2 A MARKOV SWITCHING COOKBOOK

Bruce Mizrach
Rutgers University
James Watkins
American Express

1. Introduction

Economists continue to debate the importance of nonlinearity to their discipline. When it comes to forecasting levels, unit roots seems to be quite prevalent, and there has been a great deal of skepticism about nonlinear models. See the arguments pro and con in Ramsey (1996). The time series properties of higher moments have, however, led researchers to go beyond the standard linear, normally distributed world of the textbooks. The two most widely developed lines of research in this area are the ARCH volatility models of Engle (1982), and the asymmetric Markov-switching model of Hamilton (1989). Our focus in this paper concerns numerical procedures for the estimation of the MS type of models.

Hamilton extended Goldfeld and Quandt's (1973) Markov switching regression to the time series context. He analyzed the growth rate of U.S. real GNP. Hamilton's model not only accommodated the asymmetries first noted by Neftci (1984), but also succeeded in reproducing the business cycle turning points established by the NBER. The Markov-switching framework for output was later generalized to allow for time-varying, duration-dependent, and seasonally dependent transition probabilities (as in Filardo (1994), Durland and McCurdy (1994) and Ghysels (1994) respec-

tively). Applications to interest rates by Hamilton (1988) and exchange rates by Engel and Hamilton (1990) illustrate the usefulness of the model outside of its initial application.

This paper is organized as follows. Section 2 briefly reviews Hamilton's Markov-switching framework and introduces a few general concepts. Estimation procedures are described in section 3, and the accompanying algorithms in sections 4 and 5. The relative merits of the two algorithms is discussed in section 6 and computational speed comparisons are presented in section 7. RATS code for both models is available from the authors at http://www-snde.rutgers.edu/ research.html

2. Time Series Models of Changes in Regime

A brief description of the Markov-switching (hereafter MS) framework is helpful to establish notation and vocabulary. The following description follows closely that of Hamilton (1993).

Consider for simplicity a first order autoregression where the mean value around which this series clusters may take on one of two values, $\mu^{(1)}$ and $\mu^{(2)}$:

$$y_t - \mu_t = \phi(y_{t-1} - \mu_{t-1}) + \varepsilon_t. \tag{1}$$

Suppose further that $\varepsilon_t \sim iidN(0, \sigma^2)$. A change in the value of μ alone is a change in regime (or state) in this simple case. It should be noted however, that all of the parameters of a model could be allowed to change with the state if thought appropriate. Hamilton's framework is rather agnostic regarding forces driving the change in regime: "...changes in regime are the result of processes largely unrelated to past realizations of the series and are not themselves directly observable." Hamilton (1993, p.234). It must be stressed that this does not mean that changes in regime are unrelated to the history of regimes; in fact, the state variable can have as long or longer a "memory" than the observation series y_t.

The state variable s_t is associated with the indices for the constant terms in equation 1; for instance, $s_t = 1$ is equivalent to saying $\mu_t = \mu^{(1)}$. Since the state variable is unobservable, we will need to form probabilistic inferences of its value, and in so doing form equivalent inferences regarding parameter values in 1. We assume that the state variable is governed by the Markov chain:

$$p(s_t = 1|s_{t-1} = 1) = p^{(11)}, \tag{2}$$

$$p(s_t = 2|s_{t-1} = 1) = p^{(12)}, \tag{3}$$

$$p(s_t = 1|s_{t-1} = 2) = p^{(21)}, \tag{4}$$

$$p(s_t = 2|s_{t-1} = 2) = p^{(22)}. \tag{5}$$

These transition probabilities are restricted so that $p^{(11)} + p^{(12)} = p^{(21)} + p^{(22)} = 1$. Hereafter, we will abbreviate notation when possible; the expression $p(s_t|s_{t-1})$ will refer to whichever of the above values is appropriate in the given context. The model given by 1, in conjunction with the assumptions regarding the transition probabilities,

will be referred to as an MS(1) model. By increasing the autoregressive dimension, we can consider higher-order MS(r) models.

