USING THE PENALIZED LIKELIHOOD METHOD FOR MODEL SELECTION WITH NUISANCE PARAMETERS PRESENT ONLY UNDER THE ALTERNATIVE: AN APPLICATION TO SWITCHING REGRESSION MODELS

By Arie Preminger and David Wettstein

Ben-Gurion University of the Negev

First Version received December 2003

Abstract. We study the problem of model selection with nuisance parameters present only under the alternative. The common approach for testing in this case is to determine the true model through the use of some functionals over the nuisance parameters space. Since in such cases the distribution of these statistics is not known, critical values had to be approximated usually through computationally intensive simulations. Furthermore, the computed critical values are data and model dependent and hence cannot be tabulated. We address this problem by using the penalized likelihood method to choose the correct model. We start by viewing the likelihood ratio as a function of the unidentified parameters. By using the empirical process theory and the uniform law of the iterated logarithm (LIL) together with sufficient conditions on the penalty term, we derive the consistency properties of this method. Our approach generates a simple and consistent procedure for model selection. This methodology is presented in the context of switching regression models. We also provide some Monte Carlo simulations to analyze the finite sample performance of our procedure.

Keywords. Model selection; switching regression models; penalized likelihood method; law of the iterated logarithm.
JEL classification. C12; C32; C52.

1. INTRODUCTION

Hypothesis testing plays a crucial rule in any statistical analysis. A difficulty arises when the nuisance parameters are present only under the null hypothesis. This occurs, among others, in tests for threshold type nonlinearities, tests for structural breaks and in testing for the number of states in switching regression models. In such cases, regular statistical testing methods fail due to the 'flatness' of the likelihood function rendering the standard chi-square tests inapplicable.

Davies (1977, 1987) was one of the first to analyze the problem of unidentified nuisance parameters. His work suggests viewing the test statistic as a function of the unidentified parameters, so as to apply the empirical process theory. The weakness of Davies' test is that he did not derive the exact asymptotic distribution of his test statistic but used bounds, which may have a very low power in actual testing. Hansen (1996a) proposed a similar approach in the context of testing

0143-9782/05/05 715–741 JOURNAL OF TIME SERIES ANALYSIS Vol. 26, No. 5
© 2005 Blackwell Publishing Ltd., 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

nonlinear terms in regression models. His test statistic converges to a function of a chi-square process. The critical values of his statistic are not known and have to be approximated by computationally intensive simulations with complexity increasing in the dimension of the unidentified parameter set. Furthermore, the distribution of the test statistic depends on the covariance function of a chi-square process, as well as the domain of the unidentified parameters and the functional form of the statistic. Hence, the distribution is data and model dependent, which makes general tabulation impossible. Other aspects related to this testing problem concern the choice of the functional over the nuisance parameter space in order to obtain locally more powerful tests (see e.g. Andrews and Ploberger, 1994). The distributions of these tests are not known and have to be derived in the same manner as discussed above.

Another approach is to use the Monte Carlo method to simulate the distribution of the likelihood ratio. The idea is that given a sequence of observed data, we obtain the likelihood ratio by estimating the model under the null and the alternative hypotheses. We use our estimates of the model parameters under the null hypothesis, to generate independent samples and obtain the likelihood ratio empirical distribution by fitting the models, assumed under both hypotheses, to the simulated samples. The original likelihood ratio is compared to quantiles from the empirical distribution. Such methods were used by Feng and McCulloch (1996) for mixture models and by Lam (1990) for Markov switching models.

There are, however, a few drawbacks to the use of this method. First, as mentioned by Hansen (1992), there is no reason to assume that the finite sample distribution of the likelihood ratio will be invariant to the unidentified parameters under the null hypothesis. Second, Hansen (1992) and Hamilton (1990) claim that when the data are generated under the null hypothesis, the likelihood function is ill-behaved with many local maxima. This can lead to underestimation of the likelihood ratio, and the larger the parameter set, the more likely that the tabulated likelihood ratio distribution will be a lower bound for the true distribution. Third, the Monte Carlo simulations are a time consuming procedure and when one needs to compare a set of models, it becomes less and less appealing to use this method.

In this paper, we generalize the approach discussed in Nishii (1988) and Sin and White (1996) and use the penalized likelihood method for selecting the correct model among several competing models. In this approach to model selection, a term that acts to penalize for model complexity is added to the likelihood function used to estimate the parameters of the model. We then select the model that maximizes the penalized likelihood function. As pointed out in Granger *et al.* (1995), this method avoids some problems of traditional hypothesis testing, such as the direction of the hypothesis[1] and the arbitrary choice of significance levels. They also noted that the penalized likelihood method amounts to testing each model against all other models by means of the standard likelihood ratio test and selecting that model which is accepted against all other models, with the critical values determined by the penalty term.

