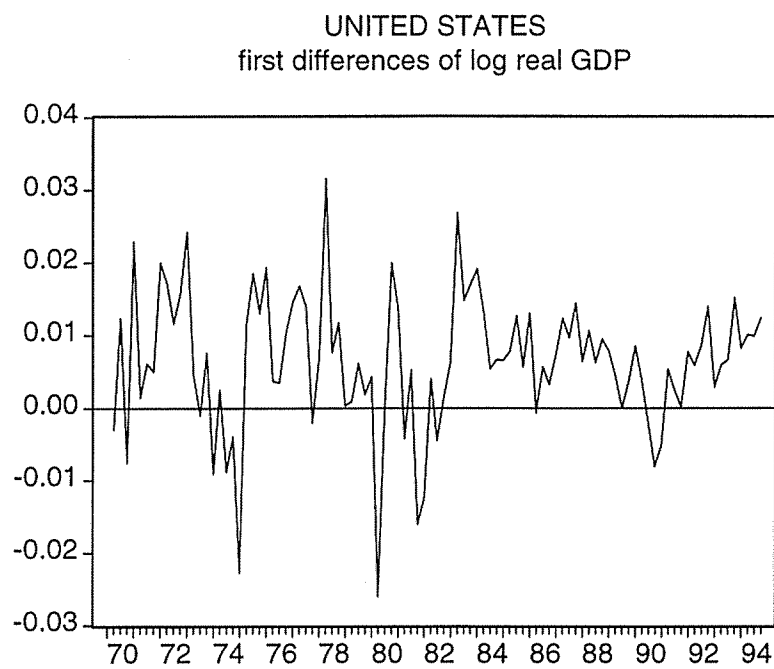
Figure 2:

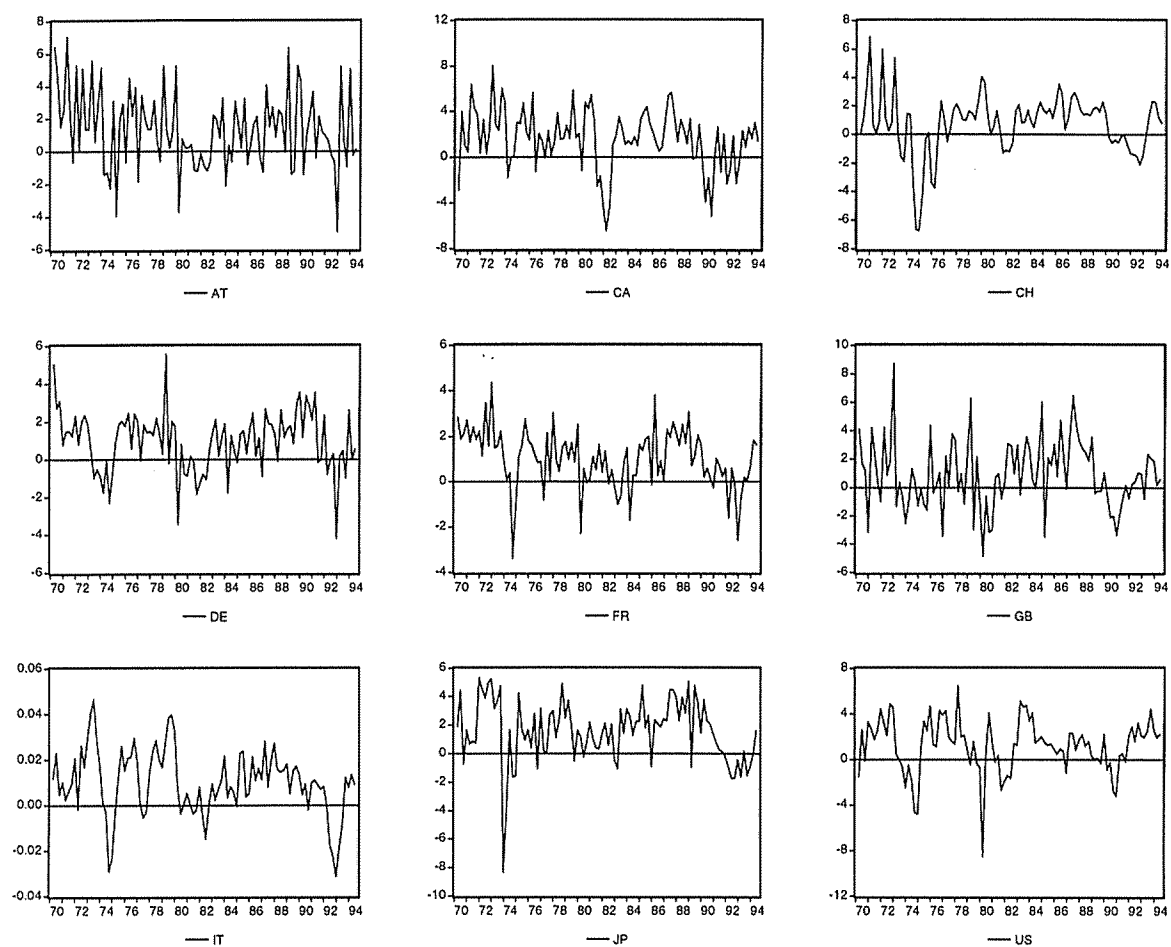100 x first differences of principal component

