Journal of Econometrics 45 (1990) 39–70. North-Holland

ANALYSIS OF TIME SERIES SUBJECT TO CHANGES IN REGIME*

James D. HAMILTON

University of Virginia, Charlottesville, VA 22901, USA

This paper introduces an EM algorithm for obtaining maximum likelihood estimates of parameters for processes subject to discrete shifts in autoregressive parameters, with the shifts themselves modeled as the outcome of a discrete-valued Markov process. The simplicity of the EM algorithm permits potential application of the approach to large vector systems.

1. Introduction

The premise of this paper is that many of the important movements in asset prices arise from specific identifiable events. On a week-to-week basis, discrete shifts in the Federal Reserve's target band for the Federal funds rate are often associated with dramatic moves in Treasury bill rates [Cook and Hahn (1989)]. Over a longer time horizon, more fundamental changes in monetary, fiscal, or incomes policies, such as the change in Fed operating procedure in October 1979, also have come in the form of a discontinuous change in the financial environment. One might further argue that many of the major exogenous economic events influencing financial time series, such as World War II or the OPEC oil shock, should also be viewed as episodes of identifiable duration in which the dynamic behavior of key economic series might be expected to differ significantly from that seen outside these episodes.

This paper builds on an approach introduced in my 1989 paper for analyzing such discrete qualities of time series. I regard the parameters of a vector autoregression as subject to occasional discrete shifts. The probability law governing these shifts is also stated explicitly and presumed to exhibit dynamic behavior of its own. The econometrician's task is then to determine when the shifts occurred and to estimate parameters characterizing the different regimes and the probability law for the transition between regimes. The expectations for the future that one would make in a world character-

*This material is based upon work supported by the National Science Foundation under Grant No. SES-8720731. I am grateful for insights offered by Robert Engle, Angelo Melino, Kerk Phillips, Paul Ruud, James Stock, and Mark Watson.

ized by such changes can also be calculated in a straightforward manner, permitting tests of sophisticated rational expectations hypotheses.

Modeling this discreteness can have important payoffs. For example, Fama and French (1988) and Poterba and Summers (1988) reported evidence of mean reversion in stock prices, implying predictable profit opportunities for a risk-neutral investor. However, Cecchetti, Lam, and Mark (forthcoming) noted that these results cannot be considered statistically significant if dividend growth rates are subject to the dramatic, abrupt shifts that appear to be in the data. And Perron (1989) and Lam (1988) have shown that allowing the possibility of occasional discrete shifts in trend can make a major difference for hypotheses about the persistence of innovations in key economic and financial time series.

Despite the promise of this approach, applications of the method proposed in my 1989 paper were basically limited to relatively small systems, owing to the computational difficulty of maximizing numerically an often ill-behaved likelihood surface with respect to a large number of unknown parameters. This paper introduces some new technical tools that may be of help in handling these problems.

The paper provides an analytic characterization of the derivative of the sample log-likelihood function for this class of models. While one could calculate analytic derivatives from rote adaptation of the recursion used to evaluate the likelihood function in my 1989 paper, that approach would require burdensome additional computer programming and calculation time for each parameter. By contrast, the expressions in this paper permit analytic derivatives to be calculated quite trivially from the smoothed inferences about the unobserved regime, a series the econometrician may well have already calculated. Adding more parameters requires no changes in the routine for calculating smoothed probabilities, and thus has essentially no effect on the computation time required to calculate the gradient. Indeed, the methods described below can maximize the likelihood function for a large vector system in less time than required for scalar systems, because the time required per iteration is basically independent of the size of the system and the number of iterations required can be lower.

The paper further shows how this class of models can be estimated by using the EM principle of Dempster, Laird, and Rubin (1977). This offers a means of maximizing the sample likelihood that is an alternative to such methods as Newton–Raphson or Davidon–Fletcher–Powell. The principal advantage of the EM algorithm over these methods is its numerical robustness. The likelihood functions for switching regression models can be plagued with multiple local maxima, essential singularities, and local increases in the likelihood function as boundary conditions are approached. In regions where the likelihood surface is not concave, methods that seek to approximate the

sample Hessian can easily go astray and crash the system with numerical underflows or overflows. By contrast, the EM algorithm by construction finds an analytic interior solution to a particular subproblem. My experience has been that the EM algorithm is quite robust with respect to poorly chosen starting values, and quickly moves to a reasonable region of the likelihood surface.

It is well-known that the EM algorithm has slower convergence than most popular hill-climbing methods once it is in the vicinity of the maximum. I have not found this to be a significant problem in my applications. The large, well-chosen initial steps that the algorithm selects and its ease of use more than make up for the time lost on the last few iterations. If (as seems desirable) one explores a large number of possible starting values for maximum likelihood estimation, the EM algorithm offers a vast improvement in efficiency, since its numerical robustness permits execution of hundreds of maximizations with no adjustments by the user.

The use of i.i.d. switching regressions was introduced in econometrics by Quandt (1958). Switching regressions in which the regime follows an unobserved Markov process first appeared in Goldfeld and Quandt (1973). Cosslett and Lee (1985) studied a regression where an unobserved dichotomous explanatory variable was presumed to follow a Markov process, and the principles they use to evaluate the likelihood function are those adapted in my 1989 paper to study a time series autoregression subject to time-varying coefficients.

The EM algorithm offers an alternative method for maximizing the likelihood function in such models. Previous applications of the EM algorithm to econometric problems have been surveyed by Kiefer (1980), Watson and Engle (1983), and Ruud (1988). Kiefer (1980) considered the case of i.i.d. switching regressions, and the algorithm presented below might be viewed as a natural generalization of his approach. Alternatively, Baum et al. (1970) introduced an algorithm for estimation of a scalar system with no explanatory variables or autoregressive dynamics [$m = 0$ in eq. (2.1) below], but with an unobserved Markov switching process for the mean and variance; Liporace (1982) discussed the vector case, again with no explanatory variables and for $m = 0$. EM estimation for the general systems considered below has not previously been discussed.

The plan of the paper is as follows. Section 2 provides a statement of the basic framework. Section 3 summarizes the general principles behind the EM algorithm, while section 4 summarizes the particular algebraic results necessary to apply the EM principle in the present context. Examples that clarify the logic and simplicity of this algorithm are provided in section 5. Section 6 comments on use of Bayesian priors, alternative approaches to modeling nonstationary processes, and hypothesis testing.

2. A stochastic model of changes in regime

Suppose we have a sample of size T $(y_1, \ldots, y_T)$ from a vector-valued process $(y_t \in \mathbb{R}^n)$. The econometrician believes that there may be occasional discrete shifts in the level, variance, or autoregressive dynamics of y. Suppose that there are a total of K possible regimes from which a particular observation y_t might have been drawn. To model this concept, I introduce an unobserved scalar random variable which I refer to as the 'state' of the process. This unobserved state, denoted s_t, takes on an integer value in $\{1, \ldots, K\}$. I assume that there is a maximal autoregressive lag order m such that y_t depends only on the current and m most recent values of s_t, on m lags of y_t, and on a vector of parameters $\boldsymbol{\theta}$:

$$\begin{aligned} & p(y_t|s_t, s_{t-1}, \ldots, y_{t-1}, y_{t-2}, \ldots; \boldsymbol{\theta}) \\ & = p(y_t|s_t, s_{t-1}, \ldots s_{t-m}, y_{t-1}, y_{t-2}, \ldots, y_{t-m}; \boldsymbol{\theta}) \\ & \equiv p(y_t|z_t; \boldsymbol{\theta}), \end{aligned} \tag{2.1}$$

where

$$z_t \equiv (s_t, s_{t-1}, \ldots s_{t-m}, y_{t-1}', y_{t-2}', \ldots, y_{t-m}')'.$$

For example, my (1989) paper considered a scalar-valued fourth-order autoregression around one of two constants, μ_1 or μ_2:

$$\begin{aligned} (y_t - \mu_{s_t}) = {} & \phi_1(y_{t-1} - \mu_{s_{t-1}}) + \phi_2(y_{t-2} - \mu_{s_{t-2}}) + \phi_3(y_{t-3} - \mu_{s_{t-3}}) \\ & + \phi_4(y_{t-4} - \mu_{s_{t-4}}) + \varepsilon_t, \end{aligned}$$

with $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$. The proposal is thus that there might be an occasional shift in the constant term around which the autoregression clusters (μ_1 or μ_2). In this case, $n = 1$, $m = 4$, $K = 2$, $s_t = 1$ or 2 depending on the state at data t, $\boldsymbol{\theta} = (\mu_1, \mu_2, \phi_1, \phi_2, \phi_3, \phi_4, \sigma)$, and

$$\begin{aligned} p(y_t|z_t; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big[& -\big((y_t - \mu_{s_t}) - \phi_1(y_{t-1} - \mu_{s_{t-1}}) \\ & - \phi_2(y_{t-2} - \mu_{s_{t-2}}) - \phi_3(y_{t-3} - \mu_{s_{t-3}}) \\ & - \phi_4(y_{t-4} - \mu_{s_{t-4}})\big)^2 / [2\sigma^2]\Big]. \end{aligned} \tag{2.2}$$

As a second example of (2.1), Engel and Hamilton (forthcoming) evaluated vector systems with no autoregressive dynamics in which both the mean

vector and the variance–covariance matrix were functions of the state:

$$y_t|s_t \sim \mathrm{N}(\boldsymbol{\mu}_{s_t}, \boldsymbol{\Omega}_{s_t}), \qquad s_t = 1,2.$$

Here n denotes the number of variables, $m = 0$, $K = 2$, and $\boldsymbol{\theta}$ consists of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and the diagonals and lower triangular blocks of $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$, with

$$p(y_t|z_t;\boldsymbol{\theta}) = \frac{1}{[2\pi]^{(n/2)}|\boldsymbol{\Omega}_{s_t}|^{(1/2)}} \exp\left[\frac{-[y_t - \boldsymbol{\mu}_{s_t}]'\boldsymbol{\Omega}_{s_t}^{-1}[y_t - \boldsymbol{\mu}_{s_t}]}{2}\right]. \tag{2.3}$$

The formulation (2.1) also includes a generalization of Engle's (1982) ARCH model to allow for occasional discrete shifts in the ARCH parameters. As another example of (2.1), one could generalize a vector autoregression so as to allow the constant terms, the autoregressive coefficients and innovation variance–covariance matrix to be functions of the state s_t.

I postulate that the transition between states is governed by a Markov chain whose realizations take on values in $\{1, \ldots, K\}$:

$$p(s_t = j|s_{t-1} = i) = p_{ij}, \tag{2.4}$$

with, of course,

$$\sum_{j=1}^{K} p_{ij} = 1 \quad \text{for} \quad i = 1, \ldots, K. \tag{2.5}$$

This Markov chain is presumed to be independent of lagged y once one conditions on s_{t-1}; for example, s_t is independent of ε_{t-1} in (2.2). Let $\boldsymbol{p} = (p_{1,1}, p_{1,2}, \ldots, p_{K,K})'$ denote the $(K^2 \times 1)$ vector of Markov transition probabilities.