As will become apparent through the exposition of the estimation techniques, numerous probabilistic inferences can be computed at different points throughout the sample. For example, the inference $p(s_t = 1, \ldots, s_{t-r} = 2|Y_t)$ will refer to the probability that the unobserved state variable took on the values $2, \ldots, 1$ at times $t - r$ through t respectively, conditioned on data up to and including that from date t. Such a value may be abbreviated $p(s_t, \ldots, s_{t-r}|Y_t)$ when possible, but note that the chronological subset represented by Y_t will always be included to distinguish between filter inferences, smoothed inferences, and r-lag smoothed inferences (all to be defined later). Likewise, an example of an observation density conditional on states would be $p(y_t|s_t = 1, \ldots, s_{t-r} = 2, Y_{t-1})$ and as may be apparent from the notation, may depend on a finite history of past regimes. It will also be necessary to compute joint densities of states and observations, given by expressions such as $p(y_t, s_t = 1, \ldots, s_{t-r} = 2|Y_{t-1})$. Both the observation densities and joint densities are abbreviated in the same manner as the inferences regarding states. Finally, more familiar looking observation densities (i.e. $p(y_t|Y_{t-1})$) will become available as functions of the above densities.

Often the conditional density depends not only on the current regime, but also on past regimes. This requires us to make assumptions regarding the memory of the state variable. For simplicity, the state process is generally given the same time dimension as the observation process. In other words, it will be assumed that the state variable is AR(r). To still satisfy the property of a Markov system, we must restructure the two regimes as 2^{r+1} distinct regimes, to exhaust the combinations of states. For example, in an MS(2) model, the expression

$$(y_t - \mu^{(1)}) = \phi^{(11)}(y_{t-1} - \mu^{(2)}) + \phi^{(12)}(y_{t-2} - \mu^{(1)}) + \varepsilon_t^{(1)} \tag{6}$$

is associated with the event: $(s_t = 1, s_{t-1} = 2, s_{t-2} = 1)$.

3. The Filter

Recall the assumption that the state variable s_t is generally unobservable. In order to estimate the parameters of a MS model with this uncertainty, we must compute probabilities associated with each possible regime. Further, in the case of a MS model where the conditional density depends on both current and past regimes

$$p(y_t|s_{t,}Y_{t-1}) \neq p(y_t|s_t, s_{t-1}, s_{t-2}, \ldots, Y_{t-1}), \tag{7}$$

these inferences need to extend over several periods

$$p(s_t, s_{t-1}, \ldots, s_{t-r}|Y_t). \tag{8}$$

Such probabilities are estimated using Hamilton's recursive filter; We discuss the filter in the general case of an MS(r) model. This procedure will compute r and $r + 1$ period inferences (2^r and 2^{r+1}distinct numbers for each t) and as a by-product, the conditional likelihood function. It is the conditional likelihood function that we seek for techniques such as Newton-Raphson (NR) or Davidon-Fletcher-Powell

(DFP), since the burdensome exact-likelihood method provides only a marginal improvement. The exposition is confined to the case where there are two states, and the initialization of the filter is reserved to the end of the discussion.

An arbitrary iteration of the filter begins by advancing an r period inference, available to us from the prior iteration,

$$p(s_{t+1}, s_t, \ldots, s_{t-r+1}|Y_t) = p(s_{t+1}|s_t) \cdot p(s_t, s_{t-1}, \ldots, s_{t-r+1}|Y_t). \tag{9}$$

We then use the appropriate density to find the joint probability inference of the current observation and the $r+1$ most recent states, conditional on last period's datum,

$$\begin{aligned} & p(y_{t+1}, s_{t+1}, s_t, \ldots, s_{t-r+1}|Y_t) \\ = \quad & p(y_{t+1}|s_{t+1}, s_t, \ldots, s_{t-r+1}, Y_t) \cdot p(s_{t+1}, s_t, \ldots, s_{t-r+1}|Y_t). \end{aligned} \tag{10}$$

Integrating over states, we find a density conditional only on prior data,

$$p(y_{t+1}|Y_t) = \sum_{s_{t+1}=1}^{2} \sum_{s_t=1}^{2} \sum_{s_{t-r+1}=1}^{2} (y_{t+1}, s_{t+1}, s_t, \ldots, s_{t-r+1}|Y_t). \tag{11}$$