© Blackwell Publishing Ltd 2005

While the method we use resolves the model selection problem in a variety of cases where the nuisance parameters are not identified under the null hypothesis, we choose, without loss of generality, to present it within the context of switching regression models. An important hypothesis that we will address in this paper is the validity of the model latent structure i.e. the number of states the state variable can assume.

Determining the number of states in a switching regression model is a difficult problem. One can perform a formal test of the null hypothesis in which a process with N − 1 states generates the data against the alternative that it came from an N-states model. Unfortunately, this hypothesis cannot be tested using the likelihood ratio test, with an asymptotic chi-square distribution, since under the null hypothesis the nuisance parameters that describe the Nth state are unidentified. That is, the likelihood function under the null hypothesis is nonquadratic and flat with respect to the nuisance parameter at the optimum, therefore the score function is identically zero at extreme points of the likelihood function and the information matrix is singular.

However, the difficulty associated with testing in these models is not just the problem of unidentified nuisance parameters under the null. The difficulty, as was pointed out by Andrews (1993), is that for mixture models or more generally switching regression models with constant probabilities, the singularity problem exists even if we fix the unidentified parameters. Jeffries (1998) showed that having the state probabilities random and change over time is sufficient in order to overcome the singularity problem, which validates the usage of the empirical process theory for testing. Therefore, we analyze in this paper switching regression models with time varying probabilities.

We specify general conditions under which the use of the penalized likelihood method will lead to selecting the correct model with probability one (strong consistency of selection), or with probability approaching one (weak consistency of selection) as the sample size increases. Thus, our conditions will guarantee that the correct number of states will be chosen.

This paper is organized as follows: in Section 2 we describe the general set-up and define the penalized likelihood statistic used for model selection. In Sections 3 and 4 we address the consistency issue. Weak consistency is established using the empirical process theory and strong consistency is established using the uniform law of the iterated logarithm (ULIL). To illustrate the small sample performance of our statistic, the findings of Monte Carlo simulations are reported in Section 5. Section 6 offers concluding remarks.

2. BASIC FRAMEWORK

We describe the general set-up and define the penalized likelihood statistic used for model selection. More specifically, the observed data is a realization of a stochastic process $\{Z_t : \Omega \to R^v,\ t = 1, 2, \ldots\}$ on a complete probability space

© Blackwell Publishing Ltd 2005

$(\Omega, \Im, P_0)$, where $\Omega = \times_{t=1}^{\infty} R^v$ and $\Im$ is the Borel σ-field generated by measurable finite dimensional product cylinders, and p_0 is the probability measure governing the behaviour of the data.

ASSUMPTION 1. *The random vectors $\{Z_t\}_{t\in N}$ are strictly stationary and ergodic and the true model is in the class of models being investigated.*

Let $\Im_t$ be the σ-field generated by current and past Z_t, i.e. $\Im_t = \sigma(\ldots, Z_{t-1}, Z_t)$, where $\Im_{t-1} \subset \Im_t \cdots \subset \Im$. The vector Z_t is partitioned into $Z_t = (Y_t, X_t)$ where Y_t is the dependent variable and X_t is the $1 \times \ell$ dimensional vector of explanatory variables with $v = 1 + \ell$. We are interested in a parametric family of conditional probability distributions indexed by $\psi \in \Psi$, a compact set and conditioned on $\bar{\Im}_{t-1} \equiv \sigma(Z_{t-\tau}, \ldots, Z_{t-1}, X_t)$, $\tau < \infty$, which is given by $\{P_t^{\psi}(y_t|\bar{\Im}_{t-1}; \psi), \psi \in \Psi, \bar{\Im}_{t-1} \subset \Im_t\}$. These measures exist by Jirina's theorem (Bauer, 1972, p. 319) and our parametric models include explicitly a finite number of lags. We also assume that these conditional distributions have Radon–Nikodym densities with respect to the usual Lebesgue measure that is $\tilde{f}(y_t|w_t, \psi) \equiv \mathrm{d}P_t^{\psi}(y_t|w_t; \psi)/\mathrm{d}\upsilon$, where w_t denotes the variables that the analyst has chosen to explain or forecast y_t (these might include X_t and lagged values of the dependent variables).

