ELSEVIER

Journal of Econometrics 70 (1996) 127–157

JOURNAL OF Econometrics

Specification testing in Markov-switching time-series models

James D. Hamilton

Department of Economics, University of California, San Diego, La Jolla, CA 92093-0508, USA

Abstract

This paper develops a series of specification tests of Markov-switching time-series models. Tests for omitted autocorrelation, omitted ARCH, misspecification of the Markovian dynamics, and omitted explanatory variables are proposed. All of the tests can be constructed as a natural byproduct of the routine used to calculate the 'smoothed' probability that a given observation came from a particular regime, and do not require estimation of additional parameters. The paper performs Monte Carlo analysis of the tests and briefly illustrates their use with an empirical application.

Key words: Regime-switching models; Markov-switching models; Specification tests
JEL classification: C22

1. Introduction

A number of researchers have recently become interested in modeling economic and financial time series as subject to occasional, discrete shifts in parameters. When the Markov-switching regression framework of Goldfeld and Quandt (1973) and Cosslett and Lee (1985) is applied to a time-series autoregression, the result is a model that allows for nonlinear dynamics and sudden changes in the variability of a series, yet still is very tractable for rational-expectations econometrics. This approach has provided some new perspectives

This material is based upon work supported by the National Science Foundation under Grant SES-8920752. I am grateful to Neil Ericsson, René Garcia, Eric Ghysels, and anonymous referees for helpful comments on an earlier draft of this paper. Data and software used in this paper can be obtained at no charge from the author. Alternatively, data and software can be obtained by writing to ICPSR, Institute for Social Research, P.O. Box 1248, Ann Arbor, MI 48106.

SSDI 0304407694168 6 T

on the nature of business cycles and the long-run trend in GNP (Hamilton, 1989; Lam, 1990), the effect of Federal Reserve actions on interest rates (Hamilton, 1988; Garcia and Perron, 1989), the 'mean reversion' property of stock returns (Cecchetti, Lam, and Mark, 1990b), financial panics (Schwert, 1989, 1990), risk aversion in financial markets (Cecchetti, Lam, and Mark, 1990a; Turner, Startz, and Nelson, 1989), and the behavior of foreign exchange rates (Engel and Hamilton, 1990).

Despite the interest in this approach, to date there has been no clearly articulated guide for verifying the adequacy of the models' fit to the data. This paper attempts to fill this gap, applying the Lagrange multiplier principle popularized by Breusch and Pagan (1980), Godfrey and Wickens (1981), and Engle (1982a, 1984) and the general approach to specification testing developed by Newey (1985), Tauchen (1985), and White (1987).

The paper provides an expression for the score in these models, defined as the derivative of the conditional log-likelihood of the tth observation with respect to the parameter vector. This permits easy calculation of all of the necessary test statistics as well as an intuitive interpretation of what each test is based on. The score turns out to be a natural byproduct of the routine used to calculate the 'smoothed' probability that a given observation came from a particular regime. From the same calculations that produce these smoothed probabilities, this paper shows how to construct asymptotic standard errors for the parameter vector and specification tests for a variety of forms of autocorrelation, generalized ARCH effects, higher-order Markovian dynamics, violation of the presumed independence of the Markov process, and omitted explanatory variables for both the mean and variance. Once one has maximized the likelihood function of a specification lacking all these considerations, these tests can all be calculated together from a single pass through the data.

The general class of models considered in this paper is as follows. The 'state' or 'regime' that a process is in at date t is indexed by an unobserved random variable s_t – regime 1 is represented by $s_t = 1$, regime 2 by $s_t = 2$, and so on. If there are K different possible regimes, we assume that transitions between regimes are governed by a K-state Markov chain:

$$p(s_t = j | s_{t-1} = i) = p_{ij}, \qquad i, j = 1, \ldots, K. \tag{1.1}$$

Within a given regime, the dynamics of the observed data are determined by a conventional dynamic model. To take the simplest example, suppose that conditional on a particular value of s_t the observed datum y_t is Gaussian,

$$y_t | s_t \sim \mathrm{N}(\mu_{s_t}, \sigma^2_{s_t}). \tag{1.2}$$

The results in Section 3 imply that for this simple case the derivative of the log-likelihood of the observed sample of data through date t with respect to the parameter that characterizes the mean in regime 1 is given by

$$\frac{\partial \log p(y_t, y_{t-1}, \ldots, y_1)}{\partial \mu_1} = \sum_{\tau=1}^{t} \frac{(y_\tau - \mu_1)}{\sigma_1^2} \cdot p(s_\tau = 1 | y_t, y_{t-1}, \ldots, y_1), \tag{1.3}$$

where $p(s_\tau = 1 | y_t, y_{t-1}, \ldots, y_1)$ denotes the smoothed probability that the observation for date τ was drawn from regime 1, with this probabilistic inference based on the observations on y through date t. It is straightforward to deduce from (1.3) that the score of observation t is given by

$$\begin{aligned}\frac{\partial \log p(y_t | y_{t-1}, y_{t-2}, \ldots, y_1)}{\partial \mu_1} &= \frac{(y_t - \mu_1)}{\sigma_1^2} \cdot p(s_t = 1 | y_t, y_{t-1}, \ldots, y_1) \\ &\quad + \sum_{\tau=1}^{t-1} \frac{(y_\tau - \mu_1)}{\sigma_1^2} \cdot [p(s_\tau = 1 | y_t, y_{t-1}, \ldots, y_1) \\ &\quad - p(s_\tau = 1 | y_{t-1}, y_{t-2}, \ldots, y_1)], \end{aligned} \tag{1.4}$$

which is a simple function of how the inference about date τ's state changes with the addition of observation t of y. This change in turn is naturally calculated as a byproduct of the routine for generating the inference about date τ's state using the full sample of observations on y,

$$p(s_\tau = 1 | y_T, y_{T-1}, \ldots, y_1),$$

which one would often want to calculate in these applications even if there were no interest in specification diagnostics.

Once we have the score from expressions such as (1.4), we can calculate an estimate of the information matrix as in Berndt, Hall, Hall, and Hausman (1974), permitting construction of asymptotic standard errors for the maximum-likelihood estimates of parameters. Lagrange multiplier tests are similarly easy to construct. For example, one can test for the presence of autocorrelation within regime 1 on the basis of the size of the Lagrange multiplier. This takes the form of the derivative of the log-likelihood with respect to the autoregressive coefficient (denoted ϕ_1) applicable when the process is in regime 1

$$\begin{aligned}\frac{\partial \log p(y_T, y_{T-1}, \ldots, y_1)}{\partial \phi_1} &= \sum_{\tau=1}^{T} \frac{(y_\tau - \mu_1)(y_{\tau-1} - \mu_1)}{\sigma_1^2} \\ &\quad \cdot p(s_\tau = 1, s_{\tau-1} = 1 | y_T, y_{T-1}, \ldots, y_1).\end{aligned}$$

The score in this case is

$$\frac{\partial \log p(y_t|y_{t-1}, y_{t-2}, \cdots, y_1)}{\partial \phi_1} = \frac{(y_t - \mu_1)(y_{t-1} - \mu_1)}{\sigma_1^2} \cdot p(s_t = 1, s_{t-1} = 1|y_t, y_{t-1}, \cdots, y_1) + \sum_{\tau=1}^{t-1} \frac{(y_\tau - \mu_1)(y_{\tau-1} - \mu_1)}{\sigma_1^2} \cdot [p(s_\tau = 1, s_{\tau-1} = 1|y_t, y_{t-1}, \cdots, y_1) - p(s_\tau = 1, s_{\tau-1} = 1|y_{t-1}, y_{t-2}, \cdots, y_1)], \tag{1.5}$$

with (1.5) again easy to calculate from the routine for updating the smoothed probabilities.

The tests proposed here employ an estimate of the information matrix based on the average outer product of the gradient. This form has the advantage that it is very simple to compute, but has the disadvantage that it may have substantially poorer finite-sample properties than an estimate based on the analytically evaluated expectation of the matrix of second derivatives; see Davidson and MacKinnon (1983) and Godfrey, McAleer, and McKenzie (1988). Orme (1988, 1990) has found convenient ways to calculate better estimates of the information matrix for models related to those investigated here.

The paper is organized as follows. Section 2 provides a summary and brief motivation for the general principles of specification testing employed in this paper. Section 3 presents the analogs to (1.4) and (1.5) for a general Markov-switching time-series model with a variety of possible departures from the null specification, and discusses the intuitive basis for the tests. Section 4 reports Monte Carlo evidence on the small-sample performance of the various test statistics proposed. These tests are applied to an analysis of exchange rate data in Section 5. Section 6 briefly summarizes.

2. Description of specification tests employed

Suppose we want to estimate an $(m \times 1)$ vector of parameters λ based on a time series of T observations on a scalar y_t. Consider the distribution of y_t conditional on values of y for dates $t-1, t-2, \ldots, 0, -1, \ldots, -r+1$ and on realizations of a vector of observable exogenous variables $\boldsymbol{w}_t$,

$$p(y_t|\mathcal{X}_t; \lambda), \tag{2.1}$$

where

$$\mathcal{X}_t' \equiv (\boldsymbol{w}_t', \boldsymbol{w}_{t-1}', \ldots, \boldsymbol{w}_1', y_{t-1}, y_{t-2}, \ldots, y_0, y_{-1}, \ldots, y_{-r+1}). \tag{2.2}$$

The task is to choose $\boldsymbol{\lambda}$ so as to maximize

$$\sum_{t=1}^{T} \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda}). \tag{2.3}$$

The score of the tth observation is defined as the derivative of the log of the conditional likelihood (2.1) with respect to the parameter vector $\boldsymbol{\lambda}$. We will let the $(m \times 1)$ vector-valued function $\boldsymbol{h}_t(\boldsymbol{\lambda})$ denote this derivative, and will focus on the value of this derivative at the true parameter value $(\boldsymbol{\lambda}_0)$,

$$\boldsymbol{h}_t(\boldsymbol{\lambda}_0) \equiv \frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0}, \tag{2.4}$$

and at the MLE $(\hat{\boldsymbol{\lambda}})$:

$$\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}}) \equiv \frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}. \tag{2.5}$$

There are T such $(m \times 1)$ score vectors $\boldsymbol{h}_t(\boldsymbol{\lambda}_0)$ associated with the sample $(y_T, y_{T-1}, \ldots, y_1)$.