Figure 3:
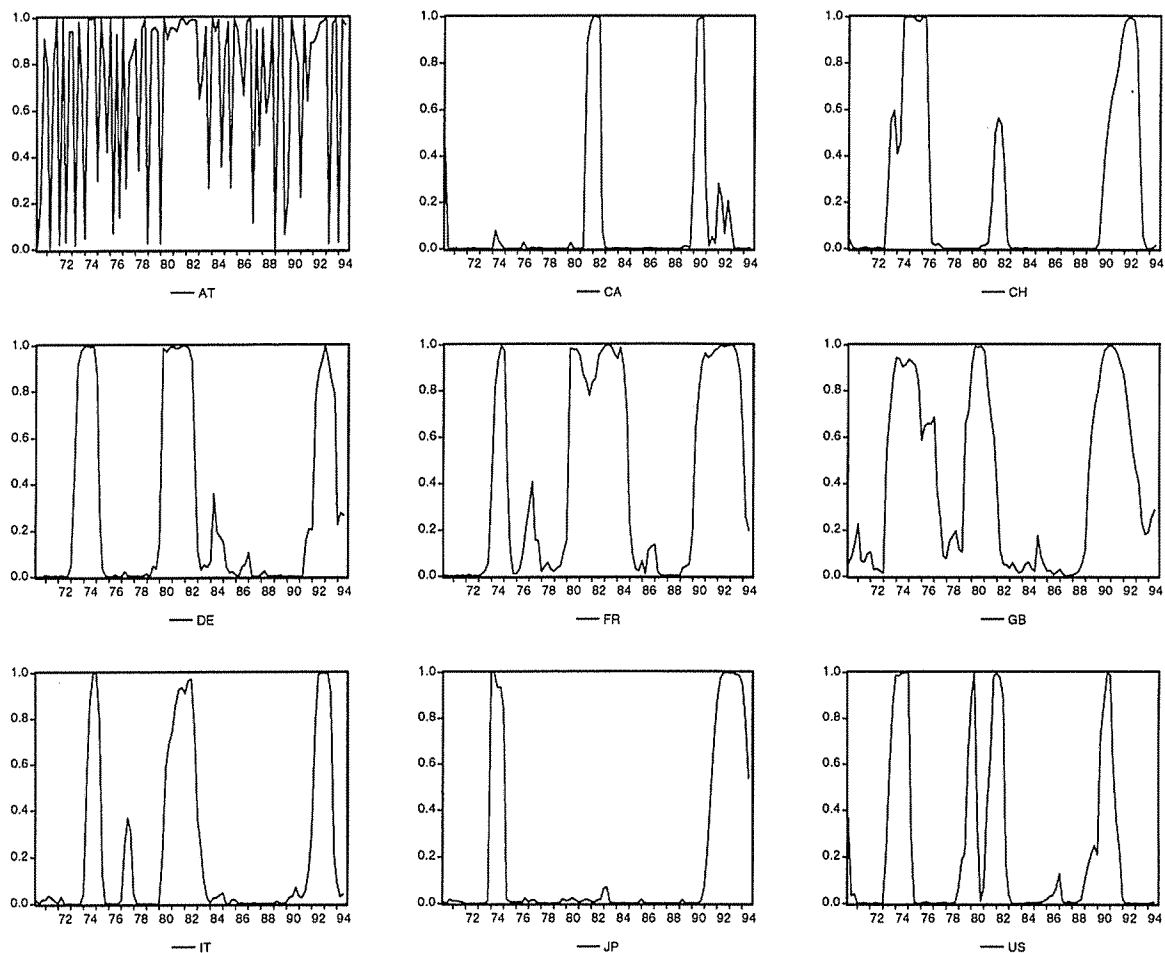
Principal components
smoothed prob(st=1), AR(1) term excluded