The assumption of scalar-valued, first-order Markovian dynamics for the unobserved regimes is not restrictive. A vector-valued or higher-order system can always be rewritten as a scalar-valued, first-order process.

The simplest example of (2.4) is that used in Hamilton (1989) and Engel and Hamilton (forthcoming), who consider a two-state, first-order Markov process:

$$\begin{aligned} p(s_t = 1|s_{t-1} = 1) &= p_{11}, \\ p(s_t = 2|s_{t-1} = 1) &= p_{12}, \\ p(s_t = 1|s_{t-1} = 2) &= p_{21}, \\ p(s_t = 2|s_{t-1} = 2) &= p_{22}, \end{aligned} \tag{2.6}$$

for which $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22})'$. Hassett (1988) employed a four-state system to capture the dynamics of a bivariate vector y_t. Again, the remarks here are based on (2.4) and are not restricted to the special case (2.6).

The econometrician is presumed to observe y_t but not the state s_t. Our task is to maximize the likelihood function of the observed data $p(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{p}, \boldsymbol{\theta})$ by choice of the population parameters $(\boldsymbol{p}, \boldsymbol{\theta})$.

It usually proves computationally simplest in time series autoregressions to replace the exact likelihood function $p(y_T, y_{T-1}, \ldots, y_1)$ with the likelihood conditional on the first m observations: $p(y_T, y_{T-1}, \ldots, y_{m+1} | y_m, y_{m-1}, \ldots, y_1)$. The same turns out to be true for the EM algorithm suggested here. In forming this conditional likelihood, one also needs to make assumptions about the probability law governing the initial unobserved states. It turns out to be computationally simplest to assume that the initial states were drawn from a separate probability distribution, whose parameters $(\boldsymbol{\rho})$ are unrelated to $\boldsymbol{p}$ and $\boldsymbol{\theta}$:

$$p(s_m, s_{m-1}, \ldots, s_1 | y_m, y_{m-1}, \ldots, y_1) = \rho_{s_m, s_{m-1}, \ldots, s_1}. \tag{2.7}$$

These population probabilities are collected in the vector $\boldsymbol{\rho}$:

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

Thus $\boldsymbol{\rho}$ is a $(K^m \times 1)$ vector when $m > 0$ and a $(K \times 1)$ vector when $m = 0$. The elements of $\boldsymbol{\rho}$ sum to unity and are to be estimated by maximum likelihood along with $\boldsymbol{p}$ and $\boldsymbol{\theta}$.

In addition to computational simplicity, allowing the probability distribution for initial unobserved states $(s_m, s_{m-1}, \ldots, s_1)$ to be freely parameterized through the vector $\boldsymbol{\rho}$ rather than tied down by the values indicated by an ergodic draw from the underlying Markov chain represented by $\boldsymbol{p}$ has the further advantage of allowing for the possibility of a permanent change in regime.

We collect the unknown parameters to be estimated in a single vector $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} \equiv (\boldsymbol{p}', \boldsymbol{\theta}', \boldsymbol{\rho}')'.$$

As noted above, the econometrician is presumed to observe y_t but not the state s_t. There are thus three problems of interest. The first is the problem of *inference*: what does the econometrician conclude about the value of the regime at date t based on observations of y through date $\tau \geq t$? The second

is the problem of *forecasting*: what is the best forecast of y_{t+j} given observation of $y_t, y_{t-1}, \ldots, y_1$? The third is the problem of *parameter estimation*: what values of p, θ, and ρ are most consistent with the observed data $y_1, \ldots, y_T$?

My 1989 paper proposed a solution to these three problems based on a recursive, nonlinear filter. A simple version of this filter would calculate the probability that the unobserved state took on some particular value s_t at date t based on observation of y through that date:

$$p(s_t|y_t, y_{t-1}, \ldots, y_1; \boldsymbol{\lambda}). \tag{2.8}$$

A byproduct of this recursion is evaluation of the sample likelihood,

$$p(y_T, y_{T-1}, \ldots, y_{m+1}|y_m, y_{m-1}, \ldots, y_1; \boldsymbol{\lambda}). \tag{2.9}$$

My earlier papers recommended that (2.9) be maximized numerically with respect to $\boldsymbol{\lambda}$ to generate estimates of the unknown parameters.

This paper proposes an alternative algorithm for obtaining the same maximum likelihood estimates $\hat{\boldsymbol{\lambda}}$. This algorithm makes use of the 'smoothed' inference about unobserved states, for example, the probability that the unobserved state took on some particular value s_t at date t based on observation of y over the entire sample:

$$p(s_t|y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\lambda}). \tag{2.10}$$

All that is necessary to implement the EM algorithm is evaluation of smoothed probabilities such as (2.10), which can be calculated from a simple iterative processing of the data. No auxiliary software for numerical optimization or calculation of gradients is needed. One simply uses probabilities such as (2.10) to reweight the observed data y_t. Calculation of simple sample statistics of OLS regressions on the weighted data then generate new estimates of $\boldsymbol{\lambda}$. These new estimates are then used to recalculate the smoothed probabilities (2.10), and the data are reweighted with the new probabilities. Each such calculation of probabilities and reweighting the data can be shown to increase the value of the likelihood function. The process is repeated until a fixed point for $\boldsymbol{\lambda}$ is found, and this turns out to be the maximum likelihood estimate.

Several examples presenting the details of these calculations are provided in section 5, and the reader is invited to skip there immediately if uninterested in the formal demonstration that such an iterative reweighting of the data leads to maximum likelihood estimates. Section 3 states the general nature of the EM algorithm and reviews why it leads to maximum likelihood estimates [the reader is referred to Ruud (1988) for additional details and

some alternative perspectives]. Section 4 summarizes the algebraic results necessary to implement the EM principle on the general class of Markov switching dynamic models presented above.

3. A general characterization of the EM algorithm and its properties

3.1. Notation

Let $\mathscr{Y}$ denote the vector of observations:

$$\mathscr{Y} = (y_T', y_{T-1}', \ldots, y_1')'.$$

Our objective is to choose the vector of parameters $\boldsymbol{\lambda} = (\boldsymbol{p}', \boldsymbol{\theta}', \boldsymbol{\rho}')'$ so as to maximize the conditional likelihood,

$$p(\mathscr{Y}; \boldsymbol{\lambda}) \equiv p(y_T, y_{T-1}, \ldots, y_{m+1} | y_m, y_{m-1}, \ldots, y_1; \boldsymbol{\lambda}). \tag{3.1}$$

We can most easily characterize the structure of this maximum likelihood estimation if we consider the hypothetical joint likelihood function for unobserved states (s_t) and observed data (y_t). Define the $(T \times 1)$ vector $\mathscr{S}$ to be the realization of the unobserved states for the entire sample:

$$\mathscr{S} = (s_T, s_{T-1}, \ldots, s_1)'.$$

Though $\mathscr{S}$ is not observed by the econometrician, it is straightforward to characterize what the joint distribution of $\mathscr{Y}$ and $\mathscr{S}$ would look like if $\mathscr{S}$ were observed [see eq. (A.1) in appendix A]:

$$p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda})$$

$$\equiv p(y_T, y_{T-1}, \ldots, y_{m+1}, s_T, s_{T-1}, \ldots, s_1 | y_m, y_{m-1}, \ldots, y_1; \boldsymbol{\lambda}). \tag{3.2}$$

I should emphasize that one does *not* need to calculate (3.2) in order to use the EM algorithm; I will use expression (3.2) solely as a theoretical construct for expositing what the EM algorithm is and why it works. From this perspective, one can think of the marginal likelihood function, $p(\mathscr{Y}; \boldsymbol{\lambda})$, as simply the summation of the joint likelihood $p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda})$ over all possible values of $\mathscr{S}$:

$$p(\mathscr{Y}; \boldsymbol{\lambda}) = \int_{\mathscr{S}} p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}), \tag{3.3}$$

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

the elements of $\mathscr{S}$:

$$\int_{\mathscr{S}} f(\mathscr{S}) \equiv \sum_{s_T=1}^{K} \sum_{s_{T-1}=1}^{K} \cdots \sum_{s_1=1}^{K} f(s_T, s_{T-1}, \ldots, s_1).$$

Again, (3.3) is not the expression one would use to evaluate the actual likelihood, but is a representation of the sample likelihood in terms of the theoretical construct $p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda})$.

As a final piece of notation, it will prove useful to use the expression $Q(\boldsymbol{\lambda}_{l+1}; \boldsymbol{\lambda}_l, \mathscr{Y})$ to denote the expected log-likelihood, where the log-likelihood is parameterized by $\boldsymbol{\lambda}_{l+1}$ and the expectation is taken with respect to a second distribution parameterized by $\boldsymbol{\lambda}_l$:

$$Q(\boldsymbol{\lambda}_{l+1}; \boldsymbol{\lambda}_l, \mathscr{Y}) = \int_{\mathscr{S}} \log p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}_{l+1}) \cdot p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}_l). \tag{3.4}$$

3.2. *Two views of the EM algorithm*

There are two ways to characterize the EM algorithm for arriving at the MLE $\hat{\boldsymbol{\lambda}}$. The first characterization conceives of a sequence of optimization problems (indexed by $l = 1, 2, \ldots$), each of whose analytic solution $(\hat{\boldsymbol{\lambda}}_l)$ is found exactly. The solution to optimization problem $l + 1$ [denoted $(\hat{\boldsymbol{\lambda}}_{l+1})$] by construction increases the value of the likelihood function relative to the value for $\hat{\boldsymbol{\lambda}}_l$ (see Observation 1 below). The limit of this sequence of estimators achieves a local maximum of the likelihood function (see Observation 2):

$$\lim_{l \to \infty} \hat{\boldsymbol{\lambda}}_l = \hat{\boldsymbol{\lambda}}_{\text{MLE}}.$$

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

I now take up each of these interpretations in turn.

3.3. *The EM algorithm as the analytic solution to a sequence of optimization problems*

Let $\hat{\boldsymbol{\lambda}}_l$ denote the estimate of the parameter vector resulting from our previous iteration, with $\hat{\boldsymbol{\lambda}}_0$ an arbitrary initial guess at the parameter vector. We choose for $\hat{\boldsymbol{\lambda}}_{l+1}$ the value of $\boldsymbol{\lambda}_{l+1}$ that maximizes $Q(\boldsymbol{\lambda}_{l+1}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y})$ given in (3.4); that is, $\hat{\boldsymbol{\lambda}}_{l+1}$ satisfies

$$\int_{\mathscr{S}} \frac{\partial \log p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}_{l+1})}{\partial \boldsymbol{\lambda}_{l+1}}\bigg|_{\boldsymbol{\lambda}_{l+1}=\hat{\boldsymbol{\lambda}}_{l+1}} \cdot p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_l) = \mathbf{0}. \tag{3.5}$$

I show in section 4 how (3.5) can be solved analytically for $\hat{\boldsymbol{\lambda}}_{l+1}$ as a function of $\mathscr{Y}$ and $\hat{\boldsymbol{\lambda}}_l$.