Notice that since (2.1) is a density, it integrates to unity:

$$\int p(y_t|\mathcal{X}_t; \boldsymbol{\lambda}) \mathrm{d}y_t = 1. \tag{2.6}$$

Differentiating (2.6) (and assuming that regularity conditions discussed below permit exchanging the order of operations), we see $\int [\partial p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})/\partial \boldsymbol{\lambda}] \mathrm{d}y_t = \mathbf{0}$, or

$$\int \frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \cdot p(y_t|\mathcal{X}_t; \boldsymbol{\lambda}) \mathrm{d}y_t = \mathbf{0}. \tag{2.7}$$

If the data were really generated from the distribution $p(y_t|\mathcal{X}_t; \boldsymbol{\lambda}_0)$, then (2.7) states that

$$\mathrm{E}[\boldsymbol{h}_t(\boldsymbol{\lambda}_0)|\mathcal{X}_t] = \mathbf{0}. \tag{2.8}$$

Thus if the model is correctly specified, the score $\boldsymbol{h}_t(\boldsymbol{\lambda}_0)$ should be impossible to forecast on the basis of any information available at date $t-1$, such as elements of the lagged score $\boldsymbol{h}_{t-1}(\boldsymbol{\lambda}_0)$.

White (1987) therefore proposed tests for serial correlation of the scores using the conditional moment tests of Newey (1985) and Tauchen (1985). To conduct this test we collect in an $(l \times 1)$ vector $\boldsymbol{c}_t(\boldsymbol{\lambda})$ those elements of the $(m \times m)$ matrix $[\boldsymbol{h}_t(\boldsymbol{\lambda})] \cdot [\boldsymbol{h}_{t-1}(\boldsymbol{\lambda})]'$ that we want to confirm have a zero mean when evaluated at $\boldsymbol{\lambda}_0$. If the model is correctly specified, then

$$\left[T^{-1/2} \sum_{t=1}^{T} \boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})\right]' \cdot \hat{A}^{22} \cdot \left[T^{-1/2} \sum_{t=1}^{T} \boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})\right] \xrightarrow{\mathrm{d}} \chi^2(l), \tag{2.9}$$

where $\hat{A}^{22}$ denotes the (2, 2) subblock of the inverse of the following partitioned matrix:

$$\hat{A} = (1/T)\cdot\begin{bmatrix} \sum_{t=1}^{T}[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})]\cdot[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})]' & \sum_{t=1}^{T}[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})]\cdot[\boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})]' \\ \sum_{t=1}^{T}[\boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})]\cdot[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})]' & \sum_{t=1}^{T}[\boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})]\cdot[\boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})]' \end{bmatrix}. \tag{2.10}$$

A further use made of the scores is evaluation of Lagrange multiplier tests (see Rao, 1948; Aitchison and Silvey, 1958; Breusch and Pagan, 1980; Godfrey and Wickens, 1981; Engle, 1982a, 1984). Suppose that the $(m \times 1)$ parameter vector $\boldsymbol{\lambda}$ is estimated subject to the constraint that the last m_0 elements are zero. Then at the constrained MLE $\tilde{\boldsymbol{\lambda}}$, the first $(m - m_0)$ elements of the average score $(1/T)\cdot\sum_{t=1}^{T}\boldsymbol{h}_t(\tilde{\boldsymbol{\lambda}})$ are zero (by virtue of the first-order conditions for constrained maximization of the likelihood function), whereas the last m_0 are nonzero. The magnitude of these last m_0 elements reflects how much the likelihood function might increase if the constraints were relaxed, and can be used to assess the validity of the constraints. Asymptotically,

$$\left[T^{-1/2}\sum_{t=1}^{T}\boldsymbol{h}_t(\tilde{\boldsymbol{\lambda}})\right]'\cdot\left[(1/T)\cdot\sum_{t=1}^{T}[\boldsymbol{h}_t(\tilde{\boldsymbol{\lambda}})]\cdot[\boldsymbol{h}_t(\tilde{\boldsymbol{\lambda}})]'\right]^{-1}\cdot\left[T^{-1/2}\sum_{t=1}^{T}\boldsymbol{h}_t(\tilde{\boldsymbol{\lambda}})\right]$$

$$\xrightarrow[\text{d}]{} \chi^2(m_0) \tag{2.11}$$

As in Godfrey and Wickens (1981, p. 490), the statistic in (2.11) can conveniently be calculated as T times the uncentered R^2 from a regression of the constant unity on $\boldsymbol{h}_t(\tilde{\boldsymbol{\lambda}})$. Similarly, (2.9) can be calculated as T times the uncentered R^2 from a regression of 1 on $([\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})]', [\boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})]')'$.

A final use of the scores is to construct standard errors for the MLE $\hat{\boldsymbol{\lambda}}$ using an estimate of the information matrix from the scores' average outer product, as in Berndt, Hall, Hall, and Hausman (1974). Let

$$\hat{\mathcal{I}}_{\text{OP}} = \frac{1}{T}\sum_{t=1}^{T}[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})]\cdot[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})]'. \tag{2.12}$$

We might then treat

$$(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) \approx \text{N}(\boldsymbol{0}, (1/T)\cdot\hat{\mathcal{I}}_{\text{OP}}^{-1}). \tag{2.13}$$

It would also be possible to calculate 'robust' standard errors as in White (1982),

$$(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) \approx \text{N}(\boldsymbol{0}, (1/T)\cdot\{\hat{\mathcal{I}}_{\text{2D}}\hat{\mathcal{I}}_{\text{OP}}^{-1}\hat{\mathcal{I}}_{\text{2D}}\}^{-1}),$$

where

$$\hat{\mathcal{I}}_{\text{2D}} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial\boldsymbol{h}_t(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}'}.$$

Sufficient conditions for interchanging the order of differentiation and integration and deriving the asymptotic distributions above are discussed in White (1987) and Gallant and White (1988). Verifying that related conditions hold here and cataloging all the possible cases is beyond the scope of this paper. We note here a few key exceptions to these conditions likely to arise in practice. The derivations assume that both $\hat{\lambda}$ and λ_0 are interior; the formulas below will not be valid at corner conditions such as $\hat{p}_{ii} = 0$ or 1. For the model

$$y_t = \phi_{1,s_t} y_{t-1} + \phi_{2,s_t} y_{t-2} + \cdots + \phi_{m,s_t} y_{t-m} + \varepsilon_t,$$

the mixing condition is satisfied if roots of $1 - \phi_{1,j} z - \cdots - \phi_{m,j} z^m = 0$ lie outside the unit circle for all j; for

$$y_t = \mu_{s_t} + \varepsilon_t,$$

the mixing condition is satisfied if the Markov chain for s_t is ergodic. In the $K = 2$ case, the Markov chain is ergodic whenever p_{11} and p_{22} are both interior points in (0, 1). Restrictions or modifications on maximum-likelihood estimation are required when the variance of innovations is made a function of the state – see the discussion in Hamilton (1991).

3. Form of the specification tests for Markov-switching models

3.1. The score for the basic switching model

Eq. (2.1) characterized the likelihood function for the tth observation conditional on an observed vector $\mathcal{X}_t$, where $\mathcal{X}_t$ included a growing number of lags on y and an exogenous vector $\boldsymbol{w}$. This representation summarized the observable implications of a model that is now instead characterized in terms of an unobserved regime. Let s_t denote an unobserved random variable that takes an integer value in $(1, \ldots, K)$, corresponding to K separate possible states or regimes. The switching model describes the distribution of y_t conditional on s_t, a finite number of its own lags $(y_{t-1}, y_{t-2}, \ldots, y_{t-r})$ and the current value of the observable vector of exogenous variables $\boldsymbol{w}_t$,

$$p(y_t|\boldsymbol{x}_t, s_t; \boldsymbol{\theta}), \tag{3.1}$$

where

$$\boldsymbol{x}_t' \equiv (\boldsymbol{w}_t', y_{t-1}, y_{t-2}, \ldots, y_{t-r}). \tag{3.2}$$

As an example of (3.1) we could postulate that parameters of a regression relation between y and $\boldsymbol{x}$ shift with the realization of s_t,

$$p(y_t|\boldsymbol{x}_t, s_t; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{s_t}} \cdot \exp\left[\frac{-(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_{s_t})^2}{2\sigma_{s_t}^2}\right], \tag{3.3}$$

in which case θ consists of the K regression coefficient vectors and variances:

$$\theta' = (\beta_1', \ldots, \beta_K', \sigma_1^2, \ldots, \sigma_K^2).$$

To get from the specification (3.1) involving unobservables to the observed likelihood (2.1), it is necessary to specify the probability law that governs changes in regime. It is assumed that the state s_t follows a Markov chain that is independent of w and lagged y. Denote the transition probabilities for this process by (p_{ij}):

$$\begin{aligned} \text{prob}[s_t = j | s_{t-1} = i, w_{t+l}, y_{t-m}] &= \text{prob}[s_t = j | s_{t-1} = i] \\ &= p_{ij}, \quad i, j = 1, \ldots, K, \\ &\qquad l = \ldots, -1, 0, 1, \ldots, \\ &\qquad m = 1, 2, \ldots . \end{aligned} \tag{3.4}$$

From (3.1) and (3.4) it is straightforward to calculate the observed likelihood (2.1) using algorithms developed by Cosslett and Lee (1985) and Hamilton (1989). The observed likelihood (2.1) is parameterized by λ, which includes both the parameters θ appearing in (3.1) and the parameters p_{ij} in (3.4),

$$\lambda' = (\theta', p'), \tag{3.5}$$

where

$$p' = (p_{11}, p_{12}, \ldots, p_{1,K-1}, p_{21}, p_{22}, \ldots, p_{K,K-1}).$$

Note we omit from the definition of p the redundant parameters

$$p_{iK} = 1 - p_{i1} - p_{i2} - \cdots - p_{i,K-1}, \qquad i = 1, \ldots, K. \tag{3.6}$$

The score $h_t(\lambda)$ discussed in Section 2 refers to the derivative of the observed likelihood (2.1). It turns out that the elements of the score corresponding to θ have a simple relation to derivatives of the unobserved conditional density (3.1). This relation is captured by the probabilistic inference the econometrician makes about the value of s for any given date. Let Ω_t denote the information available to the econometrician at date t,

$$\Omega_t \equiv (y_t, y_{t-1}, \ldots, y_0, y_{-1}, \ldots, y_{-r+1}, w_t, w_{t-1}, \ldots, w_1),$$

and let $p(s_\tau | \Omega_t)$ denote the econometrician's assessed probability that observation τ came from regime s_τ, where this assessment is based on information

observed through date t. Appendix A shows that for data generated by the model (3.1) and (3.4), the score with respect to θ is given by

$$\frac{\partial \log p(y_t|\mathcal{X}_t;\lambda)}{\partial \theta} = \sum_{j=1}^{K} \psi_{t,j}\, p(s_t = j|\Omega_t) + \sum_{\tau=1}^{t-1} \sum_{j=1}^{K} \psi_{\tau,j}[p(s_\tau = j|\Omega_t) - p(s_\tau = j|\Omega_{t-1})], \quad t = 1, 2, \ldots, T, \tag{3.7}$$

where

$$\psi_{t,j} = \frac{\partial \log p(y_t|x_t, s_t = j;\theta)}{\partial \theta}.$$

To evaluate (3.7), one simply needs to note how the econometrician's inference about the state the process was in at date τ changes with the addition of date t's data:

$$p(s_\tau|\Omega_t) - p(s_\tau|\Omega_{t-1}). \tag{3.8}$$

This number is then used to weight the derivatives of the log of (3.1):

$$\frac{\partial \log p(y_\tau|x_\tau, s_\tau;\theta)}{\partial \theta}.$$

Calculation of (3.8) is a natural byproduct of a routine for calculating the full-sample inferences

$$p(s_\tau|\Omega_T),$$

which one would typically want to evaluate even if no specification testing were desired. Appendix B describes how to calculate the terms in (3.8) by means of a simple recursion.