We then have at our disposal an $r+1$ period inference conditional on current data

$$p(s_{t+1}, s_t, \ldots, s_{t-r+1}|Y_{t+1}) = \frac{p(y_{t+1}, s_{t+1}, s_t, \ldots, s_{t-r+1}|Y_t)}{p(y_{t+1}|Y_t)} \tag{12}$$

and by integration, an updated r period inference

$$\begin{aligned} p(s_{t+1}, s_t, \ldots, s_{t-r+2}|Y_{t+1}) \quad &= \\ p(s_{t+1}, s_t, \ldots, s_{t-r+1} \quad &= \quad 1|Y_{t+1}) + p(s_{t+1}, s_t, \ldots, s_{t-r+1} = 2|Y_{t+1}). \end{aligned} \tag{13}$$

The updated inference is then used as input for the next iteration. In later discussion, when we refer to "filter execution", this will mean that the entire sample is passed through the above process.

The filter is initialized with r-period unconditional probabilities,

$$p(s_r, s_{r-1}, \ldots, s_1) = p(s_t, s_{t-1}, \ldots, s_{t-r+1}). \tag{14}$$

To find these, we start by computing the ergodic probabilities, which are simply the unconditional estimates that the process will fall into each regime at an arbitrary date

$$\pi^{(j)} \equiv p(s_t = j) \text{ for } j = 1, 2. \tag{15}$$

These are found by solving the following set of equations:

$$p^{(1j)} \cdot \pi^{(1)} + p^{(2j)} \cdot \pi^{(2)} = \pi^{(j)} \text{ for } j = 1, 2 \tag{16}$$

$$\pi^{(1)} + \pi^{(2)} = 1. \tag{17}$$

Employing the appropriate transition probabilities, we can compute the necessary r-period unconditional probabilities, e.g.:

$$\begin{aligned} p(s_t \quad &= \quad 1, s_{t-1} = 2, s_{t-2} = 1) \\ &= \quad (1 - p(s_t = 2|s_{t-1} = 2)) \cdot (1 - p(s_{t-1} = 1|s_{t-2} = 1)) \cdot \pi^{(1)}. \end{aligned} \tag{18}$$

In the case of an MS(r) system, one needs to compute 2^r of these probabilities to initialize the filter.

After the entire sample has been passed through the filter, the computed observation densities can be used to form the conditional likelihood function

$$p(y_T, y_{T-1}, \ldots, y_{r+1}) = \prod_{t=r+1}^{T} p(y_t|Y_{t-1}). \tag{19}$$

3.1 The Full-Sample Smoother

The filter inferences provided above form a time series that can be volatile; they may in fact indicate changes in regime that have not occurred.. Smoothed inferences are more extensively conditioned by utilizing both past and future observations. This feature reduces the chance that we will misinterpret an outlier occurring in a particular regime for an actual change of state. Smoothed inferences can be distinguished from filter inferences by the time subscript on the relevant information set, Y_T instead of Y_t.

Suppose now that we wish to compute $r+1$ period smoothed inferences for our MS(r) model. After executing the filter to obtain $r+1$ period filter inferences, we expand those inferences to length $r+2$

$$\begin{aligned} & p(s_{t+1}, s_t, \ldots, s_{t-r}|Y_{t+1}) \qquad (20) \\ = \quad & \frac{p(y_{t+1}|s_{t+1}, s_t, \ldots, s_{t-r+1}, Y_t) \cdot p(s_{t+1}|s_t) \cdot p(s_t, s_{t-1}, \ldots, s_{t-r}|Y_t)}{p(y_{t+1}|Y_t)} \end{aligned}$$

The $r+2$ period inferences can then be expanded in the same manner to $r+3$ period inferences, the $r+3$ period inferences into $r+4$ period inferences, and so on. This process is continued until we reach inferences of length $2r+2$ (2^{2r+2} distinct numbers for each t). At this point, we can integrate to find what might be referred to as non-adjacent probability inferences

$$\begin{aligned} & p(s_{t+r+1}, s_{t+r}, \ldots, s_{t+2}, s_t, \ldots, s_{t-r}|Y_{t+r+1}) \qquad (21) \\ = \quad & \sum_{s_{t+1}=1}^{2} (s_{t+r+1}, s_{t+r}, \ldots, s_{t-r}|Y_{t+r+1}). \end{aligned}$$