The following results are well-known from other applications of the EM principle and are easy to demonstrate here.

Observation 1. $\hat{\boldsymbol{\lambda}}_{l+1}$ *is associated with a higher value of the likelihood function than is* $\hat{\boldsymbol{\lambda}}_l$*; that is,*

$$p(\mathscr{Y}; \hat{\boldsymbol{\lambda}}_{l+1}) \geq p(\mathscr{Y}; \hat{\boldsymbol{\lambda}}_l),$$

with equality only if $\hat{\boldsymbol{\lambda}}_{l+1} = \hat{\boldsymbol{\lambda}}_l$.

Proof. Following the arguments in Liporace (1982, p. 731), we know by construction that $\hat{\boldsymbol{\lambda}}_{l+1}$ maximizes $Q(\boldsymbol{\lambda}_{l+1}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y})$, so in particular

$$Q(\hat{\boldsymbol{\lambda}}_{l+1}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y}) \geq Q(\hat{\boldsymbol{\lambda}}_l; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y}),$$

with equality only if $\hat{\boldsymbol{\lambda}}_{l+1} = \hat{\boldsymbol{\lambda}}_l$. Recall that for any positive scalar x, $\log(x) \leq (x-1)$, with equality only if $x = 1$ [this follows from the strict concavity of $\log(.)$ and the point of tangency between $y = \log(x)$ and the line $y = x - 1$ at the point $x = 1$]. Thus

$$\begin{aligned}
&Q(\hat{\boldsymbol{\lambda}}_{l+1}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y}) - Q(\hat{\boldsymbol{\lambda}}_l; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y}) \\
&= \int_{\mathscr{S}} \log\left[\frac{p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_{l+1})}{p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_l)}\right] \cdot p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_l) \\
&\leq \int_{\mathscr{S}} \left[\frac{p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_{l+1})}{p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_l)} - 1\right] \cdot p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_l) \\
&= \int_{\mathscr{S}} \left[p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_{l+1}) - p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_l)\right] \\
&= p(\mathscr{Y}; \hat{\boldsymbol{\lambda}}_{l+1}) - p(\mathscr{Y}; \hat{\boldsymbol{\lambda}}_l).
\end{aligned}$$

Thus if $Q(\hat{\boldsymbol{\lambda}}_{l+1}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y}) > Q(\hat{\boldsymbol{\lambda}}_l; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y})$, then $p(\mathscr{Y}; \hat{\boldsymbol{\lambda}}_{l+1}) > p(\mathscr{Y}; \hat{\boldsymbol{\lambda}}_l)$, which was to be shown. Q.E.D.

Next I demonstrate that the sequence $\{\hat{\boldsymbol{\lambda}}_l\}_{l=1}^{\infty}$ converges to the (local) MLE. More accurately, if the first-order conditions for maximizing $Q(\boldsymbol{\lambda}_{l+1}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y})$ are satisfied by choosing $\boldsymbol{\lambda}_{l+1} = \hat{\boldsymbol{\lambda}}_l$, then the first-order conditions for maximizing $p(\mathscr{Y}; \boldsymbol{\lambda})$ are satisfied by choosing $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_l$.

Observation 2. If

$$\left.\frac{\partial Q(\boldsymbol{\lambda}_{l+1}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y})}{\partial \boldsymbol{\lambda}_{l+1}}\right|_{\boldsymbol{\lambda}_{l+1} = \hat{\boldsymbol{\lambda}}_l} = \mathbf{0},$$

then

$$\left.\frac{\partial p(\mathscr{Y}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_l} = \mathbf{0}.$$

Proof.

$$\left.\frac{\partial Q(\boldsymbol{\lambda}_{l+1}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y})}{\partial \boldsymbol{\lambda}_{l+1}}\right|_{\boldsymbol{\lambda}_{l+1} = \hat{\boldsymbol{\lambda}}_l}$$

$$= \int_{\mathscr{S}} \left\{ \frac{\partial p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}_{l+1})}{\partial \boldsymbol{\lambda}_{l+1}} \cdot \frac{1}{p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}_{l+1})} \right\} \Bigg|_{\boldsymbol{\lambda}_{l+1} = \hat{\boldsymbol{\lambda}}_l} \cdot p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_l)$$

$$= \int_{\mathscr{S}} \left. \frac{\partial p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}_{l+1})}{\partial \boldsymbol{\lambda}_{l+1}} \right|_{\boldsymbol{\lambda}_{l+1} = \hat{\boldsymbol{\lambda}}_l}$$

$$= \left. \frac{\partial p(\mathscr{Y}; \boldsymbol{\lambda}_{l+1})}{\partial \boldsymbol{\lambda}_{l+1}} \right|_{\boldsymbol{\lambda}_{l+1} = \hat{\boldsymbol{\lambda}}_l}.$$

So if the left-hand side is zero, so must be the right-hand-side as well. Q.E.D.

This completes the first perspective on what the EM algorithm is and why it yields the maximum likelihood estimate $\hat{\boldsymbol{\lambda}}$. I now turn to the second perspective.

3.4. *The EM algorithm as replacing unobserved scores with their conditional expectation*

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

$$\left.\frac{\partial \log p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}(\mathscr{S})} = \mathbf{0}. \tag{3.6}$$

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

$$p(\mathscr{S}|\mathscr{Y}; \hat{\boldsymbol{\lambda}}_l) = \frac{p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_l)}{p(\mathscr{Y}; \hat{\boldsymbol{\lambda}}_l)}. \tag{3.7}$$

For each of the K^T possible values for $\mathscr{S}$, there is a corresponding particular first-order condition (3.6). If we weight each of these first-order conditions (3.6) by the probability (3.7) that $\mathscr{S}$ took on that particular value, we would be choosing $\boldsymbol{\lambda}$ so as to satisfy

$$\int_{\mathscr{S}} \frac{\partial \log p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \cdot \frac{p(\mathscr{Y}, \mathscr{S}; \hat{\boldsymbol{\lambda}}_l)}{p(\mathscr{Y}; \hat{\boldsymbol{\lambda}}_l)} = \mathbf{0}$$

or

$$\frac{1}{p(\mathscr{Y}; \hat{\boldsymbol{\lambda}}_l)} \cdot \frac{\partial Q(\boldsymbol{\lambda}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y})}{\partial \boldsymbol{\lambda}} = \mathbf{0},$$

which of course is equivalent to

$$\frac{\partial Q(\boldsymbol{\lambda}; \hat{\boldsymbol{\lambda}}_l, \mathscr{Y})}{\partial \boldsymbol{\lambda}} = \mathbf{0}.$$

Thus the estimate $\hat{\boldsymbol{\lambda}}_{l+1}$ defined earlier as the value that maximizes $Q(\boldsymbol{\lambda}; \hat{\boldsymbol{\lambda}}_l; \mathscr{Y})$ with respect to $\boldsymbol{\lambda}$ is now seen also to be the estimate that would result if we weighted the first-order conditions associated with direct observa-

tion of $\mathscr{S}$ [eq. (3.6)] by the probability that $\mathscr{S}$ took on each of its feasible values.[1]

This completes our review of general principles of the EM algorithm. The next section derives the algebraic results necessary to apply it in the class of Markov switching vector autoregressions considered in this paper.

4. The particular form of the EM algorithm

Appendix A shows that for the model given by (2.1) and (2.4), the EM expression (3.4) is maximized by choosing $\boldsymbol{\lambda}_{l+1} = (\boldsymbol{p}'_{l+1}, \boldsymbol{\theta}'_{l+1}, \boldsymbol{\rho}'_{l+1})'$ to satisfy

$$p_{ij}^{(l+1)} = \frac{\sum_{t=m+1}^{T} p(s_t = j, s_{t-1} = i | \mathscr{Y}; \boldsymbol{\lambda}_l)}{\sum_{t=m+1}^{T} p(s_{t-1} = i | \mathscr{Y}; \boldsymbol{\lambda}_l)}, \qquad i, j = 1, \ldots, K, \tag{4.1}$$

$$\sum_{t=m+1}^{T} \sum_{s_t=1}^{K} \cdots \sum_{s_{t-m}=1}^{K} \frac{\partial \log p(y_t | z_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{l+1}} \cdot p(s_t, \ldots, s_{t-m} | \mathscr{Y}; \boldsymbol{\lambda}_l) = \mathbf{0}, \tag{4.2}$$

$$\rho_{i_m, i_{m-1}, \ldots, i_1}^{(l+1)} = p(s_m = i_m, s_{m-1} = i_{m-1}, \ldots, s_1 = i_1 | \mathscr{Y}; \boldsymbol{\lambda}_l), \tag{4.3}$$

$$i_1, \ldots, i_m = 1, \ldots, K.$$

Thus the EM algorithm begins at iteration $l = 0$ with an arbitrary guess for the parameter vector $\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_0$. For this guess we calculate the smoothed probabilities $p(s_t, \ldots, s_{t-m} | \mathscr{Y}; \boldsymbol{\lambda}_0)$. Eqs. (4.1)–(4.3) are then solved for $\boldsymbol{\lambda}_{l+1} = \boldsymbol{\lambda}_1$. The next iteration ($l = 1$) takes $\boldsymbol{\lambda}_l$ to be the value $\boldsymbol{\lambda}_1$ calculated from the previous iteration, and solves eqs. (4.1)–(4.3) for $\boldsymbol{\lambda}_{l+1} = \boldsymbol{\lambda}_2$. The process continues until a fixed point $\boldsymbol{\lambda}_{l+1} = \boldsymbol{\lambda}_l$ is satisfactorily approximated.

Calculation of $\boldsymbol{\lambda}_{l+1}$ as a function of $\boldsymbol{\lambda}_l$ is quite straightforward. Once one has calculated smoothed probabilities such as $p(s_t = j, s_{t-1} = i | \mathscr{Y}; \boldsymbol{\lambda}_l)$, eqs.

[1] Obviously when the scores $\partial \log p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}) / \partial \boldsymbol{\lambda}$ are not a linear function of $\mathscr{S}$, this is not the same as simply replacing the unobserved variable $\mathscr{S}$ in (3.6) with its expectation. This seems to be the basis for Ruud's (1988) objection to describing the EM algorithm as replacing unobserved latent variables with their expectation.

(4.1) and (4.3) allow calculation of $\boldsymbol{p}_{l+1}$ and $\boldsymbol{\rho}_{l+1}$ quite trivially. As shown in the examples in section 5, eq. (4.2) often also has a simple closed-form solution for $\boldsymbol{\theta}_{l+1}$.

Thus it should be clear that in order to *implement* the EM algorithm, it is not at all necessary to calculate such cumbersome expressions as $p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda})$ or $Q(\boldsymbol{\lambda}_{l+1}; \boldsymbol{\lambda}_l, \mathscr{Y})$. Rather, all one ever needs to evaluate are the smoothed inferences about the unobserved state:

$$p(s_t, s_{t-1}, \ldots, s_{t-m} | \mathscr{Y}; \boldsymbol{\lambda}_l).$$

Appendix B presents an algorithm for calculating these smoothed probabilities.