If the data were really generated by (3.1) and (3.4), then the scores given by (3.7) have expectation zero and form a martingale difference sequence. To see this, take expectations of (3.7):

$$\mathrm{E}\left[\frac{\partial \log p(y_t|\mathcal{X}_t;\lambda)}{\partial \theta}\,\middle|\,\Omega_{t-1}\right]$$

$$= \int \sum_{j=1}^{K} \psi_{t,j}\, p(s_t = j|\Omega_t)\cdot p(\Omega_t|\Omega_{t-1})\,\mathrm{d}\Omega_t$$

$$= \sum_{j=1}^{K} \left[\int \psi_{t,j}\, p(\Omega_t|s_t = j,\ \Omega_{t-1})\,\mathrm{d}\Omega_t\right]\cdot p(s_t = j|\Omega_{t-1}).$$

This is zero whenever the term in brackets is zero, or whenever

$$\mathrm{E}[\boldsymbol{\psi}_{t,j}|s_t = j, \Omega_{t-1}] = \mathbf{0},$$

which is simple to verify directly.

For the switching regression (3.3), the score with respect to $\boldsymbol{\beta}_1$ is given by (3.7) with

$$\boldsymbol{\psi}_{t,j} = \frac{\partial \log p(y_t|\boldsymbol{x}_t, s_t = j; \boldsymbol{\theta})}{\partial \boldsymbol{\beta}_1}$$

or

$$\begin{aligned} \boldsymbol{\psi}_{t,j} &= \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_1)\cdot \boldsymbol{x}_t}{\sigma_1^2} \quad \text{if} \quad j = 1 \\ &= \mathbf{0} \qquad\qquad\qquad\quad \text{otherwise.} \end{aligned} \tag{3.9}$$

The score with respect to $\boldsymbol{\beta}_2$ is

$$\begin{aligned} \frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\beta}_2} = \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_2)\cdot \boldsymbol{x}_t}{\sigma_2^2} \cdot p(s_t = 2|\Omega_t) \\ + \sum_{\tau=1}^{t-1} \frac{(y_\tau - \boldsymbol{x}_\tau'\boldsymbol{\beta}_2)\cdot \boldsymbol{x}_\tau}{\sigma_2^2} \cdot [p(s_\tau = 2|\Omega_t) - p(s_\tau = 2|\Omega_{t-1})]. \end{aligned} \tag{3.10}$$

The score with respect to the variance parameter for regime 1 is similarly found from (3.7) by setting

$$\begin{aligned} \boldsymbol{\psi}_{t,j} &= \frac{\partial \log p(y_t|\boldsymbol{x}_t, s_t = j; \boldsymbol{\theta})}{\partial \sigma_1^2} \\ &= \frac{-1}{2\sigma_1^2} + \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_1)^2}{2\sigma_1^4} \quad \text{if} \quad j = 1 \\ &= 0 \qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{3.11}$$

The score with respect to the transition probabilities takes a similar form to (3.7):

$$\begin{aligned} &\frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})}{\partial p_{ij}} \\ &= p_{ij}^{-1} \cdot p(s_t = j, s_{t-1} = i|\Omega_t) - p_{iK}^{-1} \cdot p(s_t = K, s_{t-1} = i|\Omega_t) \\ &\quad + p_{ij}^{-1} \cdot \left\{ \sum_{\tau=2}^{t-1} [p(s_\tau = j, s_{\tau-1} = i|\Omega_t) - p(s_\tau = j, s_{\tau-1} = i|\Omega_{t-1})] \right\} \end{aligned}$$

$$- p_{iK}^{-1} \cdot \left\{ \sum_{\tau=2}^{t-1} [p(s_\tau = K, s_{\tau-1} = i|\Omega_t) - p(s_\tau = K, s_{\tau-1} = i|\Omega_{t-1})] \right\}$$

$$+ \sum_{s_1=1}^{K} \frac{\partial \log p(s_1; p)}{\partial p_{ij}} \cdot [p(s_1|\Omega_t) - p(s_1|\Omega_{t-1})], \tag{3.12}$$

$$i = 1, \ldots, K, \quad j = 1, \ldots, K-1, \quad t = 2, \ldots, T.$$

$$\frac{\partial \log p(y_1|\mathcal{X}_1; \lambda)}{\partial p_{ij}} = \sum_{s_1=1}^{K} \frac{\partial \log p(s_1; p)}{\partial p_{ij}} \cdot p(s_1|\Omega_1). \tag{3.13}$$

One approach is to assume that the initial state s_1 represents a draw from the ergodic distribution of the Markov chain.[1] For example, for $K = 2$,

$$p(s_1 = 1) = \frac{(1 - p_{22})}{(1 - p_{11}) + (1 - p_{22})}, \tag{3.14a}$$

$$p(s_1 = 2) = \frac{(1 - p_{11})}{(1 - p_{11}) + (1 - p_{22})}. \tag{3.14b}$$

For the case $K = 2$, it is simplest to exploit the symmetry of the problem and parameterize the two nonredundant transition probabilities as p_{11} and p_{22} (rather than p_{11} and p_{21}, which would be the convention in all of the general K-state formulas elsewhere in this paper). For this parameterization, differentiating (3.14) and substituting into (3.12) and (3.13) yields

$$\frac{\partial \log p(y_t|\mathcal{X}_t; \lambda)}{\partial p_{11}}$$

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

$$+ p_{11}^{-1} \cdot \left\{ \sum_{\tau=2}^{t-1} [p(s_\tau = 1, s_{\tau-1} = 1|\Omega_t) - p(s_\tau = 1, s_{\tau-1} = 1|\Omega_{t-1})] \right\}$$

$$- (1 - p_{11})^{-1} \cdot \left\{ \sum_{\tau=2}^{t-1} [p(s_\tau = 2, s_{\tau-1} = 1|\Omega_t) - p(s_\tau = 2, s_{\tau-1} = 1|\Omega_{t-1})] \right\}$$

$$+ \frac{p(s_1 = 1|\Omega_t) - p(s_1 = 1|\Omega_{t-1})}{1 - p_{11}}, \quad t = 2, \ldots, T, \tag{3.15a}$$

$$= \frac{p(s_1 = 1|\Omega_1) - [(1 - p_{22})/(1 - p_{11} + 1 - p_{22})]}{1 - p_{11}}, \quad t = 1, \tag{3.16a}$$

[1] See Hamilton (1990) for a discussion of this and other approaches.

and

$$\frac{\partial \log p(y_t|\mathcal{X}_t;\boldsymbol{\lambda})}{\partial p_{22}}$$

$$= p_{22}^{-1}\cdot p(s_t=2, s_{t-1}=2|\Omega_t) - (1-p_{22})^{-1}\cdot p(s_t=1, s_{t-1}=2|\Omega_t)$$

$$+ p_{22}^{-1}\cdot\left\{\sum_{\tau=2}^{t-1}[p(s_\tau=2, s_{\tau-1}=2|\Omega_t) - p(s_\tau=2, s_{\tau-1}=2|\Omega_{t-1})]\right\}$$

$$-(1-p_{22})^{-1}\cdot\left\{\sum_{\tau=2}^{t-1}[p(s_\tau=1, s_{\tau-1}=2|\Omega_t) - p(s_\tau=1, s_{\tau-1}=2|\Omega_{t-1})]\right\}$$

$$+\frac{p(s_1=2|\Omega_t) - p(s_1=2|\Omega_{t-1})}{1-p_{22}}, \quad t=2,\ldots,T, \tag{3.15b}$$

$$=\frac{p(s_1=2|\Omega_1) - [(1-p_{11})/(1-p_{11}+1-p_{22})]}{1-p_{22}}, \quad t=1. \tag{3.16b}$$

To summarize, for the $K=2$ case, the parameter vector is given by $\boldsymbol{\lambda}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \sigma_1^2, \sigma_2^2, p_{11}, p_{22})$. The score $\boldsymbol{h}_t(\boldsymbol{\lambda})$ is obtained by vertically stacking the results from (3.9), (3.10), the terms coming from (3.11) for σ_1^2 and σ_2^2, and (3.15a) and (3.15b). The average outer product of these scores can be used to construct an estimate of the information matrix as in (2.12) and used to generate standard errors for $\hat{\boldsymbol{\lambda}}$ from (2.13). All of the elements of the score take the form of very simple functions of the data weighted by the change in the smoothed probabilities.

3.2. *Implementing the Newey–Tauchen–White test for dynamic misspecification*

Having shown how to evaluate the score $\boldsymbol{h}_t(\boldsymbol{\lambda})$, it is simple to calculate the $(l\times 1)$ vector $\boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})$ consisting of the products of selected elements of the score for observation t with elements for observation $t-1$. This subsection comments on what is being examined by this test in the Markov-switching instance and on which elements of $[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})]\cdot[\boldsymbol{h}_{t-1}(\hat{\boldsymbol{\lambda}})]'$ are of particular interest.

One can most clearly see the intuition behind these tests in the case where the econometrician has no difficulty identifying which regime a given observation was drawn from. In this case, $p(s_\tau = j|\Omega_t)$ is unity when $s_\tau = j$ and is zero otherwise. Considering the score with respect to the constant term (say, the first element of $\boldsymbol{x}_t$), (3.9) would in this case imply

$$\frac{\partial \log p(y_t|\mathcal{X}_t;\boldsymbol{\lambda})}{\partial \beta_1(1)} = \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_1)}{\sigma_1^2} \quad \text{if} \quad s_t = 1,$$
$$= 0 \qquad\qquad\qquad\quad \text{otherwise,} \tag{3.17}$$

where the scalar $\beta_j(1)$ denotes the first element of the vector $\boldsymbol{\beta}_j$ (associated with the constant term). Expression (3.17) characterizes this component of the score as the scaled OLS residual whenever the process is in regime 1. The model holds these to be uncorrelated, and choosing $\boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})$ to be the (1, 1) element of $[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})] \cdot [\boldsymbol{h}_{t-1}(\hat{\boldsymbol{\lambda}})]'$ would in this case just be a test for first-order autocorrelation of the residuals from those observations from regime 1.