Figure 4:

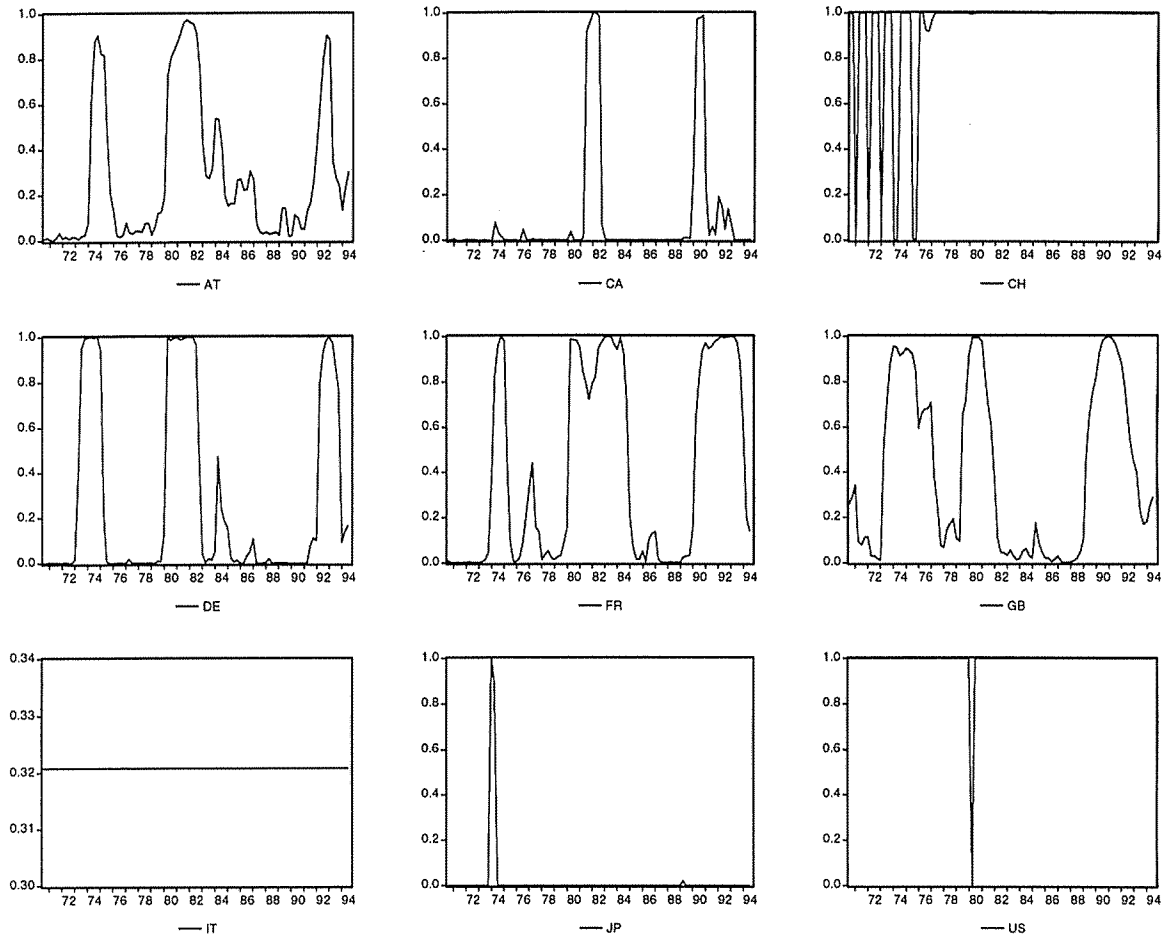Principal components
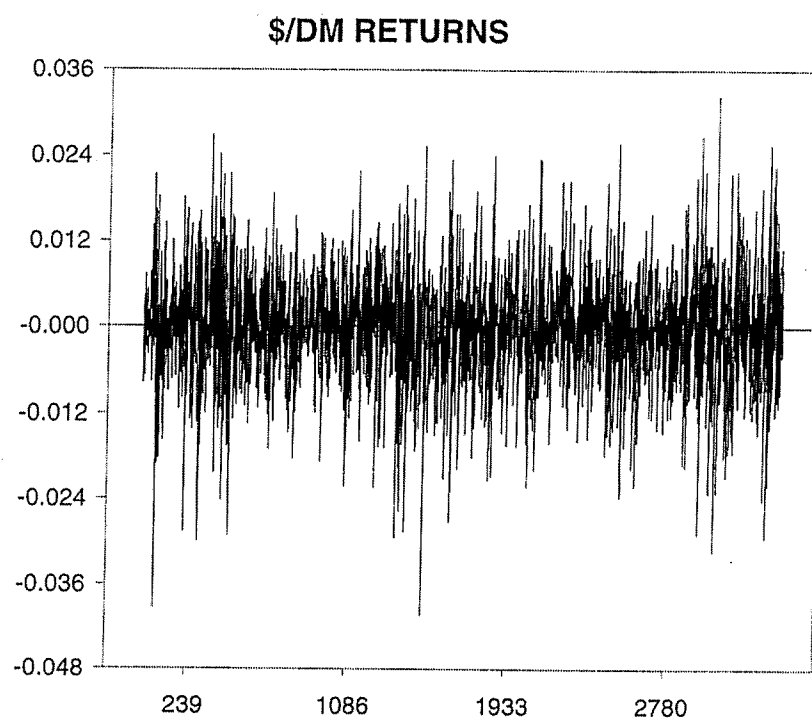