One convergence criterion is to stop when the maximal element of $|\boldsymbol{\lambda}_{l+1} - \boldsymbol{\lambda}_l|$ is less than, say, 10^{-8}. I have encountered no difficulties with this as a convergence criterion in practice. However, experience with the EM algorithm in other settings suggests that it might be wise to confirm convergence to the MLE by verifying directly that the first-order conditions for maximization of the log-likelihood are satisfied. As Ruud (1988) has emphasized, the derivative of the log-likelihood has effectively already been calculated in implementing the EM algorithm, since, as we saw in the proof of Observation 2,

$$\left.\frac{\partial p(\mathscr{Y}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_l} = \left.\frac{\partial Q(\boldsymbol{\lambda}_{l+1}; \boldsymbol{\lambda}_l, \mathscr{Y})}{\partial \boldsymbol{\lambda}_{l+1}}\right|_{\boldsymbol{\lambda}_{l+1}=\boldsymbol{\lambda}_l},$$

implying

$$\left.\frac{\partial \log p(\mathscr{Y}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_l} = \frac{1}{p(\mathscr{Y}; \boldsymbol{\lambda}_l)} \cdot \left.\frac{\partial Q(\boldsymbol{\lambda}_{l+1}; \boldsymbol{\lambda}_l, \mathscr{Y})}{\partial \boldsymbol{\lambda}_{l+1}}\right|_{\boldsymbol{\lambda}_{l+1}=\boldsymbol{\lambda}_l}. \tag{4.4}$$

Thus stopping criteria based on the gradient of the log-likelihood function can be implemented in a straightforward fashion.

5. Examples

Example 1

In this first example, I consider y_t a vector process in which s_t follows a two-state, first-order Markov process. Both the mean and the variance of $y_t | s_t$ may be functions of the state s_t, though there are no autoregressive

dynamics ($m = 0$). Thus $\boldsymbol{\theta}$ consists of the elements of the mean vectors $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and variance–covariance matrices $(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2)$ associated with the two states, $\boldsymbol{p}$ is summarized by the two parameters p_{11} and p_{22} with $p_{ij} = p(s_t = j | s_{t-1} = i)$, and $\boldsymbol{\rho}$ is summarized by the scalar $\rho = p(s_1 = 1)$ [the other probability parameters are known from $p_{12} = 1 - p_{11}$, $p_{21} = 1 - p_{22}$, and $p(s_1 = 2) = 1 - \rho$]. For this example $p(\boldsymbol{y}_t | \boldsymbol{z}_t; \boldsymbol{\theta}) = p(\boldsymbol{y}_t | s_t; \boldsymbol{\theta})$ and is given by (2.3). Differentiating (2.3) we find[2]

$$\frac{\partial \log p(\boldsymbol{y}_t | \boldsymbol{z}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_j} = \boldsymbol{\Omega}_j^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_j) \quad \text{if} \quad s_t = j,$$
$$= \mathbf{0} \qquad \text{otherwise}, \tag{5.1}$$

$$\frac{\partial \log p(\boldsymbol{y}_t | \boldsymbol{z}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\Omega}_j^{-1}} = \tfrac{1}{2}\boldsymbol{\Omega}_j - \tfrac{1}{2}(\boldsymbol{y}_t - \boldsymbol{\mu}_j)(\boldsymbol{y}_t - \boldsymbol{\mu}_j)' \quad \text{if} \quad s_t = j,$$
$$= \mathbf{0} \qquad \text{otherwise}. \tag{5.2}$$

Thus for this example expressions (4.2) take the form

$$\sum_{t=1}^{T} \left[\boldsymbol{\Omega}_j^{(l+1)}\right]^{-1} \left(\boldsymbol{y}_t - \boldsymbol{\mu}_j^{(l+1)}\right) \cdot p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_l) = \mathbf{0} \tag{5.3}$$

and

$$\sum_{t=1}^{T} \left(\tfrac{1}{2}\boldsymbol{\Omega}_j^{(l+1)} - \tfrac{1}{2}\left(\boldsymbol{y}_t - \boldsymbol{\mu}_j^{(l+1)}\right)\left(\boldsymbol{y}_t - \boldsymbol{\mu}_j^{(l+1)}\right)'\right) \cdot p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_l) = \mathbf{0}, \tag{5.4}$$

for $j = 1, 2$. Solving (5.3) and (5.4) for $\boldsymbol{\mu}_j^{(l+1)}$ and $\boldsymbol{\Omega}_j^{(l+1)}$, respectively, the EM

[2]Here I have followed the recommendation of Theil (1971, p. 32) in first taking derivatives and then imposing the symmetry condition [the (i, j)th element of $\boldsymbol{\Omega}_k$ is equal to the (j, i)th element]. I further write the derivative in the form of a matrix for convenience. Formulas for matrix derivatives are from Theil (1971, pp. 31–32).

equations are thus

$$\boldsymbol{\mu}_j^{(l+1)} = \frac{\sum_{t=1}^{T} \mathbf{y}_t \cdot p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_l)}{\sum_{t=1}^{T} p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_l)}, \quad j = 1, 2, \tag{5.5}$$

$$\boldsymbol{\Omega}_j^{(l+1)} = \frac{\sum_{t=1}^{T} \left(\mathbf{y}_t - \boldsymbol{\mu}_j^{(l+1)}\right)\left(\mathbf{y}_t - \boldsymbol{\mu}_j^{(l+1)}\right)' \cdot p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_l)}{\sum_{t=1}^{T} p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_l)}, \quad j = 1, 2, \tag{5.6}$$

$$p_{11}^{(l+1)} = \frac{\sum_{t=2}^{T} p(s_t = 1, s_{t-1} = 1 | \mathscr{Y}; \boldsymbol{\lambda}_l)}{\sum_{t=2}^{T} p(s_{t-1} = 1 | \mathscr{Y}; \boldsymbol{\lambda}_l)}, \tag{5.7a}$$

$$p_{22}^{(l+1)} = \frac{\sum_{t=2}^{T} p(s_t = 2, s_{t-1} = 2 | \mathscr{Y}; \boldsymbol{\lambda}_l)}{\sum_{t=2}^{T} p(s_{t-1} = 2 | \mathscr{Y}; \boldsymbol{\lambda}_l)}, \tag{5.7b}$$

$$\rho^{(l+1)} = p(s_1 = 1 | \mathscr{Y}; \boldsymbol{\lambda}_l), \tag{5.8}$$

where (5.7) comes from (4.1) and (5.8) comes from (4.3).

The EM algorithm for calculating the maximum likelihood estimates for Example 1 is thus as follows. Begin with an initial guess for $\boldsymbol{\lambda}$ (denoted $\boldsymbol{\lambda}_0$), consisting of the means and variances for the two different states, $(\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \boldsymbol{\Omega}_1^{(0)}, \boldsymbol{\Omega}_2^{(0)})$, the Markov transition probabilities ($p_{11}^{(0)}$ and $p_{22}^{(0)}$), and the initial state probability $\rho^{(0)}$. For these parameter values and for $\mathscr{Y} \equiv (\mathbf{y}_T, \ldots, \mathbf{y}_1)$, the full sample of observations on $\mathbf{y}$, calculate the smoothed probabilities $p(s_t = j, s_{t-1} = i | \mathscr{Y}; \boldsymbol{\lambda}_0)$ and $p(s_{t-1} = i | \mathscr{Y}; \boldsymbol{\lambda}_0)$ as described in appendix B. These smoothed probabilities are then used in eq. (5.5) to calculate $\boldsymbol{\mu}_1^{(1)}$ as the weighted average of the raw data $\mathbf{y}$, with weights on $\mathbf{y}_t$ proportional to the smoothed probability that that date's observation came

from regime 1: $p(s_t = 1|\mathscr{Y}; \boldsymbol{\lambda}_0)$. Similarly, $\mu_2^{(1)}$ is given by a weighted average of y with weights proportional to $p(s_t = 2|\mathscr{Y}; \boldsymbol{\lambda}_0)$. The values $\boldsymbol{\mu}_1^{(1)}$ and $\boldsymbol{\mu}_2^{(1)}$ are next used in (5.6) to construct $\boldsymbol{\Omega}_1^{(1)}$ and $\boldsymbol{\Omega}_2^{(1)}$. We further obtain $p_{11}^{(1)}$, $p_{22}^{(1)}$, and $\rho^{(1)}$ from (5.7) and (5.8). We have now arrived at the new complete vector $\boldsymbol{\lambda}_1$, which is then used to calculate a new set of smoothed probabilities $p(s_t = j, s_{t-1} = i|\mathscr{Y}; \boldsymbol{\lambda}_1)$ and $p(s_{t-1} = i|\mathscr{Y}; \boldsymbol{\lambda}_1)$. These are now used on the right-hand sides of (5.5)–(5.8) to generate the updated estimates $\boldsymbol{\lambda}_2$. The process continues until (i) the change in parameter values between iteration l and $l+1$ is less than some target convergence criterion, e.g., the maximal element of $|\boldsymbol{\lambda}^{l+1} - \boldsymbol{\lambda}^l|$ is less than 10^{-8}, or (ii) the first-orderconditions for maximization of the likelihood function [found from (4.4)] are satisfied within some tolerance level. For this example, the gradient is simply given by (5.3) and (5.4), with '$l+1$' replaced by 'l':

$$\left.\frac{\partial \log p(\mathscr{Y}; \boldsymbol{\lambda})}{\partial \boldsymbol{\mu}_j}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_l} = \sum_{t=1}^{T} \left[\boldsymbol{\Omega}_j^{(l)}\right]^{-1} \left(\boldsymbol{y}_t - \boldsymbol{\mu}_j^{(l)}\right) \cdot p(s_t = j|\mathscr{Y}; \boldsymbol{\lambda}_l),$$

$$\left.\frac{\partial \log p(\mathscr{Y}; \boldsymbol{\lambda})}{\partial \boldsymbol{\Omega}_j^{-1}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_l} = \sum_{t=1}^{T} \left(\tfrac{1}{2}\boldsymbol{\Omega}_j^{(l)} - \tfrac{1}{2}\left(\boldsymbol{y}_t - \boldsymbol{\mu}_j^{(l)}\right)\left(\boldsymbol{y}_t - \boldsymbol{\mu}_j^{(l)}\right)'\right)$$

$$\cdot p(s_t = j|\mathscr{Y}; \boldsymbol{\lambda}_l).$$

Example 2

Consider now a vector-valued process with K states and again no autoregressive dynamics ($m = 0$). In contrast to Example 1, the variance of the process is the same for all states ($\boldsymbol{\Omega}_j = \boldsymbol{\Omega}$ for $j = 1, \ldots, K$). Then (5.1) and (5.2) become

$$\begin{aligned} \frac{\partial \log p(\boldsymbol{y}_t|\boldsymbol{z}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_j} &= \boldsymbol{\Omega}^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_j) \quad \text{if} \quad s_t = j, \\ &= \mathbf{0} \qquad\qquad\quad \text{otherwise}, \\ \frac{\partial \log p(\boldsymbol{y}_t|\boldsymbol{z}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\Omega}^{-1}} &= \tfrac{1}{2}\boldsymbol{\Omega} - \tfrac{1}{2}(\boldsymbol{y}_t - \boldsymbol{\mu}_{s_t})(\boldsymbol{y}_t - \boldsymbol{\mu}_{s_t})'. \end{aligned} \tag{5.9}$$