When the econometrician does not know for certain which regimes governed which observations, the first element of the score $\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})$ is the residual (3.17) weighted by the current probability that observation t came from regime 1, plus the innovation in the assessment of the categorization of regime 1 residuals $(y_\tau - \boldsymbol{x}_\tau'\boldsymbol{\beta}_1)/\sigma_1^2$ for $\tau < t$. Again, if the model is correctly specified these innovations should be impossible to forecast.

The element of $[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})] \cdot [\boldsymbol{h}_{t-1}(\hat{\boldsymbol{\lambda}})]'$ corresponding to

$$\frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})}{\partial \beta_1(1)} \cdot \frac{\partial \log p(y_{t-1}|\mathcal{X}_{t-1}; \boldsymbol{\lambda})}{\partial \beta_2(1)}$$

reflects whether the residual from regime 2 is useful in forecasting the residual from regime 1 in those periods t when $s_{t-1} = 2$ may have been followed by $s_t = 1$. This along with its mirror image

$$\frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})}{\partial \beta_2(1)} \cdot \frac{\partial \log p(y_{t-1}|\mathcal{X}_{t-1}; \boldsymbol{\lambda})}{\partial \beta_1(1)}$$

are both interesting candidates for inclusion in $\boldsymbol{c}_t(\boldsymbol{\lambda})$.

Looking for autocorrelation in the elements of $\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})$ corresponding to derivatives with respect to σ_i^2 (as in 3.11) is a check for generalized ARCH effects. For example, the element of $[\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})] \cdot [\boldsymbol{h}_{t-1}(\hat{\boldsymbol{\lambda}})]'$ corresponding to

$$\frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})}{\partial \sigma_1^2} \cdot \frac{\partial \log p(y_{t-1}|\mathcal{X}_{t-1}; \boldsymbol{\lambda})}{\partial \sigma_2^2}$$

is basically checking whether, in those instances when regime 2 was followed by regime 1, the squared residuals from regime 1,

$$\left[\frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_1)^2}{2\sigma_1^4} - \frac{1}{2\sigma_1^2}\right],$$

are correlated with the lagged squared residuals from regime 2.

Finally, looking for predictability of the elements of $\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})$ corresponding to p_{ij} [as in (3.15)] is a check for violation of the Markov specification for the unobserved states. Consider first a two-state i.i.d. switching regression ($K = 2$ with $p_{11} = p_{21} = p$). In this case Eq. (3.15a) simplifies to

$$\frac{\partial \log p(y_t|\boldsymbol{x}_t)}{\partial p} = p^{-1} \cdot p(s_t = 1|y_t, \boldsymbol{x}_t) - (1 - p)^{-1} \cdot p(s_t = 2|y_t, \boldsymbol{x}_t).$$

The score is positive when $p(s_t = 1|y_t, x_t) > p$, that is, when the data persuade us that observation t was more likely to have come from regime 1 than we would have anticipated with no prior information. If one has no prior information about the value s_t will take on, the score has expectation zero conditional on any information available at date $t-1$. Positive serial correlation in the scores in this i.i.d. switching regression case would mean that having useful information that y_{t-1} came from regime 1 helps us to predict that y_t will be more likely than usual to come from regime 1 as well, and thus would constitute evidence against the i.i.d. switching assumption.

In the more general Markov-switching specification ($p_{11} \neq p_{21}$), the leading terms in the score in (3.15) are

$$p_{11}^{-1} \cdot p(s_t = 1, s_{t-1} = 1|\Omega_t) - (1 - p_{11})^{-1} \cdot p(s_t = 2, s_{t-1} = 1|\Omega_t).$$

The forecast of the score based on information available at $t-1$ can be written

$$p_{11}^{-1} \cdot p(s_t = 1|s_{t-1} = 1, \Omega_{t-1}) \cdot p(s_{t-1} = 1|\Omega_{t-1})$$
$$- (1 - p_{11})^{-1} \cdot p(s_t = 2|s_{t-1} = 1, \Omega_{t-1}) \cdot p(s_{t-1} = 1|\Omega_{t-1}).$$

Again, if the model is correctly specified, $p(s_t = 1|s_{t-1} = 1, \Omega_{t-1}) = p_{11}$ and the score should have expectation zero. The score is biggest for those observations for which we are most sure that state 1 followed itself. Positive serial correlation in the scores means that if we learn that s_{t-1} and s_{t-2} were both 1, the probability that $s_t = 1$ is greater than p_{11}. Thus serial correlation of the scores indicates that $p(s_t = 1|s_{t-1} = 1)$ is different from $p(s_t = 1|s_{t-1} = 1, s_{t-2} = 1)$, a violation of the first-order Markov assumption. A specification test based on

$$\frac{\partial \log p(y_t|\mathcal{X}_t; \lambda)}{\partial p_{11}} \cdot \frac{\partial \log p(y_{t-1}|\mathcal{X}_{t-1}; \lambda)}{\partial p_{11}}$$

should therefore be quite useful.

By contrast, looking for cross-correlations of the form

$$\frac{\partial \log p(y_t|\mathcal{X}_t; \lambda)}{\partial p_{11}} \cdot \frac{\partial \log p(y_{t-1}|\mathcal{X}_{t-1}; \lambda)}{\partial p_{22}} \tag{3.18}$$

may be unreliable in all but the largest samples. Suppose for example that the sample separation were sufficiently good that one essentially knows the value of s_t with certainty once we observe y_t. In this case (3.18) would be given by

$$\begin{cases} -p_{11}^{-1} \cdot (1 - p_{22})^{-1} & \text{if } s_{t-2} = 2, \quad s_{t-1} = 1, \quad s_t = 1, \\ (1 - p_{11})^{-1}(1 - p_{22})^{-1} & \text{if } s_{t-2} = 2, \quad s_{t-1} = 1, \quad s_t = 2, \\ 0 & \text{otherwise.} \end{cases}$$

In empirical estimation one often finds the regimes to be fairly persistent. When this is the case, the event ($s_{t-2} = 2$, $s_{t-1} = 1$, $s_t = 2$) will be rare, but, when it

occurs, produce a huge value for the product (3.18). Tests based on the presumed asymptotic normality of the sample mean of such a random variable could have quite poor performance, and indeed Monte Carlo investigation suggested that White's test applied to moments such as (3.18) can be quite unreliable.

Testing for predictability of the score with respect to p_{11} on the basis of other elements of $\boldsymbol{h}_{t-1}(\boldsymbol{\lambda})$ does seem useful, however. Of particular interest is whether the score with respect to the constant term for regime i helps predict the score with respect to the transition probability from state i to i. This is basically a means of verifying that

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

as presumed in the formulation investigated here.

To summarize, three specification statistics seem particularly interesting:

(i) *Dynamic specification test for autocorrelation.* This test chooses for $\boldsymbol{c}_t(\boldsymbol{\lambda})$ the K^2 elements of the form

$$\frac{\partial \log p(y_t | \mathcal{X}_t; \boldsymbol{\lambda})}{\partial \beta_j(1)} \cdot \frac{\partial \log p(y_{t-1} | \mathcal{X}_{t-1}; \boldsymbol{\lambda})}{\partial \beta_i(1)} \quad \text{for} \quad i, j = 1, \ldots, K, \tag{3.19}$$

where $\beta_i(1)$ denotes the coefficient associated with the constant term in the regression vector $\boldsymbol{\beta}_i$ appropriate when the process is in regime i.

(ii) *Dynamic specification test for ARCH effects.* This test chooses for $\boldsymbol{c}_t(\boldsymbol{\lambda})$ the K^2 elements of the form

$$\frac{\partial \log p(y_t | \mathcal{X}_t; \boldsymbol{\lambda})}{\partial \sigma_j^2} \cdot \frac{\partial \log p(y_{t-1} | \mathcal{X}_{t-1}; \boldsymbol{\lambda})}{\partial \sigma_i^2} \quad \text{for} \quad i, j = 1, \ldots, K. \tag{3.20}$$

(iii) *Dynamic specification test for validity of Markov assumptions.* This test chooses for $\boldsymbol{c}_t(\boldsymbol{\lambda})$ the $2K$ elements of the form

$$\frac{\partial \log p(y_t | \mathcal{X}_t; \boldsymbol{\lambda})}{\partial p_{ii}} \cdot \frac{\partial \log p(y_{t-1} | \mathcal{X}_{t-1}; \boldsymbol{\lambda})}{\partial \beta_i(1)}, \qquad i = 1, \ldots, K, \tag{3.21a}$$

$$\frac{\partial \log p(y_t | \mathcal{X}_t; \boldsymbol{\lambda})}{\partial p_{ii}} \cdot \frac{\partial \log p(y_{t-1} | \mathcal{X}_{t-1}; \boldsymbol{\lambda})}{\partial p_{ii}}, \qquad i = 1, \ldots, K. \tag{3.21b}$$

In each case the vector $\boldsymbol{c}_t(\hat{\boldsymbol{\lambda}})$ is used in conjunction with the score vector $\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})$ to construct the statistic (2.9). The first two statistics are asymptotically distributed as χ^2 with K^2 degrees of freedom; the last is $\chi^2(2K)$.

3.3. Lagrange multiplier tests

Suppose we wanted to test the Markov-switching regression model (3.3) against an alternative that allowed for autocorrelation of the regression residuals:

$$p(y_t|\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, y_{t-1}, s_t, s_{t-1}; \boldsymbol{\theta}, \phi)$$
$$= \frac{1}{\sqrt{2\pi}\sigma_{s_t}} \cdot \exp\left[\frac{-[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_{s_t}) - \phi(y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta}_{s_{t-1}})]^2}{2\sigma_{s_t}^2}\right]. \tag{3.22}$$

We wish to test the null hypothesis of no autocorrelation ($\phi = 0$). Calculations similar to those above reveal that

$$\left.\frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda}, \phi)}{\partial \phi}\right|_{\phi=0} = \sum_{i=1}^{K}\sum_{j=1}^{K} \psi_{t,j,i}\, p(s_t = j, s_{t-1} = i|\Omega_t; \boldsymbol{\lambda})$$
$$+ \sum_{\tau=1}^{t-1}\sum_{i=1}^{K}\sum_{j=1}^{K} \psi_{\tau,j,i}[p(s_\tau = j, s_{\tau-1} = i|\Omega_t; \boldsymbol{\lambda})$$
$$- p(s_\tau = j, s_{\tau-1} = i|\Omega_{t-1}; \boldsymbol{\lambda}], \tag{3.23}$$

where

$$\psi_{t,j,i} = \left.\frac{\partial \log p(y_t|\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, y_{t-1}, s_t = j, s_{t-1} = i; \boldsymbol{\theta}, \phi)}{\partial \phi}\right|_{\phi=0}$$
$$= \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_j)(y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta}_i)}{\sigma_j^2}.$$

Let $\hat{\boldsymbol{\lambda}}$ denote the MLE of

$$\boldsymbol{\lambda}' = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_K', \sigma_1^2, \ldots, \sigma_K^2, \boldsymbol{p}'),$$

ignoring any autocorrelation. Let $\boldsymbol{\lambda}^*$ denote the parameter vector of the more general model

$$\boldsymbol{\lambda}^* = (\boldsymbol{\lambda}', \phi)'$$

with associated constrained MLE $\tilde{\boldsymbol{\lambda}}^* = (\hat{\boldsymbol{\lambda}}', 0)'$. It is straightforward to verify that under the null hypothesis $\phi = 0$, the score for the more general model $\boldsymbol{h}_t(\tilde{\boldsymbol{\lambda}}^*)$ is obtained by vertically stacking that for the basic model $\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})$ on top of (3.23). The LM test using $\boldsymbol{h}_t(\tilde{\boldsymbol{\lambda}}^*)$ in Eq. (2.11) is then asymptotically $\chi^2(1)$.