smoothed prob(st=1), AR(1) term included
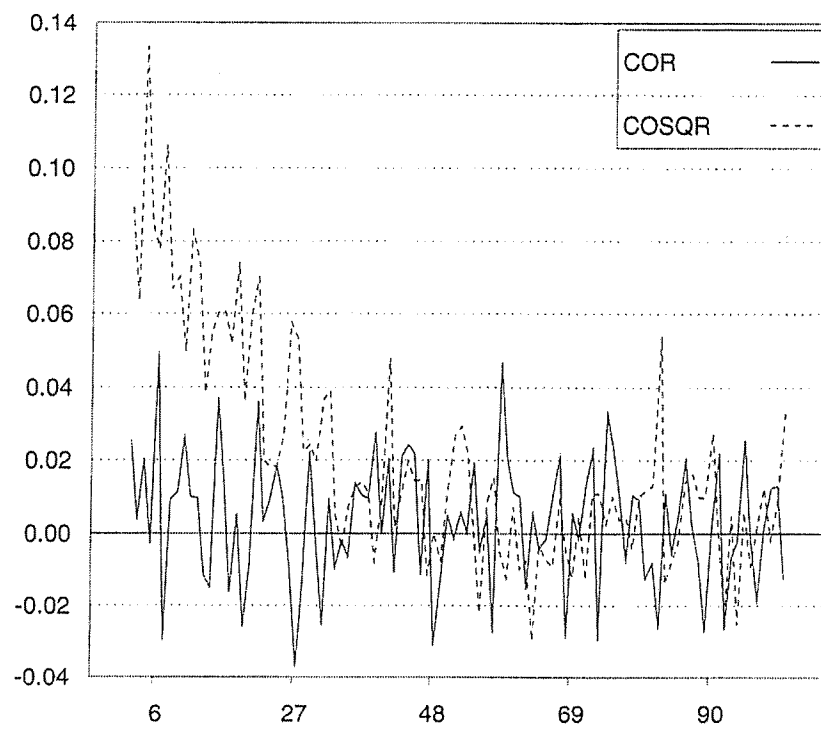
Figure 5:

$/DM RETURNS

Figure 6:

Figure 7:

Normalized Estimated Unconditional Distribution of DM
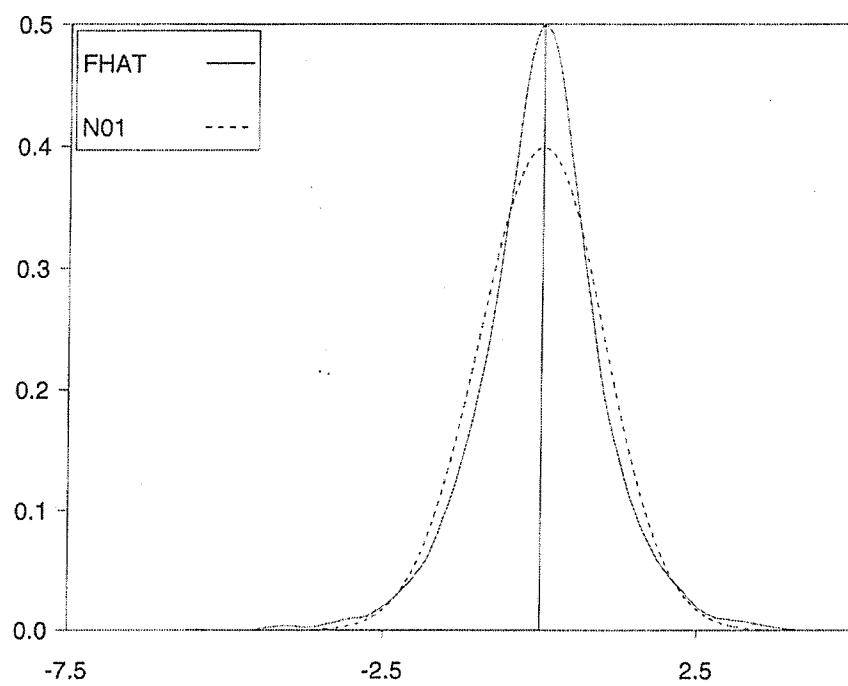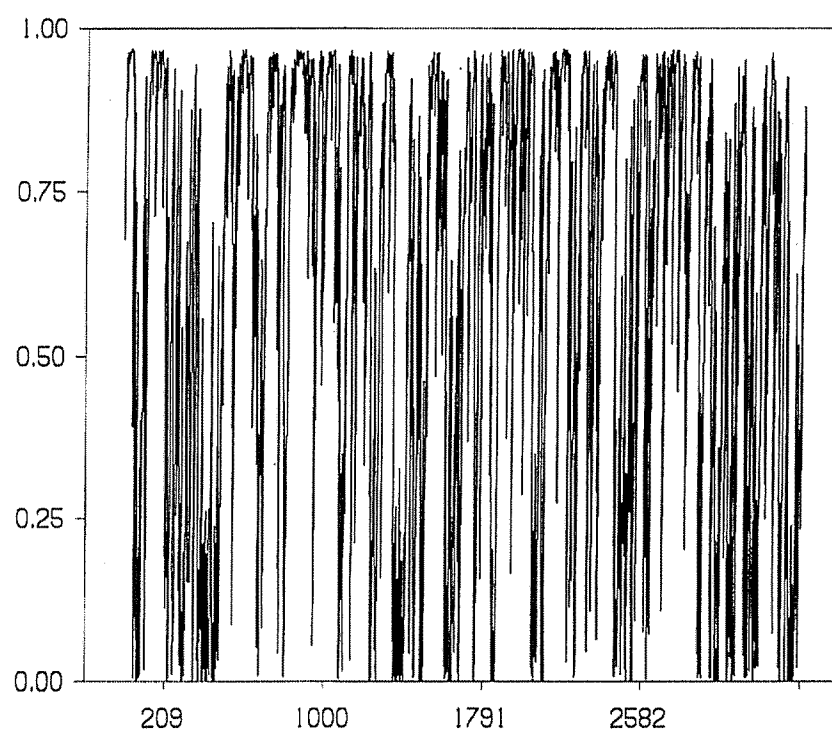
Figure 8:

C   EM algorithm

Looking at the first two iterations through 19 and 20 in a detailed way will help understand the implementation of Hamilton's EM algorithm. We take example I on page 6 and set the starting value for $\widehat{xi}_{1|0}$ to the vector of unconditional state probabilities π. Remember that y_t is a scalar, there are no exogenous variables and the number of regimes is 2. First note that

$$
\begin{aligned}
\widehat{\xi}_{1|0} &= \pi \\
&= P \cdot \widehat{\xi}_{0|0} \\
&= \left[\begin{array}{c} p_{11}\pi_1 + p_{21}\pi_2 \\ p_{12}\pi_2 + p_{22}\pi_2 \end{array} \right]
\end{aligned}
$$

The algorithm then yields:

$$
\widehat{\xi}_{1|1} = \frac{\left(\widehat{\xi}_{1|0} \odot \eta_1 \right)}{\mathbf{1}' \left(\widehat{\xi}_{1|0} \odot \eta_1 \right)}
$$

$$
\widehat{\xi}_{2|1} = P \cdot \widehat{\xi}_{1|1}
$$

$$
\widehat{\xi}_{1|0} \odot \eta_1 = \left[\begin{array}{c} \left(p_{11}\widehat{\xi}_{0|0,1} + p_{21}\widehat{\xi}_{0|0,2} \right) \cdot f \left(y_1 | s_1 = 1; \theta \right) \\ \left(p_{12}\widehat{\xi}_{0|0,1} + p_{22}\widehat{\xi}_{0|0,2} \right) \cdot f \left(y_1 | s_1 = 2; \theta \right) \end{array} \right]
$$

$$
\widehat{\xi}_{1|1} = \frac{\left(\widehat{\xi}_{1|0} \odot \eta_1 \right)}{\mathbf{1}' \left(\widehat{\xi}_{1|0} \odot \eta_1 \right)} = \left[\begin{array}{c} \frac{(p_{11}\widehat{\xi}_{0|0,1} + p_{21}\widehat{\xi}_{0|0,2}) \cdot f(y_1|s_1=1;\theta)}{\mathbf{1}'(\widehat{\xi}_{1|0} \odot \eta_1)} \\ \frac{(p_{12}\widehat{\xi}_{0|0,1} + p_{22}\widehat{\xi}_{0|0,2}) \cdot f(y_1|s_1=2;\theta)}{\mathbf{1}'(\widehat{\xi}_{1|0} \odot \eta_1)} \end{array} \right]
$$