These quantities can be advanced one period

$$\begin{aligned} & p(s_{t+r+2}, s_{t+r+1}, \ldots, s_{t+2}, s_t, \ldots, s_{t-r}|Y_{t+r+2}) \qquad (22) \\ = \quad & \frac{p(y_{t+r+2}|s_{t+r+2}, s_{t+r+1}, \ldots, s_{t+2}, Y_{t+r+1})}{p(y_{t+r+2}|Y_{t+r+1})} \times \\ & p(s_{t+r+2}|s_{t+r+1}) \cdot p(s_{t+r+1}, s_{t+r}, \ldots, s_{t+2}, s_t, \ldots, s_{t-r}|Y_{t+r+1}), \end{aligned}$$

then integrated again to expand the gap in the inferences

$$\begin{aligned} & p(s_{t+r+2}, s_{t+r}, \ldots, s_{t+3}, s_t, \ldots, s_{t-r}|Y_{t+r+2}) \qquad (23) \\ = \quad & \sum_{s_{t+2}=1}^{2} p(s_{t+r+2}, s_{t+r}, \ldots, s_{t+2}, s_t, \ldots, s_{t-r}|Y_{t+r+2}). \end{aligned}$$

Such iterations continue until the end of the sample is reached

$$p(s_T, s_{T-1}, \ldots, s_{T-r}, s_t, \ldots, s_{t-r}|Y_T),$$

allowing us, via integration, to find full-sample the smoothed inferences

$$p(s_t, s_{t-1}, \ldots, s_{t-r}|Y_T) \tag{24}$$

$$= \sum_{s_T=1}^{2}\sum_{s_{T-1}=1}^{2}\cdots\sum_{s_{T-r}=1}^{2} p(s_T, s_{T-1}, \ldots, s_{T-r}, s_t, s_{t-1}, \ldots, s_{t-r}|Y_T).$$

3.2 Approximations to Full-Sample Smoothed Inferences

Approximations to the above inference are available with significantly less computation. If we use partial conditioning on future observations, we may still arrive at a stable sequence of inferences. This possibility was recognized by Hamilton (1989). After executing the filter, one expands the inferences as would be done in the first step of the full-sample smoother. When inferences of length $2r+1$ are reached, one integrates over $s_{t+1}, s_{t+2,} \ldots, s_{t+r}$ to find the r-lag smoothed inference

$$\begin{aligned} & p(s_t, s_{t-1}, \ldots, s_{t-r}|Y_{t+r}) \qquad (25)\\ = & \sum_{s_{t+r}=1}^{2}\cdots\sum_{s_{t+1}=1}^{2} p(s_{t+r}, s_{t+r-1}, \ldots, s_{t+1}, s_t, s_{t-1}, \ldots, s_{t-r}|Y_{t+r}). \end{aligned}$$

These inferences get their name from the quantity of future data used for conditioning. The justification for this method is that we have considered data sufficiently far in the future that any further conditioning of the inference for date t does not utilize the observation from date t. Any further conditioning should have a negligible impact.

4. Hill Climbing

It is almost always the case that we are unsure as to when each regime was active in our sample. We therefore need to handle two types of uncertainty during estimation: uncertainty regarding parameter values and uncertainty regarding the path of the state variable. Recall that the filter computes, as a by-product, the conditional likelihood function

$$L(\theta) = \prod_{t=r+1}^{T} p(y_t|Y_{t-1}), \qquad (26)$$

where

$$\begin{aligned} p(y_t|Y_{t-1}) = & \sum_{s_t=1}^{2}\cdots\sum_{s_{t-r}=1}^{2} p(y_t|s_t, s_{t-1}, \ldots, s_{t-r}, Y_{t-1}) \qquad (27)\\ & \times p(s_t|s_{t-1}) \cdot p(s_{t-1}, s_{t-2}, \ldots, s_{t-r}|Y_{t-1}). \end{aligned}$$

The dual uncertainty of our estimation problem makes maximization of the above likelihood function is more complicated than it may first appear. If we knew the path of the state variable, we could simply maximize the above function with only small modifications to a canned software package. However, each chronological element of the likelihood function is a mixture distribution, composed of the densities representing each state. The weights for these densities (the probabilistic inferences associated with each regime or recent history of regimes) are functions of the parameters being estimated. As a result, the complete set of filter inferences will change with every perturbation of the parameter vector during gradient computation. For the sake of clarity, consider an arbitrary iteration of a DFP routine. We have at our disposal a tentative value for the set of parameters. Each element of the gradient requires that we execute the filter twice. Next, every stepsize to be considered requires a new execution of the filter, and only then can the updated parameter vector be chosen. We

leave it to the reader to consider what is involved to compute the Hessian numerically.