Lagrange multiplier tests against a variety of other alternative specifications also take the form of (3.23). Table 1 presents the formulas necessary to implement LM tests against alternatives such as autocorrelation, heteroskedasticity, and omitted variables. To implement any of the tests, one only needs to estimate the model under the null and calculate $\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})$ in the manner described in Section 3.1 using the restricted estimates $\hat{\boldsymbol{\lambda}}$ to form the smoothed probabilities. One then

Table 1
Value to use for $\psi_{t,j,i}$ in Eq. (3.23) for Lagrange multiplier tests

To test for autocorrelation across regimes:

H_A: $(y_t|\mathcal{X}_t, s_t, s_{t-1}; \lambda^*) \sim N[\{x_t'\beta_{s_t} + \phi(y_{t-1} - x_{t-1}'\beta_{s_{t-1}})\}, \sigma^2_{s_t}]$

H_0: $\phi = 0$

$$\psi_{t,j,i} = \frac{(y_t - x_t'\beta_j)\cdot(y_{t-1} - x_{t-1}'\beta_i)}{\sigma_j^2}$$

Distribution of test statistic (2.11): $\chi^2(1)$

To test for autocorrelation within regime 1:

H_A: $(y_t|\mathcal{X}_t, s_t, s_{t-1}; \lambda^*) \sim N[\{x_t'\beta_{s_t} + \delta_{[s_t=1, s_{t-1}=1]}\phi_1(y_{t-1} - x_{t-1}'\beta_{s_{t-1}})\}, \sigma^2_{s_t}]$

H_0: $\phi_1 = 0$

$$\psi_{t,j,i} = \frac{(y_t - x_t'\beta_1)\cdot(y_{t-1} - x_{t-1}'\beta_1)}{\sigma_1^2} \quad \text{if} \quad i = 1, \quad j = 1$$

$$= 0 \qquad \text{otherwise}$$

Distribution of test statistic (2.11): $\chi^2(1)$

To test for ARCH effects:

H_A: $(y_t|\mathcal{X}_t, s_t, s_{t-1}; \lambda^*) \sim N[x_t'\beta_{s_t}, h_t]$

$$h_t = \gamma_{s_t}\left[1 + \frac{\xi\cdot(y_{t-1} - x_{t-1}'\beta_{s_{t-1}})^2}{\gamma_{s_{t-1}}}\right]$$

H_0: $\xi = 0$

$$\psi_{t,j,i} = \left[-1 + \frac{(y_t - x_t'\beta_j)^2}{\sigma_j^2}\right]\left[\frac{(y_{t-1} - x_{t-1}'\beta_i)^2}{2\sigma_i^2}\right]$$

Distribution of test statistic (2.11): $\chi^2(1)$

To test for variables omitted from mean:

H_A: $(y_t|\mathcal{X}_t, s_t; \lambda^*) \sim N[(x_t'\beta_{s_t} + z_t'\delta), \sigma^2_{s_t}]$

H_0: $\boldsymbol{\delta} = \mathbf{0}$

$$\boldsymbol{\psi}_{t,j,i} = \frac{(y_t - x_t'\beta_j)\cdot z_t}{\sigma_j^2}$$

Distribution of test statistic (2.11): $\chi^2(l)$ $z_t = (l \times 1)$

To test for variables omitted from variance:

H_A: $(y_t|\mathcal{X}_t, s_t; \lambda^*) \sim N[x_t'\beta_{s_t}, \sigma^2_{s_t}[1 + z_t'\eta]]$

H_0: $\boldsymbol{\eta} = \mathbf{0}$

$$\boldsymbol{\psi}_{t,j,i} = \left[-1 + \frac{(y_t - x_t'\beta_j)^2}{\sigma_j^2}\right]\cdot(z_t/2)$$

Distribution of test statistic (2.11): $\chi^2(l)$ $z_t = (l \times 1)$

stacks this vector $\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})$ on top of the function of the smoothed probabilities given in (3.23), and plugs the resulting expression into (2.11). For example, to test whether an $(l \times 1)$ vector of variables z_t has been left out of the description of the mean of y_t, one puts the $(l \times 1)$ vector

$$\sum_{j=1}^{K} \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_j)\cdot \boldsymbol{z}_t}{\sigma_j^2} \cdot p(s_t = j|\Omega_t; \hat{\boldsymbol{\lambda}}) + \sum_{\tau=1}^{t-1} \sum_{j=1}^{K} \frac{(y_\tau - \boldsymbol{x}_\tau'\boldsymbol{\beta}_j)\cdot \boldsymbol{z}_\tau}{\sigma_j^2} \cdot [p(s_\tau = j|\Omega_t; \hat{\boldsymbol{\lambda}}) - p(s_\tau = j|\Omega_{t-1}; \hat{\boldsymbol{\lambda}})] \tag{3.24}$$

at the bottom of $\boldsymbol{h}_t(\hat{\boldsymbol{\lambda}})$, and substitutes the resulting $\boldsymbol{h}_t(\tilde{\boldsymbol{\lambda}}^*)$ into (2.11) to obtain a χ^2 statistic with l degrees of freedom under the null hypothesis that z_t does not matter.

The common form of the different tests makes it easy to implement Andrews's (1993) test for structural change on an estimated Markov-switching time-series model. Let $z_t(\tau)$ be a scalar that is equal to zero for all $t < \tau$ and equal to unity for all $t \geqslant \tau$. In this case the statistic constructed from (3.24) is a Lagrange multiplier test of the null that the data are accurately modeled by the Markov-switching model against the alternative that there is a further shift in the mean of the process at date τ beyond that captured by the estimated model. One can calculate this LM test statistic for all τ in the range $0.15\ T \leqslant \tau \leqslant 0.85\ T$, where T is the sample size, and find the value of τ that produces the largest statistic. This maximum LM test statistic has an asymptotic distribution corresponding to the $\pi_0 = 0.15$, $p = 1$ entry in Table 1 of Andrews (1993).

One issue which comes up in practice (but is irrelevant asymptotically) is that evaluation of (3.23) requires observation of y_0 and $\boldsymbol{x}_0$ which were not used in estimation of the parameters $\hat{\boldsymbol{\lambda}}$. The solution adopted in the Monte Carlo experiments reported in this paper was simply to set the score for observation 1 equal to its expected value of zero, and let the summation in (3.23) begin with $\tau = 2$ rather than $\tau = 1$. Another solution would be not to use some of the data for estimation, saving it for specification testing.

4. Monte Carlo analysis of tests

The small-sample properties of the above tests were investigated for the following parameterization of the simple two-state process (1.1)–(1.2):

$$\mu_1 = 2, \quad \mu_2 = -2, \quad p_{11} = p_{22} = 0.8, \quad \sigma_1^2 = \sigma_2^2 = 1.$$

For any constants k_μ and k_σ, changing the mean parameters to $k_\sigma(\mu_i + k_\mu)$ and the variance parameters to $[k_\sigma]^2 \cdot \sigma_i^2$ for $i = 1, 2$ would yield numerically

identical results to those reported in Tables 3 through 5.[2] Also, the chosen parameterization is symmetric between states 1 and 2, permitting pooling some of the results across parameters to improve the accuracy of the Monte Carlo estimates.

One thousand Monte Carlo realizations from samples of size $T = 50$ and 100 were generated from this model.[3] For each realization, the EM algorithm described in Hamilton (1990) was used to calculate the maximum likelihood estimates for that sample.[4]

One design difficulty with Monte Carlo analysis is the presence of multiple local maxima of the sample likelihood function. The single-state solution ($\hat{\mu}_1 = \hat{\mu}_2 = \bar{y}$, $\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \sum_{t=1}^{T}(y_t - \bar{y})^2/T$, $\hat{p}_{11} = \hat{p}_{22} = 0.5$) is always a local maximum (see Hamilton, 1990), and there are troublesome corner conditions and singularities in the likelihood surface such as $\hat{p}_{ii} = 0$ and $\hat{\sigma}_i^2 \to 0$. In practice, an econometrician should investigate several hundred values from which to start the maximization iterations. Hamilton (1991) suggested minor adjustments to the objective function that may also be helpful in choosing among local maxima. Unfortunately, completing this careful evaluation of each Monte Carlo draw is at the present time computationally infeasible. Instead, it is possible to avoid these issues to some degree by choosing a parameterization where the difference in means is quite dramatic relative to the standard errors, and by starting the EM iterations for each sample at the true parameter values. Thus only one local maximum was found for each sample, and this could conceivably either underestimate or overestimate the errors a practical econometrician would make after a more exhaustive investigation of the same data.

Table 2 reveals that maximum-likelihood estimates of all the parameters are slightly biased toward zero; in a sample of size 50, this bias is 0.3% for the mean and 2% for the variance. Table 2 further summarizes the standard errors of the maximum-likelihood estimates, calculated as the square root of the average

[2] Following the suggestion of Hendry (1984), this was confirmed numerically as one of many checks of the validity of the Monte Carlo experiment.

[3] Pseudo-uniform random numbers were generated by the routine RNDU in version 1.49b of the GAUSS programming language. The algorithm generates uniform variates via the multiplicative-congruential method. Starting with a random seed chosen from the clock, the routine generates the variable u_{i+1} from

$$s_{i+1} = b \cdot s_i \quad [\text{mod } r], \qquad u_{i+1} = s_{i+1}/r,$$

with $b = 397204094$ and $r = 2^{31} - 1$. Such uniform variates were used to generate realizations of the state variable s_t. Gaussian variates were generated from separate uniform variates by the GAUSS routine RNDN, which uses the fast acceptance–rejection algorithm proposed in Kinderman and Ramage (1976).

[4] The EM iteration proceeded until the maximal change in the parameter vector was less than 1×10^{-5} or until 400 iterations had been completed, whichever came first.