We assume that the true data generating process (DGP) is an autoregressive switching regression (SR) model. In this model, in each period, there exists a model selection procedure, which picks a specific parametric model. More specifically, in each point in time the unobserved selection process picks a parameter vector from the set $\{\psi_1, \ldots, \psi_k\}$. The selection is random and dependent on the realization of a latent state variable, s_t, which can assume only an integer value $\{1, \ldots, k\}$. The state (variable) probabilities are not constant and are denoted by $\Pr(s_t = i|\tilde{z}_t, \gamma)$, where the vector $\tilde{z}_t$ contains explanatory variables that affect the state probabilities and are made up of known measurable functions of w_t. $\gamma \in \Gamma$ is the parameter vector, where the set Γ is compact and restricted to allow only time varying probabilities and these functions satisfy standard measurability and continuity requirements[2] on $\tilde{z}_t \times \Gamma$.

The conditional density of y_t can be described by

$$f(y_t|w_t, \theta) = \sum_{i=1}^{k} Pr(s_t = i|\tilde{z}_t, \gamma) \cdot \tilde{f}(y_t|w_t, \psi_i) \tag{1}$$

$\theta = (\psi_1, \ldots, \psi_k, \gamma) \in \Theta \subset R^L$ is the vector of the model parameters. We define the likelihood function of the sample as

$$L_T(\theta) = \sum_{t=1}^{T} \log f(y_t|w_t; \theta) \tag{2}$$

Next, we define the true number of states by $k_0 \in \{1, \ldots, k_{\max}\}$. When $k < k_0$ the true model is not nested in the alternative models and we minimize the

© Blackwell Publishing Ltd 2005

Kullback–Leibler information criterion (KLIC) (see e.g. White, 1982, 1994; Vuong, 1989). When $k_0 < k \leq k_{\max}$ the true model is nested in the alternative models and some of the model parameters are not identified. In this case, we divide the set of parameters into two disjoint sets $\theta(k) = (\theta_1(k), \theta_2(k)) \in \Theta_1(k) \times \Theta_2(k) = \Theta(k)$. Let $\theta_1(k) \in \Theta_1(k) \subset R^{L_1}$ be the parameters which are not identified, $\theta_2(k) \in \Theta_2(k) \subset R^{L_2}$ are the other parameters ($L_1 + L_2 = L$), where $\Theta_1(k)$, $\Theta_2(k)$ are compact sets. We note that in this case, the score function of these models is a martingale difference, since we assume the true model belongs to the set of models being considered.

For example, suppose the true model is $f_1 = f(y_t|w_t, \beta_1)$, that is $k = 1$, and we estimate a simple mixture of two parametric models $f_1 = f(y_t|w_t, \beta_1)$ and $f_2 = f(y_t|w_t, \beta_2)$ with probability π and $1 - \pi$ respectively, assuming that $k = 2$. In this case, some of the model parameters are not identifiable and this means that f_1 has different representations with different parameters:

$$\pi \cdot f_1 + (1 - \pi) \cdot f_1 = 1 \cdot f_1 + 0 \cdot f_2. \tag{3}$$

We see that under the restriction of no mixture, π might converge to one in which case β_2 can assume any value or, in another scenario, β_2 might converge to β_1 and the probability is not identifiable.

For a given k we consider a model with parameter space $\Theta(k)$ and quasi-likelihood functions $\{f(y_t|w_t; \theta(k)) : 1 \leq k \leq k_{\max}, \theta(k) \in \Theta(k), t = 1, 2, \ldots\}$.

The likelihood function is

$$L_T(\theta(k)) = \sum_{t=1}^{T} \log f(y_t|w_t; \theta(k)). \tag{4}$$

The maximum likelihood statistic is

$$Q_{T,k} = \max_{\theta(k) \in \Theta(k)} L_T(\theta(k)) = L_T(\hat{\theta}_T(k)) \tag{5}$$

where $\hat{\theta}(k)$ is the maximum likelihood estimator. The penalized likelihood function is defined by subtracting a penalty term from the maximum likelihood statistic, yielding what is called information criteria:

$$IC_T(k) = Q_{T,k} - c_{T,k}. \tag{6}$$

The penalized likelihood statistic $\hat{k}$ which estimates the true number of states k_0 is the maximum of the penalized likelihood function:

$$\hat{k} = \arg\max_{k \in \{1, \ldots, k_{\max}\}} (IC_T(k)). \tag{7}$$

This process of model selection is justified by the need to balance the increase in fit (more parameters yield a higher likelihood) obtained against the larger number of parameters estimated for models with more state variables. Basically, such criteria impose a penalty on the likelihood function that is related to the number of parameters estimated. The usage of penalized likelihood statistics for model selection was first introduced by Akaike (1974), who defined the Akaike

© Blackwell Publishing Ltd 2005

information criterion (AIC) in which the penalty term is equal to twice the number of additional parameters estimated in the bigger model. Another popular criterion, which imposes an additional penalty related to sample size, is the Bayesian information criterion (BIC), developed by Schwarz (1978). The BIC is defined as the maximized likelihood plus a penalty term, which is the logarithm of the number of observations, multiplied by the number of additional parameters.