Apart from this additional consideration, numerical maximum likelihood techniques, such as DFP are applied in the normal way.

5. The EM Algorithm

The EM algorithm (hereafter EMA), as outlined by Hamilton(1990), deals with the dual uncertainty problem in a different way. One begins with an initial guess for the vector of parameters, say $\boldsymbol{\theta}^{[0]}$. The filter and smoother, both parametrized by the extent of regime dependence, are executed to obtain inferences conditional on the entire sample of observations. The smoothed inferences are used as weights for coefficient updating, via minimization of the sum of weighted squared residuals. Improved estimates of the transition probabilities are simple functions of the smoothed probabilities. The set of updated values constitutes $\boldsymbol{\theta}^{[1]}$; we repeat the process until some convergence criterion is satisfied.

To outline the procedure in greater detail, consider a two-state MS(2) model:

$$(y_t - \mu^{(1)}) = \phi^{(11)}(y_{t-1} - \mu_{t-1}) + \phi^{(12)}(y_{t-2} - \mu_{t-2}) + \varepsilon_t^{(1)}, \tag{28}$$

$$(y_t - \mu^{(2)}) = \phi^{(21)}(y_{t-1} - \mu_{t-1}) + \phi^{(22)}(y_{t-2} - \mu_{t-2}) + \varepsilon_t^{(2)}. \tag{29}$$

$\varepsilon_{i,t} \sim N(0, (\sigma^{(i)})^2)$, $i = 1, 2$. Constructed this way, 10 parameters need to be estimated: $\mu^{(1)}$, $\mu^{(2)}$, $\phi^{(11)}$, $\phi^{(12)}$, $\phi^{(21)}$, $\phi^{(22)}$, $\sigma^{(1)}$, $\sigma^{(2)}$, $p^{(11)}$, $p^{(22)}$. A single iteration of the EM algorithm starts by executing the filter. Upon completion, the filter yields regime inferences for 2 periods

$$p(s_t, s_{t-1}|Y_t)$$

and for 3 periods

$$p(s_t^*|Y_t) = p(s_t, s_{t-1}, s_{t-2}|Y_t). \tag{30}$$

We use the $*$ to redefine the state in terms of the 8 permutations of the 3 lags, e.g. $s_t^* = 1$ implies $s_t = 1$, $s_{t-1} = 1$, $s_{t-2} = 1$. Also obtained are observational densities, both conditioned on states

$$p(y_t|s_t^*, Y_{t-1}) = p(y_t|s_t, s_{t-1}, s_{t-2}, Y_{t-1}) \tag{31}$$

and unconditional with regard to states

$$p(y_t|Y_{t-1}).$$

A typical observation density is written

$$\begin{aligned} p(y_t|s_t^* &= 3, Y_{t-1}) \\ &= \frac{1}{\sigma^{(1)}\sqrt{2\pi}} \exp\{[(y_t - \mu^{(1)}) - \phi^{(11)}(y_{t-1} - \mu^{(2)}) \\ &\quad -\phi^{(12)}(y_{t-2} - \mu^{(1)})]^2/2(\sigma^{(1)})^2\}. \end{aligned}$$