$$
\widehat{\xi}_{2|1} = P \cdot \widehat{\xi}_{1|1} = \left[\begin{array}{c} p_{11}\widehat{\xi}_{1|1,1} + p_{21}\widehat{\xi}_{1|1,2} \\ p_{12}\widehat{\xi}_{1|1,1} + p_{22}\widehat{\xi}_{1|1,2} \end{array} \right]
$$

$$
\widehat{\xi}_{2|2} = \frac{\left(\widehat{\xi}_{2|1} \odot \eta_2 \right)}{\mathbf{1}' \left(\widehat{\xi}_{2|1} \odot \eta_2 \right)}
$$

$$
\widehat{\xi}_{2|1} \odot \eta_2 = \left[\begin{array}{c} \left(p_{11}\widehat{\xi}_{1|1,1} + p_{21}\widehat{\xi}_{1|1,2} \right) \cdot f \left(y_2 | s_1 = 1; \theta \right) \\ \left(p_{12}\widehat{\xi}_{1|1,1} + p_{22}\widehat{\xi}_{1|1,2} \right) \cdot f \left(y_2 | s_1 = 2; \theta \right) \end{array} \right]
$$

etc.

The programming steps can be summed up:

- Construct for each observation $f \left(y_t | s_t = j; \theta \right)$ $j = 1, 2$.

- Multiply $f \left(y_t | s_t = j; \theta \right)$ with p_{ij} $i, j = 1, 2$.

- Calculate recursively $\widehat{\xi}_{t|t}$. Note that $\widehat{\xi}_{t+1|t}$ has not to be computed explicitly to get the inference about the probabilities of state j prevailing at date t.

One additional step is needed to get an inference about the primitive states in example II on page 6. We had an autoregressive process with switching mean:

$$y_t - c_{s_t}^* = \phi(y_{t-1} - c_{s_{t-1}}^*) + \varepsilon_t \qquad \varepsilon_t \sim i.i.d. \ \mathcal{N}(0, \sigma^2).$$

Here

$$\widehat{\xi}_{t|t} = \begin{bmatrix} P(s_t = 1|\mathcal{Y}_t; \theta) \\ \vdots \\ P(s_t = 4|\mathcal{Y}_t; \theta) \end{bmatrix},$$

and

$$P(s_t^* = 1|\mathcal{Y}_t; \theta) = P(s_t = 1|\mathcal{Y}_t; \theta) + P(s_t = 3|\mathcal{Y}_t; \theta)$$

because

$$\begin{aligned} s_t = 1 \quad &\text{if} \quad s_t^* = 1, \quad s_{t-1}^* = 1 \\ s_t = 2 \quad &\text{if} \quad s_t^* = 2, \quad s_{t-1}^* = 1 \\ s_t = 3 \quad &\text{if} \quad s_t^* = 1, \quad s_{t-1}^* = 2 \\ s_t = 4 \quad &\text{if} \quad s_t^* = 2, \quad s_{t-1}^* = 2. \end{aligned}$$

D Maximum likelihood estimators

Taking logs of 32 and differentiating with respect to α' yields:

$$\log \eta_t = \begin{bmatrix} -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(y_t - z_t'\beta_1)^2 \\ \vdots \\ -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(y_t - z_t'\beta_N)^2 \end{bmatrix}$$

$$\frac{\partial \log \eta_t}{\partial \alpha'} = \begin{bmatrix} \frac{1}{\sigma^2}(y_t - z_t'\beta_1)z_t & 0_{1\times(N-2)} & 0 & -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y_t - z_t'\beta_1)^2 \\ 0_{(N-2)\times 1} & \ddots & 0_{(N-2)\times 1} & \vdots \\ 0 & 0_{1\times(N-2)} & \frac{1}{\sigma^2}(y_t - z_t'\beta_N)z_t & -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y_t - z_t'\beta_N)^2 \end{bmatrix}$$