The EM equations for the means $\boldsymbol{\mu}_j$ for Example 2 then continue to be given by (5.5), the only difference being that now j runs from 1 to K. The estimate

of $\boldsymbol{\Omega}^{(l+1)}$ [which replaces (5.6) of Example 1] is now

$$\boldsymbol{\Omega}^{(l+1)} = \frac{\sum_{j=1}^{K}\sum_{t=1}^{T}\left(\mathbf{y}_t - \boldsymbol{\mu}_j^{(l+1)}\right)\left(\mathbf{y}_t - \boldsymbol{\mu}_j^{(l+1)}\right)' \cdot p(s_t = j|\mathscr{Y}; \boldsymbol{\lambda}_l)}{T}, \quad (5.10)$$

and the estimate of p_{ij} becomes

$$p_{ij}^{(l+1)} = \frac{\sum_{t=2}^{T} p(s_t = j, s_{t-1} = i|\mathscr{Y}; \boldsymbol{\lambda}_l)}{\sum_{t=2}^{T} p(s_{t-1} = i|\mathscr{Y}; \boldsymbol{\lambda}_l)}, \quad (5.11)$$

$$i = 1, \ldots, K, \quad j = 1, \ldots, K-1,$$

with[3] $p_{iK}^{(l+1)} = 1 - p_{i1}^{(l+1)} - p_{i2}^{(l+1)} - \cdots - p_{i,K-1}^{(l+1)}$. The initial probabilities $\rho_1, \ldots, \rho_K$ are updated as

$$\rho_j^{(l+1)} = p(s_1 = j|\mathscr{Y}; \boldsymbol{\lambda}_l), \qquad j = 1, \ldots, K-1, \quad (5.12)$$

with $\rho_K^{(l+1)} = 1 - \rho_1^{(l+1)} - \rho_2^{(l+1)} + \cdots - \rho_{K-1}^{(l+1)}$. The maximum likelihood estimates for Example 2 are thus obtained by starting with an initial guess for $[\boldsymbol{p}_0, \boldsymbol{\theta}_0, \boldsymbol{\rho}_0]$ and iterating on eqs. (5.5), (5.10), (5.11), and (5.12).

Example 3

Here I consider conditional maximum likelihood estimation of a scalar mth-order autoregression in which all of the parameters other than the variance are presumed to shift with the state s_t:

$$y_t = \alpha_{s_t} + \phi_{1,s_t} y_{t-1} + \phi_{2,s_t} y_{t-2} + \cdots + \phi_{m,s_t} y_{t-m} + \varepsilon_t, \quad (5.13)$$

with $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$. Eq. (5.13) can be compactly written

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_{s_t} + \varepsilon_t,$$

[3]One could in principle use (5.11) for $i, j = 1, \ldots, K$, since the probabilities calculated by the algebra of the algorithm sum to unity by construction. Nevertheless, given computer rounding errors and the repeated multiplications involved in the iterations, it is clearly preferable to code the estimates as indicated in the text.

where the $([m+1]\times 1)$ vectors $\boldsymbol{x}_t$ and $\boldsymbol{\beta}_{s_t}$ are defined by

$$\boldsymbol{x}_t \equiv (1, y_{t-1}, y_{t-2}, \ldots, y_{t-m})',$$

$$\boldsymbol{\beta}_{s_t} \equiv (\alpha_{s_t}, \phi_{1,s_t}, \phi_{2,s_t}, \ldots, \phi_{m,s_t})'.$$

For this example, (2.1) takes the form

$$p(y_t|z_t;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_{s_t})^2}{2\sigma^2}\right]$$

and

$$\frac{\partial \log p(y_t|z_t;\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_j} = \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_j)\cdot \boldsymbol{x}_t}{\sigma^2} \quad \text{if} \quad s_t = j,$$
$$= \mathbf{0} \qquad\qquad \text{otherwise,} \tag{5.14}$$

$$\frac{\partial \log p(y_t|z_t;\boldsymbol{\theta})}{\partial \sigma^{-2}} = [\sigma^2/2] - \left[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_{s_t})^2/2\right]. \tag{5.15}$$

Substitution of (5.14) and (5.15) into (4.2) yields

$$\sum_{t=m+1}^{T} \frac{\left(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_j^{(l+1)}\right)\cdot \boldsymbol{x}_t}{\sigma^2(l+1)} \cdot p(s_t = j|\mathscr{Y};\boldsymbol{\lambda}_l) = \mathbf{0}, \tag{5.16}$$

$$\tfrac{1}{2}[\sigma^2(l+1)]\cdot[T-m] = \sum_{t=m+1}^{T}\sum_{j=1}^{K} \frac{\left(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_j^{(l+1)}\right)^2}{2} \cdot p(s_t = j|\mathscr{Y};\boldsymbol{\lambda}_l). \tag{5.17}$$

The conditions for estimation of $\boldsymbol{p}$ continue to be given by (4.1), while the start-up probabilities $\boldsymbol{\rho}$ are estimated by

$$\rho_j^{(l+1)} = p(s_m = j|\mathscr{Y};\boldsymbol{\lambda}_l), \qquad j = 1, \ldots, K-1, \tag{5.18}$$

with $\rho_K^{(l+1)} = 1 - \rho_1^{(l+1)} - \rho_2^{(l+1)} - \cdots - \rho_{K-1}^{(l+1)}$.

The EM algorithm here is closely related to Kiefer's (1980) exposition of the i.i.d. switching regression case and proceeds as follows. Begin with initial guesses $\boldsymbol{\lambda}_0 = (\boldsymbol{p}_0, \boldsymbol{\theta}_0, \boldsymbol{\rho}_0)$, where $\boldsymbol{p}_0 = (p_{11}^{(0)}, p_{12}^{(0)}, \ldots, p_{KK}^{(0)})'$, $\boldsymbol{\theta}_0 = (\sigma^{(0)}, \boldsymbol{\beta}_1^{(0)\prime}, \boldsymbol{\beta}_2^{(0)\prime}, \ldots, \boldsymbol{\beta}_K^{(0)\prime})'$, and $\boldsymbol{\rho}_0 = (\rho_1^{(0)}, \rho_2^{(0)}, \ldots, \rho_K^{(0)})'$. Here $\boldsymbol{\beta}_j^{(0)}$ denotes the

iteration 0 estimate of the vector consisting of the constant term and autoregressive coefficients appropriate when s_t is in regime j. Consider updated parameter estimates for regime j $[\boldsymbol{\beta}_j^{(1)} = (\alpha_j^{(1)}, \phi_{1,j}^{(1)}, \phi_{2,j}^{(1)}, \ldots, \phi_{K,j}^{(1)})']$. These are found by multiplying each observation y_t and vector $\boldsymbol{x}_t$ by the square root of the probability that date t came from regime j:

$$y_t^{(1,j,*)} = y_t \cdot \sqrt{p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_0)} ,$$

$$\boldsymbol{x}_t^{(1,j,*)} = \boldsymbol{x}_t \cdot \sqrt{p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_0)} .$$

Then an OLS regression of $y_t^{(1,j,*)}$ on $\boldsymbol{x}_t^{(1,j,*)}$ yields an updated parameter vector $\boldsymbol{\beta}_j^{(1)}$ satisfying (5.16):

$$\sum_{t=m+1}^{T} \left(y_t - \boldsymbol{x}_t' \boldsymbol{\beta}_j^{(1)}\right) \cdot \boldsymbol{x}_t \cdot p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_0) = \mathbf{0}.$$

A set of K such regressions $(j = 1, \ldots, K)$ generates $(\boldsymbol{\beta}_1^{(1)}, \boldsymbol{\beta}_2^{(1)}, \ldots, \boldsymbol{\beta}_K^{(1)})$. By squaring and summing the residuals from all of these K regressions together, we obtain our new estimate of the variance, a solution to (5.17):

$$\sigma^2(1) = \sum_{j=1}^{K} \sum_{t=m+1}^{T} \frac{\left[y_t^{(1,j,*)} - \boldsymbol{x}_t^{(1,j,*)\prime} \boldsymbol{\beta}_j^{(1)}\right]^2}{T - m} .$$

This completes our iteration 1 inference for $\boldsymbol{\theta}_1$. Our updated inference for $\boldsymbol{p}_1$ and $\boldsymbol{\rho}_1$ are obtained from (4.1) and (5.18) as before.

Armed with our new estimates $\boldsymbol{p}_1, \boldsymbol{\theta}_1, \boldsymbol{\rho}_1$, we now repeat the process. Calculate the smoothed probabilities $p(s_t = j | \mathscr{Y}; \boldsymbol{\lambda}_1)$ and multiply by the square root of these to obtain $y_t^{(2,j,*)}$ and $\boldsymbol{x}_t^{(2,j,*)}$; OLS regression of $y_t^{(2,j,*)}$ on $\boldsymbol{x}_t^{(2,j,*)}$ yields $\boldsymbol{\beta}_j^{(2)}$ and summing the squared residuals over j and t provides $\sigma^2(2)$. We again use (4.1) and (5.18) to obtain $\boldsymbol{p}_2$ and $\boldsymbol{\rho}_2$. The process continues until a fixed point of the EM algorithm (or critical point of the log-likelihood function) is found. This fixed point represents a local maximum of the conditional likelihood $p(y_T, \ldots, y_{m+1} | y_m, \ldots, y_1; \boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{\rho})$ with $p(s_m; \boldsymbol{\rho})$ estimated as a vector of free parameters $\boldsymbol{\rho}$.

6. Extensions and additional results

6.1. *Bayesian priors*

When variances are allowed to differ across different states as in Example 1 above, a singularity in the likelihood function arises when, for example, the

mean of the first state is set exactly equal to the first observation ($\boldsymbol{y}_1 = \boldsymbol{\mu}_1$) and the variance of this state is allowed to vanish ($\boldsymbol{\Omega}_1 \to \mathbf{0}$). This problem is well-known in the literature on estimating mixtures of normal distributions with no autoregressive or Markovian dynamics [e.g., Everitt and Hand (1981)]. Moreover, for some data sets, the existence of multiple local maxima of the likelihood function can pose a significant problem for estimation.