We next execute the smoother to find a probability series that is less volatile than

that provided by the filter

$$p(s_t^*|Y_T) = p(s_t, s_{t-1}, s_{t-2}|Y_T), \tag{32}$$

while integration yields

$$p(s_t, s_{t-1}|Y_T) \text{ and } p(s_t|Y_T). \tag{33}$$

With smoothed inferences available, coefficients are updated numerically by minimizing the sum of weighted squared residuals

$$\left[\mu^{(1)}, \mu^{(2)}, \boldsymbol{\phi}\right] = \arg\min \left\{ \sum_{t=3}^{T} \sum_{k=1}^{8} (\nu_t^{(k)})^2 \cdot p(s_t^* = k|Y_T) \right\} \tag{34}$$

where $\boldsymbol{\phi} = [\phi^{(11)}, \phi^{(12)}, \phi^{(21)}, \phi^{(22)}]$ and

$$\nu_t^{(k)} = y_t - E\left(y_t|s_t^* = k, Y_{t-1}\right). \tag{35}$$

An example of an error series element would be

$$\nu_t^{(2)} = (y_t - \mu^{(2)}) - \phi^{(21)}(y_{t-1} - \mu^{(1)}) - \phi^{(22)}(y_{t-2} - \mu^{(2)}).$$

Estimation of the regimes' variances requires similar weighting

$$\begin{aligned}(\sigma^{(j)})^2 &= \frac{p(s_t = j, s_{t-1}, s_{t-2}|Y_T)}{\sum_{t=3}^{T} p(s_t = j)} \\ &\times \sum_{t=3}^{T} \sum_{s_{t-1}=1}^{2} \sum_{s_{t-2}=1}^{2} [(y_t - \mu^{(j)}) - \phi^{(j1)}(y_{t-1} - \mu^{(s_{t-1})}) \\ &-\phi^{(j2)}(y_{t-2} - \mu^{(s_{t-2})})]^2\end{aligned}$$

for $j = 1, 2$. Updated transition probabilities also utilize smoothed inferences

$$p_{ij} = p(s_t = j|s_{t-1} = i) = \frac{\sum_{t=r+1}^{T} p(s_t = j, s_{t-1} = i|Y_T)}{\sum_{t=r+1}^{T} p(s_{t-1} = i|Y_T)} \text{ for } i, j = 1, 2. \tag{36}$$

We have completed a single iteration of the algorithm; we repeat until convergence.

6. Relative merits of Different Algorithms

The DFP routine described earlier is the same method used by Hamilton (1989) for his analysis of the business cycle. Hamilton (1990) proposed the above EMA as an alternative, to handle systems of greater complexity.

Problems may arise during gradient computation due to the shape of the likelihood surface associated with a MS model. Mixture distributions may have as many local maxima as regimes, and likelihood functions derived from these densities may be plagued by the same features. The EMA however, does not involve the examination of likelihood surfaces, and as such, may avoid both local maxima and singularities. Another positive attribute of the EMA noted in other applications is its ability to arrive in the neighborhood of the mode of the likelihood function in a few early steps, which can prove advantageous if one is performing a rough grid search to determine optimal starting values. Hamilton also argues that an EMA may not be as demanding

numerically: "while one could calculate analytic derivatives from rote adaptation of the recursion ..., that approach would require burdensome additional computer programming and calculation time for each parameter." Hamilton (1990, p. 40).

If the EMA is indeed more robust than a DFP algorithm and computationally more efficient, one would clearly prefer the former. However, Hamilton's claim regarding the computational speed of the EMA is problematic. Consider the full-sample smoothing technique outlined earlier: this procedure requires computation of order $K^{2(r+1)}T^2$, where r is the relevant autoregressive dimension, K is the number of states and T is the sample size. In our experience, if r is 4 or larger, while $K = 2$ and T is a modest 100 to 200 points, a single iteration of the EMA can take several minutes on an 166 MHz Pentium, utilizing the 32 bit version of RATS. If several hundred iterations are necessary to achieve convergence, such investment in computer time becomes prohibitive. By contrast, a single iteration of a DFP routine requires several seconds.

Despite these concerns, the EMA may be preferable in many contexts. If no autoregressive dynamics are present, the full-sample smoother becomes much less of a problem, and the EMA's ability to avoid the difficulties of poorly shaped likelihood surfaces outweighs any additional CPU time. In their investigation of exchange rates, Engel and Hamilton (1990) successfully employed the EMA in the presence of numerous local maxima. The case for the EMA is also strengthened by the possibility of using the approximate smoother, as it is only marginally more demanding than the filter.