Table 2
Small-sample bias and standard errors of maximum-likelihood estimates for samples of size $T = 50$ and $T = 100$; true model: $\mu_1 = 2$, $\mu_2 = -2$, $p_{11} = 0.8$, $p_{22} = 0.8$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$

	$\hat{\mu}_1$	$\hat{p}_{11}$	$\hat{\sigma}_1^2$
$T = 50$			
Bias ($E[\hat{\lambda}_i - \lambda_i^0]$)	−0.006	−0.011	−0.021
Average actual standard error	0.26	0.10	0.40
Average estimated standard error	0.20	0.07	0.26
$T = 100$			
Bias ($E[\hat{\lambda}_i - \lambda_i^0]$)	−0.001	−0.008	−0.018
Average actual standard error	0.16	0.06	0.25
Average estimated standard error	0.15	0.06	0.21

value of $(\hat{\lambda}_i - \lambda_i^{(0)})^2$ across the 1000 Monte Carlo samples. The average value of $\hat{\mathcal{I}}_{OP}$ in Eq. (2.12) across Monte Carlo samples was also calculated,[5] and standard errors were constructed from the square root of $(1/T)$ times the diagonal elements of the resulting $\hat{\mathcal{I}}_{OP}^{-1}$ as an indication of the validity of standard errors based on (2.13). For a sample of size 50, expression (2.13) would typically understate the true standard errors for the mean parameters by 25%, for the transition probabilities by 30%, and for the variance parameters by 35%. For a sample of size 100, the standard errors from (2.13) are accurate on average for the means and transition probabilities, though continue to understate the true values for the variances by 16%.

Table 3 reports the consequences of using the estimated standard errors to construct t-tests of hypotheses for individual parameters. Even for $T = 50$, the distribution of $(\hat{\mu}_1 - \mu_1^{(0)})/\hat{\sigma}_{\hat{\mu}_1}$ is reasonably symmetric, and one will not go far wrong in treating it as N(0, 1). T-tests for the transition probabilities, $(\hat{p}_{11} - p_{11}^{(0)})/\hat{\sigma}_{\hat{p}_{11}}$ tend to be strongly skewed rightward, owing apparently to a negative covariance between $\hat{p}_{11}$ and $\hat{\sigma}_{\hat{p}_{11}}$. Even so, for $T = 50$, comparing $(\hat{p}_{11} - p_{11}^{(0)})/\hat{\sigma}_{\hat{p}_{11}}$ with ± 2 is not a bad summary of the evidence in the data against H_0: $p_{11} = p_{11}^{(0)}$. By contrast, t-statistics for the variances $(\hat{\sigma}_1^2 - \sigma_1^2(0))/\hat{\sigma}_{\hat{\sigma}_1^2}$ are severely skewed leftward, and cannot be trusted in small samples.

Table 4 summarizes the performance of the Newey–Tauchen–White dynamic specification tests. The suggested specification tests are more difficult for a correctly specified model to pass than their asymptotic approximations might indicate. For example, a nominal 5% test for autocorrelation would lead to

[5] Note that this calculation does not produce the plim of (2.12); rather, it converges to the expected value of (2.12) for a given fixed T.

Table 3
Monte-Carlo-generated critical points for t-tests on individual parameters

	$\hat{\mu}_1$	$\hat{p}_{11}$	$\hat{\sigma}_1^2$
$T = 50$			
$\pi = 0.01$	− 2.44	− 2.05	− 5.54
$\pi = 0.05$	− 1.70	− 1.42	− 2.90
$\pi = 0.95$	1.88	2.21	1.00
$\pi = 0.99$	2.97	3.56	1.44
$T = 100$			
$\pi = 0.01$	− 2.33	− 2.11	− 4.08
$\pi = 0.05$	− 1.62	− 1.54	− 2.44
$\pi = 0.95$	1.78	1.86	1.10
$\pi = 0.99$	2.52	2.92	1.60
$T = \infty$			
$\pi = 0.01$	− 2.33	− 2.33	− 2.33
$\pi = 0.05$	− 1.64	− 1.64	− 1.64
$\pi = 0.95$	1.64	1.64	1.64
$\pi = 0.99$	2.33	2.33	2.33

All statistics are asymptotically N(0, 1). The table reports the Monte Carlo estimate of the value ξ such that $\mathrm{P}[(\hat{\lambda}_i - \lambda_i^0)/\hat{\sigma}_{\lambda_i} < \xi] = \pi$ where λ_i^0 is the true parameter value.

rejection of a correctly specified model 18% of the time for a sample of size 50 and 10% of the time for a sample of size 100.

Table 5 reports the small-sample properties of the Lagrange multiplier diagnostics proposed in Section 3. The 'omitted variable' z_t was a dummy variable that took on the value unity for observations 26 through 38 and zero elsewhere. Table 5 suggests that Lagrange multiplier tests for autocorrelation or a variable omitted from the mean perform fairly well in small samples. Tests for omitted ARCH or other misspecification of the variance perform somewhat poorer, though still perhaps slightly better than the Newey–Tauchen–White specification tests.

The small-sample properties of specification tests are often better approximated by an F distribution than by an asymptotically equivalent χ^2 distribution – see for example MacKinnon and White (1985), Kiviet (1986), and Ericsson (1991). This small-sample adjustment involves two steps. First, there is a degree-of-freedom adjustment to the estimate of the variance–covariance matrix that appears in the denominator of the test statistics. Second, an F distribution rather than a χ^2 distribution is used to interpret the resulting test statistics. Both of these steps help improve the small-sample performance of any of the tests reported in Tables 4 or 5.

Table 4
Monte-Carlo-generated critical points and percent rejections using asymptotic 5% and 1% critical values for Newey–Tauchen–White dynamic specification tests

	Test for autocorrelation	Test for ARCH	Test for violation of first-order Markov assumption
Critical points for test with significance level given by $1-\pi$:			
$T = 50$			
$\pi = 0.90$	11.68	10.88	11.82
$\pi = 0.95$	13.98	12.91	15.36
$\pi = 0.99$	19.84	17.32	26.34
$T = 100$			
$\pi = 0.90$	9.54	9.62	9.50
$\pi = 0.95$	12.16	11.30	11.77
$\pi = 0.99$	16.84	15.66	15.49
$T = \infty$			
$\pi = 0.90$	7.78	7.78	7.78
$\pi = 0.95$	9.49	9.49	9.49
$\pi = 0.99$	13.28	13.28	13.28
Percent rejections using critical values asymptotically appropriate for test of size α:			
$T = 50$			
$\alpha = 0.05$	18%	17%	18%
$\alpha = 0.01$	6%	4%	7%
$T = 100$			
$\alpha = 0.05$	10%	10%	10%
$\alpha = 0.01$	3%	2%	3%
$T = \infty$			
$\alpha = 0.05$	5%	5%	5%
$\alpha = 0.01$	1%	1%	1%

All statistics (here denoted $\hat{\vartheta}$) are asymptotically $\chi^2(4)$. The first panel reports the Monte Carlo estimate of the value ξ such that $P[\hat{\vartheta} < \xi] = \pi$. The second panel reports the Monte Carlo estimate of the probability $P[\hat{\vartheta} > \xi_\alpha]$ where, if $X \sim \chi^2(4)$, $P[X > \xi_\alpha] = \alpha$.

To implement the first step for the White specification tests, an 'approximately bias-corrected' version of the variance–covariance matrix $\hat{A}$ in (2.10) would be divided by $T - m$ rather than by T, where m is the number of estimated parameters. Using such an estimate would amount to multiplying the White test statistic in (2.9) by $(T - m)/T$. For example, for a sample of size

Table 5
Monte-Carlo-generated critical points and percent rejections using asymptotic 5% and 1% critical values for Lagrange multiplier tests

	Autocorrelation in regime I	Autocorrelation across regimes	ARCH	Variable omitted from mean	Variable omitted from variance
Critical points for test with significance level given by $1-\pi$:					
$T = 50$					
$\pi = 0.90$	3.77	3.54	4.31	3.62	4.66
$\pi = 0.95$	5.44	4.97	5.58	5.23	6.54
$\pi = 0.99$	9.12	8.44	8.66	8.87	10.46
$T = 100$					
$\pi = 0.90$	3.26	3.26	3.49	3.24	4.16
$\pi = 0.95$	4.80	4.66	4.90	4.47	5.88
$\pi = 0.99$	8.60	7.69	9.68	7.92	9.78
$T = \infty$					
$\pi = 0.90$	2.71	2.71	2.71	2.71	2.71
$\pi = 0.95$	3.84	3.84	3.84	2.84	3.84
$\pi = 0.99$	6.63	6.63	6.63	6.63	6.63
Percent rejections using critical values asymptotically appropriate for test of size α:					
$T = 50$					
$\alpha = 0.05$	10%	9%	13%	9%	14%
$\alpha = 0.01$	3%	2%	3%	3%	5%
$T = 100$					
$\alpha = 0.05$	8%	8%	8%	7%	11%
$\alpha = 0.01$	3%	2%	2%	2%	4%
$T = \infty$					
$\alpha = 0.05$	5%	5%	5%	5%	5%
$\alpha = 0.01$	1%	1%	1%	1%	1%

All statistics (here denoted $\hat{\sigma}$) are asymptotically $\chi^2(1)$. The first panel reports the Monte Carlo estimate of the value ξ such that $P[\hat{\sigma} < \xi] = \pi$. The second panel reports the Monte Carlo estimate of the probability $P[\hat{\sigma} > \xi_\alpha]$ where, if $X \sim \chi^2(1)$, $P[X > \xi_\alpha] = \alpha$.

$T = 50$, Monte Carlo estimates of the 5% critical values for a test statistic given by $(T - m)/T$ times the magnitude in (2.9) are 12.3, 11.4, and 13.5 for the tests investigated in Table 4.

For the second step, these adjusted test statistics are then divided by $l = 4$ and compared with an $F(l, T - m)$ distribution rather than with a $\chi^2(l)$ distribution

as originally written. For example, for $T = 50$, the 5% critical value for an $F(4, 44)$ variable is 2.58, which multiplied by four is 10.3. This F critical value of 10.3 gives a much better approximation to the degree-of-freedom corrected test statistics than the $\chi^2(4)$ critical value of 9.49 gives to the unadjusted statistics.

The recommended small-sample procedure to implement the White tests for specification is thus to multiply the statistic in (2.9) by $(T - m)/(Tl)$ and compare the resulting statistic with an $F(l, T - m)$ distribution. Such a test statistic has the same asymptotic properties as does the statistic in (2.9), but has a much better small-sample performance based on the results in Table 4.

Analogously, if m is the total number of parameters and m_0 is the number of restrictions, asymptotic Lagrange multiplier tests with better small-sample performance are obtained by multiplying (2.11) by $(T - m + m_0)/(Tm_0)$ and comparing the result with an $F(m_0, T - m + m_0)$ distribution.