These difficulties are reviewed in my 1988b paper, which suggests a simple solution. This approach is based on the observation that, in the fixed-point solution to eqs. (5.5)–(5.7), the econometrician is implicitly proceeding as if observation of $\boldsymbol{y}_t$ were equivalent to $p(s_t = j|\mathscr{Y}; \boldsymbol{\lambda})$ observations from each of the distributions $j = 1, \ldots, K$. In that paper I therefore proposed a normal-gamma Bayesian prior for $\boldsymbol{\mu}_j$ and $\boldsymbol{\Omega}_j$ of the form

$$\boldsymbol{\mu}_j|\boldsymbol{\Omega}_j \sim \mathrm{N}(\boldsymbol{m}_j, \boldsymbol{\Omega}_j/\nu_j), \tag{6.1a}$$

$$\boldsymbol{\Omega}_j^{-1} \sim \mathrm{W}(\alpha_j, \boldsymbol{\Lambda}_j). \tag{6.1b}$$

Here $\mathrm{W}(\alpha_j, \boldsymbol{\Lambda}_j)$ denotes a Wishart distribution with α_j degrees of freedom and $(n \times n)$ precision matrix $\boldsymbol{\Lambda}_j$. The terms $\boldsymbol{m}_j$, ν_j, α_j, and $\boldsymbol{\Lambda}_j$ summarize our prior for $\boldsymbol{\mu}_j, \boldsymbol{\Omega}_j$, which in turn parameterize the distribution of regime j. If we confront the prior (6.1) with $p(s_t = j|\mathscr{Y}; \boldsymbol{\lambda})$ observations equal to $\boldsymbol{y}_t$ (for $t = 1, \ldots, T$), posterior Bayesian inference would lead us to replace the fixed-point solutions to (5.5) and (5.6) with[4]

$$\hat{\boldsymbol{\mu}}_j = \frac{\nu_j \cdot \boldsymbol{m}_j + \sum_{t=1}^{T} \boldsymbol{y}_t \cdot p(s_t = j|\mathscr{Y}; \boldsymbol{\lambda})}{\nu_j + \sum_{t=1}^{T} p(s_t = j|\mathscr{Y}; \boldsymbol{\lambda})}. \tag{6.2a}$$

$$\begin{aligned}\hat{\boldsymbol{\Omega}}_j = &\left[\frac{1}{\alpha_j + \sum_{t=1}^{T} p(s_t = j|\mathscr{Y}; \boldsymbol{\lambda})}\right] \\ &\times \left[\boldsymbol{\Lambda}_j + \sum_{t=1}^{T} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_j)' \cdot p(s_t = j|\mathscr{Y}; \boldsymbol{\lambda}) \right. \\ &\left. + \nu_j \cdot (\boldsymbol{m}_j - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{m}_j - \hat{\boldsymbol{\mu}}_j)'\right].\end{aligned} \tag{6.2b}$$

[4]For a general discussion of use of the normal-gamma prior, see DeGroot (1970, p. 178). Details of this application are presented in Hamilton (1988b).

Approximately, one acts as if one had ν_j additional observations equal to $\boldsymbol{m}_j$ and α_j separate observations whose sum of squared residuals was $\boldsymbol{\Lambda}_j$.

Monte Carlo simulations suggest that relatively diffuse priors [such as $\alpha_j = \nu_j = 0.1$ and $\boldsymbol{\Lambda}_j = 0.1 \cdot \boldsymbol{\kappa}$, where $\boldsymbol{\kappa}$ is an $(n \times n)$ diagonal matrix reflecting the scale of the data] can consistently lead to improved estimates. Seeing if a local maximum is robust with respect to a slight strengthening of the prior (proportional increase in α_j, ν_j, and $\boldsymbol{\Lambda}_j$) further turns out to be a good test of whether a local maximum is indeed a global maximum [Hamilton (1988b)].

6.2. *Modeling discrete trends*

My 1989 paper proposed that models with discrete parameter shifts might help describe the trends in economic and financial time series. In the empirical application in that paper, y_t in (2.2) represented the quarterly rate of growth of real GNP. Rearranging terms, one could view this model as decomposing the log of GNP ($\tilde{y}_t$) into the sum of two components, each of which requires first-differencing in order to achieve stationarity:

$$\tilde{y}_t = \tilde{n}_t + \tilde{z}_t. \tag{6.3}$$

First-differences of $\tilde{n}_t$ produce the Markov component of the process,

$$(1 - L)\tilde{n}_t = \mu_{s_t}, \tag{6.4}$$

whereas differencing $\tilde{z}_t$ yields the Gaussian component:

$$(1 - L)\tilde{z}_t = \left[1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3 - \phi_4 L^4\right]^{-1} \cdot \varepsilon_t. \tag{6.5}$$

The probability law for s_t is given by (2.6) whereas $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$; roots of $[1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3 - \phi_4 z^4] = 0$ lie outside the unit circle.

This specification thus assumes that both the Markov component $\tilde{n}_t$ and the Gaussian component $\tilde{z}_t$ possess a unit root. Lam (1988) has rightly suggested that this misses a very interesting class of models in which the Markov component is integrated but the Gaussian component is not. Lam's framework could also be viewed as the generalization of Perron's (1989) specification of a stationary process around an occasionally shifting linear trend, to which Lam has added an explicit stochastic representation for this change in trend. Lam's model is thus

$$\tilde{y}_t = \tilde{n}_t + \left[1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3\right]^{-1} \cdot \varepsilon_t, \tag{5.6}$$

where roots of $[1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3] = 0$ lie outside the unit circle and $\tilde{n}_t$ is still as in (6.4). Thus realizations of s_t have permanent consequences for

the level of $\tilde{y}_t$ whereas ε_t has none; that is, $E_t\tilde{y}_{t+j}$ for j large depends on s_t but not ε_t.

Note that (6.6) may be written

$$\left[1-\phi_1 L-\phi_2 L^2-\phi_3 L^3\right](\tilde{y}_t-\tilde{n}_t)=\varepsilon_t. \tag{6.7}$$

Lam suggests that we think of an underlying state $\tilde{s}_t$ which represents the cumulative number of times that the event $s_\tau = 1$ has occurred during dates $\tau = 1, 2, \ldots, t$; thus $\tilde{s}_t$ takes on an integer value in $(0, 1, \ldots, t)$. The probability transition law for the vector $(\tilde{s}_t, s_t)'$ is readily written down from (2.6), and the distribution of $\tilde{y}_t$ conditional on $(\tilde{y}_{t-1}, \tilde{y}_{t-2}, \tilde{y}_{t-3}, \tilde{s}_{t-1}, \tilde{s}_{t-2}, \tilde{s}_{t-3})$ and $\tilde{n}_0$ is calculated as

$$\begin{aligned}\frac{1}{\sqrt{2\pi\sigma}}\cdot\exp\Big\{-\big\{&[\tilde{y}_t-\tilde{n}_0-\mu_1\tilde{s}_t-(t-\tilde{s}_t)\cdot\mu_2]\\&-\phi_1[\tilde{y}_{t-1}-\tilde{n}_0-\mu_1\tilde{s}_{t-1}-(t-1-\tilde{s}_{t-1})\cdot\mu_2]\\&-\phi_2[\tilde{y}_{t-2}-\tilde{n}_0-\mu_1\tilde{s}_{t-2}-(t-2-\tilde{s}_{t-2})\cdot\mu_2]\\&-\phi_3[\tilde{y}_{t-3}-\tilde{n}_0-\mu_1\tilde{s}_{t-3}-(t-3-\tilde{s}_{t-3})\cdot\mu_2]\big\}^2\\&\div[2\sigma^2]\Big\}.\end{aligned}$$

Lam shows how the principle underlying the recursion (B.2) and (B.3) can be applied to evaluate the likelihood function for this case.

6.3. *Hypothesis testing*

Kiefer (1978) shows that in the i.i.d. switching regression case, there exists a solution to the normal equations which is asymptotically efficient, permitting use of the second derivatives of the log-likelihood function to construct asymptotic standard errors if that result generalizes to the dynamic case considered here. Where there are multiple local maxima of the likelihood surface, there remains a problem of picking the 'right' maximum, for which the Bayesian procedure noted above is hoped to be of some use.

Even if the asymptotic variance of estimates is known, it remains problematic to test the null hypothesis that the data are characterized by no changes in regime. Taking the scalar version of Example 1 above, consider the null hypothesis

$$H_0: \quad \mu_1=\mu_2, \quad \sigma_1^2=\sigma_2^2.$$

Under H_0, the parameters p_{11} and p_{22} are unidentified. The asymptotic information matrix is thus singular under the null and standard regularity conditions for constructing an asymptotically valid test of H_0 fail to apply.

The form of the normal equations for maximizing the likelihood function [the fixed-point solutions to eqs. (5.5) and (5.6)] show that the problem is even more troublesome. Letting $\bar{y}$ denote the full-sample mean $(\sum_{t=1}^{T} y_t/T)$ and s^2 the variance $(\sum_{t=1}^{T}(y_t-\bar{y})^2/T)$, it is clear that $\mu_1=\mu_2=\bar{y}$ and $\sigma_1^2=\sigma_2^2=s^2$ satisfy the normal equations. Thus at the constrained MLE (maximization of the likelihood function subject to H_0), the score (the derivative of the log-likelihood with respect to the full vector of parameters available under the alternative hypothesis, namely $\mu_1, \mu_2, \sigma_1, \sigma_2, p_{11}, p_{22}$) is identically zero. Testing H_0 thus combines features of the problems studied by Watson and Engle (1985) and Lee and Chesher (1986).

One approach for this problem is an adaptation of Gallant's (1987, pp. 139–143) strategy. Consider a simple AR(m) model for forecasting y_t based on $y_{t-1}, \ldots, y_{t-m}$. The regime-switching specification embodies the claim that the optimal forecast of y_t is a nonlinear function $\hat{y}_{t|t-1}=f(y_{t-1}, y_{t-2}, \ldots; \boldsymbol{\lambda})$ (see my 1989 paper for details). The difficulty with comparing this against the AR(m) specification directly is that some of the elements of $\boldsymbol{\lambda}$ may be unidentified under the null. A simple idea is to Monte Carlo a set of l parameter values $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \ldots, \boldsymbol{\lambda}^{(l)})$, calculate the optimal forecast $\hat{y}_{t|t-1;\boldsymbol{\lambda}^{(j)}}=f(y_{t-1}, y_{t-2}, \ldots; \boldsymbol{\lambda}^{(j)})$ for each [or, if this involves too much calculation, simply the lagged filter probabilities $p(s_{t-1}=j|y_{t-1}, y_{t-2}, \ldots, y_1; \boldsymbol{\lambda}^{(j)}]$, and extract a few principal components of the l series $\hat{y}_{t|t-1;\boldsymbol{\lambda}^{(1)}}, \ldots, y_{t|t-1;\boldsymbol{\lambda}^{(l)}}$. The statistical significance of these principal components when added to the autoregression might then be examined using a standard F-test.

In some applications, one may be satisfied with testing a variant of H_0 that avoids the statistical complications raised by the nuisance parameters. For example, Engel and Hamilton (forthcoming) were interested in testing whether exchange rates follow a martingale against the alternative that exchange rate changes follow a Markov switching process. Here the null hypothesis could be represented as either

$$H_0': \quad \mu_1=\mu_2, \quad \sigma_1^2\neq\sigma_2^2, \quad p_{11}, p_{22} \text{ unrestricted}$$

or

$$H_0'': \quad \mu_1\neq\mu_2, \quad \sigma_1^2\neq\sigma_2^2, \quad p_{11}=1-p_{22}.$$

For either of these null hypotheses, exchange rate changes have a constant mean but come from a mixture of two normal distributions for which the nuisance parameters are identified under the null.