7. Computational Comparisons

To explore the computational issues more thoroughly, consider the Hamilton's (1989) analysis of real GNP growth. The model was fitted by DFP:

$$y_t - \mu_t = \textstyle\sum_{p=1}^{4} \phi^{(ip)}(y_{t-p} - \mu_{t-p}) + \varepsilon_t,\ \varepsilon_t \sim iidN(0, \sigma^2). \tag{37}$$

The two constants $\mu^{(1)}$ and $\mu^{(2)}$ are associated with high and low rates of growth. We recreated Hamilton's results via three methods: DFP, the EMA with full-sample smoothing (EMA(1)), and the EMA using approximate smoothing (EMA(2)). Presented below are the computational demands of each algorithm for different convergence criteria.

Table 1 illustrates two important claims: first, that using the approximate EMA significantly reduces CPU time relative to the pure form of the EMA; second, when comparing gradient-search methods with EM methods, a tradeoff exists between the time that it takes to complete an iteration and the number of iterations needed to achieve convergence. While EM uses $2/3$ fewer iterations, each iteration takes from 5 to 20 times as long.

The usefulness of the approximate EM routine is of course dependent on the accuracy of the results. Comparing the approximate EM results with those provided by the DFP algorithm, we find that the mean absolute difference between the elements of the parameter vector decreases as the convergence criterion is tightened, and is no greater 0.0085 if we set $\delta = 10^{-5}$.

Table 1. Number of Iterations to Convergence

Criterion	**DFP**	**EMA(1)**	**EMA(2)**
$\delta = 10^{-3}$	61	22	27
$\delta = 10^{-4}$	166	50	51
$\delta = 10^{-5}$	254	69	75
sec./iter.	12	227	57

8. Conclusion

Markov-switching models have become widely used in accordance with the growing evidence of nonlinearity in economic time series. We have detailed the two leading algorithms for estimating MS models and have provided a discussion of the computational questions that arise during the course of an estimation problem.

Although the EM algorithm is usually more robust with respect to poor likelihood surfaces than gradient-based methods, the pure form of the former technique often places prohibitive demands on computer time. Fortunately, the availability of an approximation technique allows the EM method to stand as a practical alternative to commonly used hill-climbing routines.

References

Durland, J. Michael and Thomas H. McCurdy (1994). "Duration Dependent Transitions in a Markov Model of U.S. GNP Growth," *Journal of Business and Economic Statistics*, 12, 279-88.

Engel, Charles (1994). "Can the Markov Switching Model Forecast Exchange Rates?" *Journal of International Economics*, 36, 151-65.

Engel, Charles and James D. Hamilton (1990). "Long Swings in the Dollar: Are They in the Data and Do Markets Know It?" *American Economic Review*, 80, 689-713.

Engle, Robert F. (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation," *Econometrica*, 50, 987-1007.

Filardo, Andrew J. (1994). "Business Cycle Phases and their Transitional Dynamics," *Journal of Business and Economic Statistics*, 12, 299-308.

Ghysels, Eric (1994), "On the Periodic Structure of the Business Cycle," *Journal of Business and Economic Statistics*, 12, 289-98.

Goldfeld, Stephen M. and Richard E. Quandt (1973). "A Markov Model for Switching Regressions," *Journal of Econometrics*, 1, 3-16.

Goodwin, Thomas H. (1993). "Business Cycle Analysis with a Markov Switching Model," *Journal of Business and Economic Statistics*, 11, 331-39.

Hamilton, James D. (1989). "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357-84.

Hamilton, James D. (1990). "Analysis of Time Series Subject to Regime Changes,"

Journal of Econometrics, 45, 39-70.

Hamilton, James D. (1993). "Estimation, Inference and Forecasting of Time Series Subject to Changes in Regime," in G.S. Maddala, C.R. Rao and H.D. Vinod, (eds.), *Handbook of Statistics* Vol. 11, Amsterdam: North Holland.

Neftci, Salih N. (1984). "Are Economic Time Series Asymmetric over the Business Cycle?" *Journal of Political Economy,* 92, 307-28.

Ramsey, James (1996). "If Nonlinear Models Can't Forecast, What Use are They?" *Studies in Nonlinear Dynamics and Econometrics,* 1, 65-86.