Multiplying the transpose of this matrix with $\widehat{\xi}_{t|T}$ we get

$$\left(\frac{\partial \log \eta_t}{\partial \alpha'}\right)' \widehat{\xi}_{t|T} = \begin{bmatrix} \frac{1}{\sigma^2}(y_t - z_t'\beta_1)z_t \cdot P(s_t = 1|\mathcal{Y}_T; \theta) \\ \vdots \\ \frac{1}{\sigma^2}(y_t - z_t'\beta_N)z_t \cdot P(s_t = N|\mathcal{Y}_T; \theta) \\ \sum_{j=1}^N \left\{ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y_t - z_t'\beta_j)^2 \cdot P(s_t = j|\mathcal{Y}_T; \theta) \right\} \end{bmatrix}$$

Summing over T observations and setting to zero yields

$$\sum_{t=1}^T \left(\frac{\partial \log \eta_t}{\partial \alpha'}\right)' \widehat{\xi}_{t|T} = 0$$

$$\sum_{t=1}^T \left(y_t - z_t'\hat{\beta}_j\right) z_t \cdot P(s_t = j|\mathcal{Y}_T; \theta) = 0 \quad j = 1, \ldots, N \quad (89)$$

$$\sum_{t=1}^T \sum_{j=1}^N \left\{ \left[-\frac{1}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\left(y_t - z_t'\hat{\beta}_j\right)^2 \right] \cdot P(s_t = j|\mathcal{Y}_T; \theta) \right\} = 0 \qquad (90)$$

Finally, solving for $\hat{\beta}_j$

$$\sum_{t=1}^{T} z_t \left(y_t - z_t' \hat{\beta}_j \right) \cdot P\left(s_t = j | \mathcal{Y}_T; \theta \right) = 0$$

$$\left[\sum_{t=1}^{T} z_t y_t - \sum_{t=1}^{T} z_t z_t' \hat{\beta}_j \right] \cdot P\left(s_t = j | \mathcal{Y}_T; \theta \right) = 0$$

$$\frac{\sum_{t=1}^{T} z_t y_t \cdot P\left(s_t = j | \mathcal{Y}_T; \theta \right)}{\sum_{t=1}^{T} z_t z_t' \cdot P\left(s_t = j | \mathcal{Y}_T; \theta \right)} = \hat{\beta}_j , \qquad (91)$$

and for $\hat{\sigma}^2$

$$\sum_{t=1}^{T} \sum_{j=1}^{N} \left\{ \left[-\frac{1}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \left(y_t - z_t' \hat{\beta}_j \right)^2 \right] \cdot P\left(s_t = j | \mathcal{Y}_T; \theta \right) \right\} = 0$$

$$\sum_{t=1}^{T} -1 + \sum_{t=1}^{T} \sum_{j=1}^{N} \frac{1}{\hat{\sigma}^2} \left(y_t - z_t' \hat{\beta}_j \right)^2 \cdot P\left(s_t = j | \mathcal{Y}_T; \theta \right) = 0$$

$$\sum_{t=1}^{T} \sum_{j=1}^{N} \left(y_t - z_t' \hat{\beta}_j \right)^2 \cdot P\left(s_t = j | \mathcal{Y}_T; \theta \right) = T\hat{\sigma}^2$$

$$T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{N} \left(y_t - z_t' \hat{\beta}_j \right)^2 \cdot P\left(s_t = j | \mathcal{Y}_T; \theta \right) = \hat{\sigma}^2. \qquad (92)$$

Institut für Höhere Studien
Institute for Advanced Studies
Stumpergasse 56
A-1060 Vienna
Austria

Phone:+43-1-599 91-145
Fax: +43-1-599 91-163