Appendix A

A.1. Derivation of (4.1)

Here I show that (4.1) maximizes (3.4) with respect to $\boldsymbol{p}_{t+1}$. Note first that

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

Differentiating (A.1) with respect to p_{ij} (a representative element of $\boldsymbol{p}$), we see

$$\frac{\partial p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda})}{\partial p_{ij}} = \sum_{t=m+1}^{T} \frac{\partial \log p(s_t | s_{t-1}; \boldsymbol{p})}{\partial p_{ij}} \cdot p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}). \tag{A.2}$$

But recall from (2.4) that

$$\begin{aligned} \frac{\partial \log p(s_t | s_{t-1}; \boldsymbol{p})}{\partial p_{ij}} &= \frac{1}{p_{ij}} \quad \text{if} \quad s_t = j \quad \text{and} \quad s_{t-1} = i, \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Thus, using $\delta_{[\,]}$ for the Kronecker delta (that is, $\delta_{[A]} = 1$ when the event A occurs and 0 otherwise), eq. (A.2) simplifies to

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

and so

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

It follows immediately from the definition of $Q(.)$ in (3.4) that

$$\frac{\partial Q(\boldsymbol{\lambda}_{l+1};\boldsymbol{\lambda}_l,\mathscr{Y})}{\partial p_{ij}^{(l+1)}} = \int_{\mathscr{S}} \frac{\partial \log p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda}_{l+1})}{\partial p_{ij}^{(l+1)}} \cdot p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda}_l)$$

$$= \int_{\mathscr{S}} \left[p_{ij}^{(l+1)}\right]^{-1} \cdot \left\{ \sum_{t=m+1}^{T} \delta_{[s_t=j,\, s_{t-1}=i]} \right\}$$

$$\cdot p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda}_l).$$

But notice

$$\int_{\mathscr{S}} \delta_{[s_t=j,\, s_{t-1}=i]} \cdot p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda}_l) = p(s_t=j, s_{t-1}=i|\mathscr{Y};\boldsymbol{\lambda}_l) \cdot p(\mathscr{Y};\boldsymbol{\lambda}_l),$$

and so

$$\frac{\partial Q(\boldsymbol{\lambda}_{l+1};\boldsymbol{\lambda}_l,\mathscr{Y})}{\partial p_{ij}^{(l+1)}} = \left[p_{ij}^{(l+1)}\right]^{-1} \sum_{t=m+1}^{T} p(s_t=j, s_{t-1}=i|\mathscr{Y};\boldsymbol{\lambda}_l)$$

$$\cdot p(\mathscr{Y};\boldsymbol{\lambda}_l). \tag{A.3}$$

Now, our task in the EM algorithm was to find the value of $\boldsymbol{p}_{l+1}$ for which $Q(\boldsymbol{\lambda}_{l+1};\boldsymbol{\lambda}_l,\mathscr{Y})$ was maximized. Imposing the constraint

$$\sum_{j=1}^{K} p_{ij}^{(l+1)} = 1, \tag{A.4}$$

we thus wish to form the Lagrangian

$$Q(\boldsymbol{\lambda}_{l+1};\boldsymbol{\lambda}_l,\mathscr{Y}) - \mu_i \left(\sum_{j=1}^{K} p_{ij}^{(l+1)} - 1 \right),$$

from which the first-order conditions are

$$\frac{\partial Q(\boldsymbol{\lambda}_{l+1};\boldsymbol{\lambda}_l,\mathscr{Y})}{\partial p_{ij}^{(l+1)}} = \mu_i \quad \text{for} \quad j = 1,\ldots,K. \tag{A.5}$$

Substituting (A.3) into (A.5), we see

$$\sum_{t=m+1}^{T} p(s_t=j, s_{t-1}=i|\mathscr{Y};\boldsymbol{\lambda}_l) = p_{ij}^{(l+1)}\mu_i/p(\mathscr{Y};\boldsymbol{\lambda}_l). \tag{A.6}$$

If we now sum (A.6) for $j = 1,\ldots, K$, we see that

$$\sum_{t=m+1}^{T}\sum_{j=1}^{K} p(s_t=j, s_{t-1}=i|\mathscr{Y};\boldsymbol{\lambda}_l) = \sum_{j=1}^{K} p_{ij}^{(l+1)}\mu_i/p(\mathscr{Y};\boldsymbol{\lambda}_l),$$

or using (A.4),

$$\sum_{t=m+1}^{T} p(s_{t-1}=i|\mathscr{Y};\boldsymbol{\lambda}_l) = \mu_i/p(\mathscr{Y};\boldsymbol{\lambda}_l). \tag{A.7}$$

Substituting (A.7) into (A.6) yields eq. (4.1), which was to be shown.

A.2. Derivation of (4.2)

Proceeding analogously, differentiating (A.1) with respect to $\boldsymbol{\theta}$ yields

$$\frac{\partial p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda})}{\partial\boldsymbol{\theta}} = \sum_{t=m+1}^{T} \frac{\partial \log p(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \cdot p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda}),$$

and so

$$\frac{\partial \log p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda})}{\partial\boldsymbol{\theta}} = \sum_{t=m+1}^{T} \frac{\partial \log p(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}.$$

The point to note here is that $p(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\theta})$ depends on s (through $\mathbf{z}_t$) at most for dates $t, t-1,\ldots, t-m$ [see eq. (2.1)]. Thus,

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\lambda}_{l+1};\boldsymbol{\lambda}_l,\mathscr{Y})}{\partial\boldsymbol{\theta}_{l+1}} &= \int_{\mathscr{S}} \frac{\partial \log p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda}_{l+1})}{\partial\boldsymbol{\theta}_{l+1}} \cdot p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda}_l) \\
&= \sum_{t=m+1}^{T} \int_{\mathscr{S}} \frac{\partial \log p(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\theta}_{l+1})}{\partial\boldsymbol{\theta}_{l+1}} \cdot p(\mathscr{Y},\mathscr{S};\boldsymbol{\lambda}_l) \\
&= \sum_{t=m+1}^{T}\sum_{s_t=1}^{K}\sum_{s_{t-1}=1}^{K}\cdots\sum_{s_{t-m}=1}^{K} \left\{ \frac{\partial \log p(\mathbf{y}_t|\mathbf{z}_t;\boldsymbol{\theta}_{l+1})}{\partial\boldsymbol{\theta}_{l+1}} \right. \\
&\qquad \left. \cdot p(s_t, s_{t-1},\ldots, s_{t-m}|\mathscr{Y};\boldsymbol{\lambda}_l) \cdot p(\mathscr{Y};\boldsymbol{\lambda}_l) \right\}.
\end{aligned} \tag{A.8}$$

Noting that $p(\mathscr{Y}; \boldsymbol{\lambda}_l)$ is not a function of the index t, it can be taken outside of the summation operators. Setting (A.8) equal to zero then yields eq. (4.2), as desired.

A.3. Derivation of (4.3)

Finally, differentiating (A.1) with respect to $\rho_{i_m, i_{m-1}, \ldots, i_1}$, we obtain

$$\frac{\partial \log p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda})}{\partial \rho_{i_m, i_{m-1}, \ldots, i_1}} = \left[\rho_{i_m, i_{m-1}, \ldots, i_1}\right]^{-1} \cdot \delta_{[s_m = i_m, s_{m-1} = i_{m-1}, \ldots, s_1 = i_1]},$$

and so

$$\frac{\partial Q(\boldsymbol{\lambda}_{l+1}; \boldsymbol{\lambda}_l, \mathscr{Y})}{\partial \rho^{(l+1)}_{i_m, i_{m-1}, \ldots, i_1}} = \int_{\mathscr{S}} \left[\rho^{(l+1)}_{i_m, i_{m-1}, \ldots, i_1}\right]^{-1} \cdot \delta_{[s_m = i_m, s_{m-1} = i_{m-1}, \ldots, s_1 = i_1]} \cdot p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}_l). \tag{A.9}$$

Maximizing $Q(\boldsymbol{\lambda}_{l+1}; \boldsymbol{\lambda}_l, \mathscr{Y})$ subject to the constraint that the sum of the elements of $\boldsymbol{\rho}_{l+1}$ equal unity yields first-order conditions

$$\frac{\partial Q(\boldsymbol{\lambda}_{l+1}; \boldsymbol{\lambda}_l, \mathscr{Y})}{\partial \rho^{(l+1)}_{i_m, i_{m-1}, \ldots, i_1}} = \mu,$$

for μ the Lagrange multiplier associated with the summation constraint. Thus,

$$\int_{\mathscr{S}} \delta_{[s_m = i_m, s_{m-1} = i_{m-1}, \ldots, s_1 = i_1]} \cdot p(\mathscr{Y}, \mathscr{S}; \boldsymbol{\lambda}_l) = \mu \cdot \left[\rho^{(l+1)}_{i_m, i_{m-1}, \ldots, i_1}\right],$$

or

$$p(s_m = i_m, s_{m-1} = i_{m-1}, \ldots, s_1 = i_1 | \mathscr{Y}; \boldsymbol{\lambda}_l) \cdot p(\mathscr{Y}; \boldsymbol{\lambda}_l)$$

$$= \mu \cdot \left[\rho^{(l+1)}_{i_m, i_{m-1}, \ldots, i_1}\right]. \tag{A.10}$$

Summing (A.10) over all possible values of $i_m, i_{m-1}, \ldots, i_1$, we obtain

$$p(\mathscr{Y}; \boldsymbol{\lambda}_l) = \mu. \tag{A.11}$$

Substituting (A.11) into (A.10) we deduce

$$p(s_m = i_m, s_{m-1} = i_{m-1}, \ldots, s_1 = i_1 | \mathcal{Y}; \boldsymbol{\lambda}_l) = \rho^{(l+1)}_{i_m, i_{m-1}, \ldots, i_1},$$

as claimed in (4.3).

Appendix B

Our theoretical exposition of the EM algorithm in section 3 made use of an expression $p(\mathcal{S}|\mathcal{Y}; \boldsymbol{\lambda})$, denoting an inference about the full vector $\mathcal{S}$ of unobserved states. Since there are K^T possible values for this vector $\mathcal{S}$, calculation of this magnitude is obviously hopeless unless K and T are quite small. Fortunately, one never needs to calculate $p(\mathcal{S}|\mathcal{Y}; \boldsymbol{\lambda})$. This is because the equations necessary to implement the EM algorithm [eqs. (4.1)–(4.3)] require at most an inference about a block of $(m + 1)$ consecutive values of s at a time; specifically, we only need

$$p(s_t, s_{t-1}, \ldots, s_{t-m} | \mathcal{Y}). \tag{B.1}$$

This probability is of course also a function of the parameter vector $\boldsymbol{\lambda}$, but for notational simplicity indication of the dependence on $\boldsymbol{\lambda}$ is suppressed throughout this appendix. In implementing computer code of the equations that follow, expression (B.1) takes the form of K^{m+1} distinct numbers, one for each of the possible values that $(s_t, s_{t-1}, \ldots, s_{t-m})$ could take; thus there is one number for $p(s_t = 1, s_{t-1} = 1, \ldots, s_{t-m} = 1 | \mathcal{Y})$, a second number for $p(s_t = 2, s_{t-1} = 1, \ldots, s_{t-m} = 1 | \mathcal{Y})$, and so on. Each of these numbers represents the probability that the event $(s_t, s_{t-1}, \ldots, s_{t-m})$ occurred given data observed through the full sample of observations on $\boldsymbol{y}$. These K^{m+1} numbers are all nonnegative and sum to unity by construction.