To summarize, all of the tests are more difficult for a correctly specified model to pass than one would have anticipated based on the asymptotic distribution. Monte Carlo studies of other models have produced similar results for the outer product of the gradient version of the LM test (e.g., Davidson and MacKinnon, 1983; Godfrey, McAleer, and McKenzie, 1988) and for White's (1982) information matrix test (e.g., Kennan and Neumann, 1988). In a sample as small as 50 observations, one might be better off using the 1% critical value from the asymptotic distributions (rather than the 5% value) as a rough guide for a 5% small-sample test based on the Newey–Tauchen–White specification tests or Lagrange multiplier tests for misspecification of the variance. Otherwise, an F distribution and degree-of-freedom correction factor should probably be used in place of a χ^2 distribution for interpreting the test statistics for small sample sizes.

5. Application

The above tests were applied to Engel and Hamilton's (1990) study of exchange rates. A model of the form of (1.1)–(1.2) with $K = 2$ states was fit for y_t the quarterly percent change in the number of U.S. dollars required to purchase a German mark. Similar models were estimated for the dollar–franc and dollar–pound exchange rates. Each sample consisted of $T = 58$ observations beginning in 1973:IV and ending in 1988:I. Table 6 applies the tests developed here to Engel and Hamilton's specification. The p-values reported in parentheses rely on the small-sample F approximations discussed in Section 4.

None of the White specification tests or Lagrange multiplier tests reject the specification at the 5% level, although the tests for omitted heteroskedasticity for the British pound come close to rejecting at the 5% level.

Table 6
Specification tests for exchange-rate application

Test	Germany	France	U.K.
White test for autocorrelation $\approx F(4, 52)$	0.57	0.95	1.03
	(0.68)	(0.44)	(0.40)
White test for ARCH $\approx F(4, 52)$	0.79	1.63	2.36
	(0.54)	(0.18)	(0.06)
White test of Markov specification $\approx F(4, 52)$	0.61	1.00	0.36
	(0.66)	(0.42)	(0.84)
LM test for autocorrelation in regime 1 $\approx F(1, 52)$	0.00	2.92	2.71
	(1.00)	(0.09)	(0.11)
LM test for autocorrelation in regime 2 $\approx F(1, 52)$	1.29	0.22	0.27
	(0.26)	(0.64)	(0.61)
LM test for autocorrelation across regimes $\approx F(1, 52)$	0.53	0.95	0.58
	(0.47)	(0.33)	(0.45)
LM test for ARCH $\approx F(1, 52)$	1.14	0.09	4.01
	(0.29)	(0.77)	(0.05)
Andrews LM test for shift in mean during 1975:IV–1985:IV	11.2	12.5	3.3

The indicated F distribution gives the recommended small-sample guide to the asymptotic distribution of the test statistic. White statistics were calculated as $(58 - 6)/(4 \times 58)$ times the magnitude in (2.9), while LM statistics were calculated as $(58 - 6)/58$ times the magnitude in (2.11). P-values calculated using this recommended distribution are reported in parentheses. The asymptotic 5% critical value for the Andrews test is 8.85 and the 1% critical value is 12.35.

To implement Andrews's (1993) test for parameter stability, an LM test for a shift in the mean of the process was calculated for every possible change point between 1975:IV and 1985:IV, omitting the possibility of a change point in the first 15% or last 15% of the sample. The biggest test statistic occurred in 1985:I for the mark, 1985:III for the franc, and 1981:IV for the pound. Using the critical values calculated by Andrews, the evidence of a structural break for the mark is significant at the 5% level while that for the franc is significant at the 1% level. However, Andrew's test assumes that for any given break point the test statistic is approximately $\chi^2(1)$. Since the Monte Carlo results indicate that somewhat wider tails should be used than this, a larger critical value than that calculated by Andrews is probably appropriate in small samples.

To conclude, Engel and Hamilton's specification appears to have accurately captured a number of the key features of the data, although there is some suggestion in the case of the mark and the franc that there is an additional structural change in 1985 that is not captured by the model.

6. Summary

This paper provided an analytic representation for the conditional score statistic for Markov-switching time-series models, and suggested a variety of specification tests based on these scores. These tests can be implemented from the same set of calculations that are used to construct the 'smoothed' probability that the process was in regime i at date τ, and do not require estimation of additional parameters by maximum likelihood.

The specification tests appear to be too stringent; if applied to small samples, an asymptotic 5% test will lead to rejection of a correctly specified model more than 5% of the time. For all of the Newey–Tauchen–White tests and for the Lagrange multiplier tests for misspecification of the variance, this bias is sufficiently severe in a sample of size 50 that a critical value that would yield a 1% test according to the asymptotic distribution may be a reasonable guide for a 5%-size small-sample test. Alternatively, an F distribution along with a degree-of-freedom correction might be used to interpret the test statistics in place of a χ^2 distribution. The bias is less significant for a sample of size 100 or for Lagrange multiplier tests for autocorrelation or omitted variables.

Appendix A

This appendix derives Eqs. (3.7), (3.12), and (3.23).

Let script letters denote the full history of a variable through date t:

$$\mathcal{Y}_t' \equiv (y_t, y_{t-1}, \ldots, y_1)$$

$$\mathcal{S}_t' \equiv (s_t, s_{t-1}, \ldots, s_1)$$

$$\mathcal{W}_t' \equiv (\boldsymbol{w}_t', \boldsymbol{w}_{t-1}', \ldots, \boldsymbol{w}_1', y_0, y_{-1}, \ldots, y_{-r+1}).$$

Further let $\int \mathrm{d}\mathcal{S}_t$ denote summation over all of the K^t possible values of $\mathcal{S}_t$; for example,

$$\int g(\mathcal{S}_t)\,\mathrm{d}\mathcal{S}_t \equiv \sum_{s_1=1}^{K} \sum_{s_2=1}^{K} \cdots \sum_{s_t=1}^{K} g(s_t, s_{t-1}, \ldots, s_1).$$

(Such a large number of summations never needs to be computed in practice, but is only used as an expository aid here in describing the form of the score.)

Under our assumptions of exogeneity of $\boldsymbol{w}$ and lack of dependence of s on y or $\boldsymbol{w}$, the observed likelihood of observations 1 through t could be written as

$$p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda}) = \int p(\mathcal{Y}_t|\mathcal{W}_t, \mathcal{S}_t; \boldsymbol{\theta}) \cdot p(\mathcal{S}_t; \boldsymbol{p})\,\mathrm{d}\mathcal{S}_t, \tag{A.1}$$

where

$$p(\mathcal{Y}_t|\mathcal{W}_t, \mathcal{S}_t; \boldsymbol{\theta}) = \prod_{\tau=1}^{t} p(y_\tau|\boldsymbol{x}_\tau, s_\tau; \boldsymbol{\theta}), \tag{A.2}$$

$$p(\mathcal{S}_t; \boldsymbol{p}) = p(s_1; \boldsymbol{p}) \cdot \prod_{\tau=2}^{t} p(s_\tau|s_{\tau-1}; \boldsymbol{p}). \tag{A.3}$$

Then the derivative of the log-likelihood of the first t observations is

$$\begin{aligned}\frac{\partial \log p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} &= \frac{1}{p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda})} \int \frac{\partial p(\mathcal{Y}_t|\mathcal{W}_t, \mathcal{S}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} p(\mathcal{S}_t; \boldsymbol{p})\, \mathrm{d}\mathcal{S}_t \\ &= \int \frac{\partial \log p(\mathcal{Y}_t|\mathcal{W}_t, \mathcal{S}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{p(\mathcal{Y}_t|\mathcal{W}_t, \mathcal{S}_t; \boldsymbol{\theta}) \cdot p(\mathcal{S}_t; \boldsymbol{p})}{p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda})} \mathrm{d}\mathcal{S}_t \\ &= \int \frac{\partial \log p(\mathcal{Y}_t|\mathcal{W}_t, \mathcal{S}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} p(\mathcal{S}_t|\mathcal{Y}_t, \mathcal{W}_t; \boldsymbol{\lambda})\, \mathrm{d}\mathcal{S}_t. \end{aligned} \tag{A.4}$$

Using (A.2), Eq. (A.4) becomes

$$\frac{\partial \log p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} = \sum_{\tau=1}^{t} \sum_{s_\tau=1}^{K} \frac{\partial \log p(y_\tau|\boldsymbol{x}_\tau, s_\tau; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \times \{p(s_\tau|\Omega_t)\}, \tag{A.5}$$

where $\Omega_t = \{\mathcal{Y}_t, \mathcal{W}_t\}$.

For $t = 1$ we have $\mathcal{Y}_1 = y_1$ and $\mathcal{W}_1 = \mathcal{X}_1$, so (A.5) for $t = 1$ could be written

$$\frac{\partial \log p(y_1|\mathcal{X}_1; \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} = \sum_{s_1=1}^{K} \frac{\partial \log p(y_1|\boldsymbol{x}_1, s_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \times \{p(s_1|\Omega_1)\}. \tag{A.6}$$

Recall that

$$\log p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda}) = \sum_{\tau=1}^{t} \log p(y_\tau|\mathcal{X}_\tau; \boldsymbol{\lambda}),$$

meaning it must be the case that

$$\frac{\partial \log p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} = \sum_{\tau=1}^{t} \frac{\partial \log p(y_\tau|\mathcal{X}_\tau; \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}}. \tag{A.7}$$

From (A.5)–(A.7), we deduce that the score of observation t is given by (A.6) when $t = 1$ and by

$$\frac{\partial \log p(y_t|\mathcal{X}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} = \sum_{s_t=1}^{K} \frac{\partial \log p(y_t|\boldsymbol{x}_t, s_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot \{p(s_t|\Omega_t)\} + \sum_{\tau=1}^{t-1} \sum_{s_\tau=1}^{K} \frac{\partial \log p(y_\tau|\boldsymbol{x}_\tau, s_\tau; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot \{p(s_\tau|\Omega_t) - p(s_\tau|\Omega_{t-1})\}, \tag{A.8}$$

for $t = 2, 3, \ldots, T$. That is, one can show that (A.6) and (A.8) satisfy (A.7) and (A.5) by induction. This completes the derivation of (3.7).

Where the distribution of y_t depends on the current and previous state, as in a specification with autocorrelation across regimes, it is straightforward to show that (A.5) generalizes to

$$\frac{\partial \log p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} = \sum_{\tau=1}^{t} \sum_{s_\tau=1}^{K} \sum_{s_{\tau-1}=1}^{K} \frac{\partial \log p(y_\tau|\boldsymbol{x}_\tau, s_\tau, s_{\tau-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} p(s_\tau, s_{\tau-1}|\Omega_t), \tag{A.9}$$

from which Eq. (3.23) follows.