Actually, one only needs to keep track of the number of lags of s_t included in the vector $\boldsymbol{z}_t$ on which the probability $p(\boldsymbol{y}_t | \boldsymbol{z}_t; \boldsymbol{\theta})$ in eq. (2.1) depends. I analyze the most general case here (where the number of lags on s is equal to m, the number of lags in the conditional autoregression for y).

Let $\mathcal{Y}_t$ denote a vector consisting of all observations on $\boldsymbol{y}$ through date t:

$$\mathcal{Y}_t \equiv (\boldsymbol{y}_t', \boldsymbol{y}_{t-1}', \ldots, \boldsymbol{y}_1')',$$

so, for example, $\mathcal{Y} = \mathcal{Y}_T$.

As a first step in calculating the smoothed inferences (B.1) one should calculate and store the filter inferences

$$p(s_t, s_{t-1}, \ldots, s_{t-m} | \mathcal{Y}_t), \tag{B.2}$$

and conditional likelihoods

$$p(y_t|\mathscr{Y}_{t-1}). \tag{B.3}$$

Again, (B.2) represents K^{m+1} distinct values for each date t, while (B.3) is a scalar for each date t.

The values (B.3) and (B.2) are calculated by iterating on the following pair of equations for $t = m+2, m+3, \ldots, T$:

$$p(y_t|\mathscr{Y}_{t-1}) = \sum_{s_t=1}^{K} \sum_{s_{t-1}=1}^{K} \cdots \sum_{s_{t-m-1}=1}^{K} p(s_t|s_{t-1}) \cdot p(y_t|z_t)$$

$$\cdot p(s_{t-1}, s_{t-2}, \ldots, s_{t-m-1}|\mathscr{Y}_{t-1})$$

$$p(s_t, s_{t-1}, \ldots, s_{t-m}|\mathscr{Y}_t)$$

$$= \frac{\sum_{s_{t-m-1}}^{K} p(s_t|s_{t-1}) \cdot p(y_t|z_t) \cdot p(s_{t-1}, s_{t-2}, \ldots, s_{t-m-1}|\mathscr{Y}_{t-1})}{p(y_t|\mathscr{Y}_{t-1})}.$$

Here $p(s_t|s_{t-1})$ is given by the value in (2.4) that is appropriate for the particular pair (s_{t-1}, s_t) being analyzed. The term $p(y_t|z_t)$ is given by (2.1) and $p(s_{t-1}, s_{t-2}, \ldots, s_{t-m-1}|\mathscr{Y}_{t-1})$ is given by the previous step of the iteration. The iteration is initialized for $t = m+1$ by setting

$$p(y_{m+1}|\mathscr{Y}_m) = \sum_{s_{m+1}=1}^{K} \sum_{s_m=1}^{K} \cdots \sum_{s_1=1}^{K} p(s_{m+1}|s_m) \cdot p(y_{m+1}|z_{m+1})$$

$$\cdot \rho_{s_m, s_{m-1}, \ldots, s_1} \tag{B.4}$$

$$p(s_{m+1}, s_m, \ldots, s_1|\mathscr{Y}_{m+1})$$

$$= \frac{p(s_{m+1}|s_m) \cdot p(y_{m+1}|z_{m+1}) \cdot \rho_{s_m, s_{m-1}, \ldots, s_1}}{p(y_{m+1}|\mathscr{Y}_m)}, \tag{B.5}$$

where $\rho_{s_m, s_{m-1}, \ldots, s_1}$ is the exogenous parameterization of the initial start-up probabilities as in (2.7).

Let t denote the date for whose $m+1$ most recent values of s we wish to draw a smoothed inference. Let $s_t, s_{t-1}, \ldots, s_{t-m}$ denote a particular value out of the K^{m+1} possibilities, and think of studying this particular value in isolation. We expand the set of variables to include the values for m dates following t by iterating on the following expression for $\tau = t+1, t+2,$

$\ldots, t+m$:

$$p(s_\tau, s_{\tau-1}, \ldots, s_{t-m} | \mathscr{Y}_\tau)$$

$$= \frac{p(s_\tau | s_{\tau-1}) \cdot p(y_\tau | z_\tau) \cdot p(s_{\tau-1}, s_{\tau-2}, \ldots, s_{t-m} | \mathscr{Y}_{\tau-1})}{p(y_\tau | \mathscr{Y}_{\tau-1})}. \tag{B.6}$$

Eq. (B.6) denotes $K^{(\tau-t)}$ different numbers (indexed by $s_\tau, s_{\tau-1}, \ldots, s_{t+1}$) for the particular event $(s_t, s_{t-1}, \ldots, s_{t-m})$ being studied.

After reaching $\tau = t+m$, we carry forward an inference about (a) the $m+1$ most recent values of s as of date τ and (b) the $m+1$ most recent values of s as of date t:

$$p(s_\tau, s_{\tau-1}, \ldots, s_{\tau-m}, s_t, s_{t-1}, \ldots, s_{t-m} | \mathscr{Y}_\tau)$$

$$= \frac{\sum_{s_{\tau-m-1}=1}^{K} p(s_\tau | s_{\tau-1}) \cdot p(y_\tau | z_\tau) \cdot p(s_{\tau-1}, s_{\tau-2}, \ldots, s_{\tau-m-1}, s_t, s_{t-1}, \ldots, s_{t-m} | \mathscr{Y}_{\tau-1})}{p(y_\tau | \mathscr{Y}_{\tau-1})}.$$

The approach is to iterate on the above expression for $\tau = t+m+1$, $t+m+2, \ldots, T$. When we reach $\tau = T$, we can then calculate the smoothed inference as

$$p(s_t, s_{t-1}, \ldots, s_{t-m} | \mathscr{Y})$$

$$= \sum_{s_T=1}^{K} \sum_{s_{T-1}=1}^{K} \cdots \sum_{s_{T-m}=1}^{K} p(s_T, s_{T-1}, \ldots, s_{T-m}, s_t, s_{t-1}, \ldots, s_{t-m} | \mathscr{Y}_T).$$

The iterations leading from (B.6) to (B.7) are then repeated for the next possible value for $(s_t, s_{t-1}, \ldots, s_{t-m})$ until a smoothed probability of the form (B.1) has been calculated from all possible $(s_t, s_{t-1}, \ldots, s_{t-m})$ and all $t = m, m+1, \ldots, T$.

The total number of calculations required by the above algorithm is of order $K^{2(m+1)}T^2$; that is, it grows with the square of the sample size T and the $2(m+1)$th power of the number of states K. Total storage requirements are of order $2K^{m+1}T$.

References

Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss, 1970, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Annals of Mathematical Statistics 41, 164–171.

Cecchetti, Stephen G., Pok-sang Lam, and Nelson Mark, forthcoming, Mean reversion in equilibrium asset prices, American Economic Review.
Chiang, Chin Long, 1980, An introduction to stochastic processes and their applications (Krieger Publishing Co., New York, NY).
Cook, Timothy and Thomas Hahn, 1989, The effect of changes in the federal funds rate target on market interest rates in the 1970's, Journal of Monetary Economics 24, 331–351.
Cosslett, Stephen R. and Lung-Fei Lee, 1985, Serial correlation in discrete variable models, Journal of Econometrics 27, 79–97.
DeGroot, Morris H., 1970, Optimal statistical decisions (McGraw-Hill, New York, NY).
Dempster, A.P., N.M. Laird, and D.B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39, 1–38.
Engel, Charles and James D. Hamilton, forthcoming, Long swings in the exchange rate: Are they in the data and do markets know it?, American Economic Review.
Engle, Robert F., 1982, Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, Econometrica 50, 987–1007.
Everitt, B.S. and D.J. Hand, 1981, Finite mixture distributions (Chapman and Hall, London).
Fama, Eugene F. and Kenneth R. French, 1988, Permanent and temporary components of stock prices, Journal of Political Economy 96, 246–273.
Gallant, A. Ronald, 1987, Nonlinear statistical models (Wiley, New York, NY).
Goldfeld, Stephen M. and Richard M. Quandt, 1973, A Markov model for switching regressions, Journal of Econometrics 1, 3–16.
Hamilton, James D., 1988a, Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates, Journal of Economic Dynamics and Control 12, 385–423.
Hamilton, James D., 1988b, A pseudo-Bayesian approach to estimating parameters for mixtures of normal distributions, Mimeo. (University of Virginia, Charlottesville, VA).
Hamilton, James D., 1989, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica 57, 357–384.
Hassett, Kevin, 1988, Persistence and cyclicality in the aggregate labor market, Mimeo. (University of Pennsylvania, Philadelphia, PA).
Kiefer, Nicholas M., 1978, Discrete parameter variation: Efficient estimation of a switching regression model, Econometrica 46, 427–434.
Kiefer, Nicholas M., 1980, A note on switching regressions and logistic discrimination, Econometrica 48, 1065–1069.
Lam, Pok-sang, 1988, The generalized Hamilton model: Estimation and comparison with other models of economic time series, Mimeo. (Ohio State University, Columbus, OH).
Lee, Lung-Fei and Andrew Chesher, 1986, Specification testing when score statistics are identically zero, Journal of Econometrics 31, 121–149.
Liporace, Louis A., 1982, Maximum likelihood estimation for multivariate observations of Markov sources, IEEE Transactions on Information Theory IT-28, 729–734.
Lucas, Robert E., Jr., 1978, Asset prices in an exchange economy, Econometrica 66, 1429–1445.
Perron, Pierre, 1989, The Great Crash, the oil price shock and the unit root hypothesis, Econometrica 57, 1361–1401.
Poterba, James M. and Lawrence H. Summers, 1988, Mean reversion in stock prices: Evidence and implications, Journal of Financial Economics 22, 27–59.
Quandt, Richard E., 1958, The estimation of parameters of linear regression system obeying two separate regimes, Journal of the American Statistical Association 55, 873–880.
Ruud, Paul A., 1988, Extension of estimation methods using the EM algorithm, Mimeo. (University of California, Berkeley, CA).
Theil, Henri, 1971, Principles of econometrics (Wiley, New York, NY).
Watson, Mark W. and Robert F. Engle, 1983, Alternative algorithms for the estimation of dynamic factor, mimic, and varying coefficient regression models, Journal of Econometrics 23, 385–400.
Watson, Mark W. and Robert F. Engle, 1985, Testing for regression coefficient stability with a stationary AR(1) alternative, Review of Economics and Statistics 67, 341–346.