Scores with respect to the transition probabilities are found similarly from

$$\frac{\partial \log p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda})}{\partial \boldsymbol{p}} = \int \frac{\partial \log p(\mathcal{S}_t; \boldsymbol{p})}{\partial \boldsymbol{p}} p(\mathcal{S}_t|\mathcal{Y}_t, \mathcal{W}_t; \boldsymbol{\lambda}) \, \mathrm{d}\mathcal{S}_t.$$

Using (A.3) and (3.6), we find

$$\frac{\partial \log p(\mathcal{Y}_t|\mathcal{W}_t; \boldsymbol{\lambda})}{\partial p_{ij}} = \sum_{s_1=1}^{K} \frac{\partial \log p(s_1; \boldsymbol{p})}{\partial p_{ij}} p(s_1|\Omega_t) + \sum_{\tau=2}^{t} p_{ij}^{-1} \cdot p(s_\tau = j, s_{\tau-1} = i|\Omega_t) - \sum_{\tau=2}^{t} p_{iK}^{-1} \cdot p(s_\tau = K, s_{\tau-1} = i|\Omega_t),$$

from which we get Eqs. (3.12) and (3.13).

Appendix B

This appendix describes an iterative routine for calculating (3.8).

Suppose we have used information available at $t - 1$ to form an inference about the value of the state at dates $t - 1$, τ, and $\tau - 1$:

$$p(s_{t-1}, s_\tau, s_{\tau-1}|\Omega_{t-1}), \qquad \tau < t. \tag{B.1}$$

To update inference (B.1) with a new observation for date t, we simply calculate

$$p(s_t, s_\tau, s_{\tau-1}|\Omega_t) = \frac{\sum_{s_{t-1}=1}^{K} p(s_{t-1}, s_\tau, s_{\tau-1}|\Omega_{t-1}) \cdot p(s_t|s_{t-1}) \cdot p(y_t|x_t, s_t)}{\sum_{s_t=1}^{K}\sum_{s_{t-1}=1}^{K}\sum_{s_\tau=1}^{K}\sum_{s_{\tau-1}=1}^{K} p(s_{t-1}, s_\tau, s_{\tau-1}|\Omega_{t-1}) p(s_t|s_{t-1}) \cdot p(y_t|x_t, s_t)}, \quad \text{(B.2a)}$$

for $\tau = 1, 2, \ldots, t-1$. For $\tau = t$, we use

$$p(s_t = i, s_\tau = j, s_{\tau-1} = k|\Omega_t) = \sum_{l=1}^{K} p(s_t = i, s_{\tau-1} = k, s_{\tau-2} = l|\Omega_t) \quad \text{if } i = j,$$

$$= 0 \quad \text{if } i \neq j. \quad \text{(B.2b)}$$

The change in the assessed probability of the event $[s_\tau, s_{\tau-1}]$ based on date t's information is then

$$p(s_\tau, s_{\tau-1}|\Omega_t) - p(s_\tau, s_{\tau-1}|\Omega_{t-1}) = \sum_{s_t=1}^{K} p(s_t, s_\tau, s_{\tau-1}|\Omega_t) - \sum_{s_{t-1}=1}^{K} p(s_{t-1}, s_\tau, s_{\tau-1}|\Omega_{t-1}), \quad \text{(B.3)}$$

while (3.8), the change in the assessed probability of the event (s_τ), is

$$p(s_\tau|\Omega_t) - p(s_\tau|\Omega_{t-1}) = \sum_{s_{\tau-1}=1}^{K} [p(s_\tau, s_{\tau-1}|\Omega_t) - p(s_\tau, s_{\tau-1}|\Omega_{t-1})]. \quad \text{(B.4)}$$

The technique is thus to iterate on Eq. (B.2) for $t = \tau + 1, \tau + 2, \ldots, T$, and calculate the desired change in assessed probabilities, (B.3) and (B.4), as a by-product of this iteration.

In the computer programs to implement this recursion, the quantity $p(s_t, s_\tau, s_{\tau-1}|\Omega_t)$ was coded as a $(t \times K^3)$ matrix. The K^3 elements comprising the τth row of this matrix are associated with $p(s_t, s_\tau, s_{\tau-1}|\Omega_t)$ for the K^3 different values for $(s_t, s_\tau, s_{\tau-1})$. At iteration t a complete new matrix is produced, and the change over the $t-1$ value of the matrix was used as in (B.3) and (B.4) to weight the terms $\partial \log p(y_t|x_t, s_t; \theta)/\partial\theta$ in (3.7) or terms involving p_{ij}^{-1} or p_{iK}^{-1} in (3.12). Thus implementing this procedure only requires storage of two $(t \times K^3)$ matrices at any one time.

For starting values for this iteration, one could set $p(s_0 = q|\Omega_0)$ equal to the ergodic probability π_q. Then we can evaluate (B.1) for $t - 1 = \tau = 1$ as

$$p(s_1 = i, s_1 = j, s_0 = k|\Omega_1)$$

$$= \frac{p(y_1|\boldsymbol{x}_1, s_1 = j)\cdot p(s_1 = j|s_0 = k)\cdot p(s_0 = k)}{\sum_{j=1}^{K}\sum_{\ell=1}^{K} p(y_1|\boldsymbol{x}_1, s_1 = j)\cdot p(s_1 = j|s_0 = \ell)\cdot p(s_0 = \ell)} \quad \text{if } i = j,$$

$$= 0 \quad \text{if } i \neq j. \tag{B.5}$$

The approach is thus to begin with (B.5) and then iterate on (B.2a, b) for $t = 2, 3, \ldots, T$.

References

Aitchison, J. and S.D. Silvey, 1958, Maximum likelihood estimation of parameters subject to restraints, Annals of Mathematical Statistics 29, 813–828.

Andrews, Donald W.K., 1993, Tests for parameter instability and structural change with unknown change point, Econometrica 61, 821–856.

Berndt, E.K., B.H. Hall, R.E. Hall, and J.A. Hausman, 1974, Estimation and inference in nonlinear structural models, Annals of Economic and Social Measurement 3/4, 653–665.

Breusch, T.S. and A.R. Pagan, 1980, The Lagrange multiplier test and its applications to model specification in econometrics, Review of Economic Studies 47, 239–253.

Cecchetti, Stephen G., Pok-sang Lam, and Nelson Mark, 1990a, Evaluating empirical tests of asset pricing models: Alternative interpretations, American Economic Review 80, 48–51.

Cecchetti, Stephen G., Pok-sang Lam, and Nelson Mark, 1990b, Mean reversion in equilibrium asset prices, American Economic Review 80, 398–418.

Chiang, Chin Long, 1980, An introduction to stochastic processes and their applications (Krieger, New York, NY).

Cosslett, Stephen R. and Lung-Fei Lee, 1985, Serial correlation in discrete variable models, Journal of Econometrics 27, 79–97.

Davidson, Russel and James G. MacKinnon, 1983, Small sample properties of alternative forms of the Lagrange multiplier test, Economics Letters 12, 269–275.

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., 1982a, A general approach to Lagrange multiplier model diagnostics, Journal of Econometrics 20, 83–104.

Engle, Robert F., 1982b, Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, Econometrica 50, 987–1007.

Engle, Robert F., 1984, Wald, likelihood ratio, and Lagrange multiplier tests in econometrics, in: Z. Griliches and M.D. Intriligator, eds., Handbook of econometrics, Vol. II (North-Holland, Amsterdam).

Ericsson, Neil R., 1991, Monte Carlo methodology and the finite sample properties of instrumental variables statistics for testing nested and non-nested hypotheses, Econometrica 59, 1249–1277.

Gallant, A. Ronald, and Halbert White, 1988, A unified theory of estimation and inference for nonlinear dynamic models (Basil Blackwell, Oxford).

Garcia, René and Pierre Perron, 1989, An analysis of the real interest rate under regime shifts, Mimeo. (Princeton University, Princeton, NJ).

Godfrey, L.G. and M.R. Wickens, 1981, Testing linear and log-linear regressions for functional form, Review of Economic Studies 48, 487–496.

Godfrey, L.G., M. McAleer, and C.R. McKenzie, 1988, Variable addition and Lagrange multiplier tests for linear and logarithmic regression models, Review of Economics and Statistics 70, 492–503.

Goldfeld, Stephen M. and Richard M. Quandt, 1973, A Markov model for switching regressions, Journal of Econometrics 1, 3–16.

Hamilton, James D., 1988, 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., 1989, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica 57, 357–384.

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

Hamilton, James D., 1991, A quasi-Bayesian approach to estimating parameters for mixtures of normal distributions, Journal of Business and Economic Statistics 9, 27–39.

Hendry, David F., 1984, Monte Carlo experimentation in econometrics, in: Z. Griliches and M.D. Intriligator, eds., Handbook of econometrics, Vol. II (North-Holland, Amsterdam).

Kennan, John and George R. Neumann, 1988, Why does the information matrix test reject too often? A diagnosis of some Monte Carlo symptoms, Mimeo. (University of Iowa, Iowa City, IA).

Kinderman, A.J. and J.G. Ramage, 1976, Computer generation of normal random variables, Journal of the American Statistical Association 71, 893–896.

Kiviet, Jan, 1986, On the rigour of some misspecification tests for modelling dynamic relationships, Review of Economic Studies 53, 241–261.

Lam, Pok-sang, 1990, The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series, Journal of Monetary Economics 26, 409–432.

MacKinnon, James G. and Halbert White, 1985, Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties, Journal of Econometrics 29, 305–325.

Newey, Whitney K., 1985, Maximum likelihood specification testing and conditional moment tests, Econometrica 53, 1047–1070.

Orme, Chris, 1988, The calculation of the information matrix test for binary data models, The Manchester School of Economic and Social Studies 56, 370–376.

Orme, Chris, 1990, The small-sample performance of the information-matrix test, Journal of Econometrics 46, 309–331.

Rao, C.R., 1948, Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation, Proceedings of the Cambridge Philosophical Society 44, 50–57.

Schwert, G. William, 1989, Business cycles, financial crises, and stock volatility, in: Karl Brunner and Allen H. Meltzer, eds., IMF policy advice, market volatility, commodity price rules, and other essays, Carnegie–Rochester conference series on public policy, Vol. 31 (North-Holland, Amsterdam) 83–125.

Schwert, G. William, 1990, Stock volatility and the crash of '87, Review of Financial Studies 3, 77–102.

Tauchen, George, 1985, Diagnostic testing and evaluation of maximum likelihood models, Journal of Econometrics 30, 415–443.

Turner, Christopher M., Richard Startz, and Charles R. Nelson, 1989, A Markov model of heteroskedasticity, risk, and learning in the stock market, Journal of Financial Economics 25, 3–22.

White, Halbert, 1982, Maximum likelihood estimation of misspecified models, Econometrica 50, 1–26.

White, Halbert, 1987, Specification testing in dynamic models, in: Truman F. Bewley, ed., Advances in econometrics, Fifth world congress, Vol. II, (Cambridge University Press, Cambridge).