3. WEAK CONSISTENCY OF THE PENALIZED LIKELIHOOD STATISTIC

For what follows, we need the following assumptions.

ASSUMPTION 2. *For all* $k \in \{1, \ldots, k_{\max}\}$ *the functions* $f(y_t|w_t; \theta(k))$ *are positive and measurable* $\sigma(y_t, w_t) \subset \Im_t$ *for every* $\theta(k)$ *in* $\Theta(k) \subset R^L$ *a compact set, and are continuous on* $\Theta(k)$ *for each* (y_t, w_t) *a.s.* p_0 *for all* t.

ASSUMPTION 3. *For all* $1 \leq k \leq k_0$, $E(\log f(y_t|w_t; \theta(k)))$ *has a unique maximum at* $\theta^*(k)$ *an interior point of* $\Theta(k)$.

ASSUMPTION 4. *For all* $k \in \{1, \ldots, k_{\max}\}$, $|\log f(y_t|w_t; \theta(k))| < m(y_t, w_t)$ *for all* $\theta(k) \in \Theta(k)$ *and for each* (y_t, w_t) *a.s.* p_0, *and* $E(m(y_t, w_t)) < \Delta < \infty$.

ASSUMPTION 5. *The penalty term satisfies* $c_{T,\tilde{k}} > c_{T,k} > 0$ *for* $\tilde{k} > k$, $\lim_{T\to\infty} c_{T,k} = +\infty$, $c_{T,k} = o(T)$.

Assumptions 1–3 are needed to establish the existence of a measurable quasi-maximum likelihood estimate, which is uniquely identifiable. Assumption 4 imposes a moment condition, by assuming the existence of a data-dependent upper bound on $\log f(y_t|w_t, \theta(k))$ that has a finite expectation and hence allows us to apply the uniform strong law of large numbers (USLLN) for stationary and ergodic processes (Rao, 1962). For the general case, this assumption can be replaced by a 'high-level' assumption that the likelihood function obeys the USLLN.

In the SR models the true parameter set is not identifiable due to 'label switching', i.e. the parameter set for which the likelihood function has the same value, is not a singleton set. Therefore, in order to simplify the discussion as well as the notations and the assumptions made in this work, we establish the existence and consistency of the penalized likelihood statistic in the quotient topology. That is, we merge all the parameter values which define the same (penalized) likelihood function into a single equivalence class (see Leroux, 1992; and Redner, 1981 for the case of mixture distributions) and by letting Θ be the set of such equivalence

© Blackwell Publishing Ltd 2005

classes we ensure that θ^* is identifiably unique. The topology over the parameter values is translated into a topology over equivalence classes, known as the quotient topology. We will not concern ourselves further with these details.

Lemma 1 establishes that asymptotically the estimator $\hat{k}$ does not underestimate the number of states almost surely.

LEMMA 1. *Given Assumptions* 1–5 $\hat{k} \geq k_0$ *almost surely*.

PROOF. See Appendix. □

Lemma 1 is also used to show strong consistency of selection in the next section. However, so as to obtain the weak consistency of the penalized likelihood, it is sufficient to show that $\hat{k}$ provides an upper bound in probability. Hence, we can weaken Assumptions 4 and 5 and require that the likelihood function satisfies the uniform weak law of large numbers (UWLLN) and that the penalty term satisfies $c_{T,k} = o_p(T)$, $c_{T,k} \xrightarrow{p} \infty$. When the observations are dependent and heterogeneous and the likelihood function is sufficiently smooth, we can establish the UWLLN by using Andrews' (1992) results. In addition, other standard results (Gallant and White, 1988, Thm 3.18) allow us to establish the USLLN for the likelihood function under this case.

In order to prove that $\hat{k}$ is weakly consistent, we will also have to establish that it cannot overestimate the true number of states. Given Assumption 5, it is sufficient to show that for any $k > k_0$ the likelihood ratio (LR) has an asymptotic distribution. Sin and White (1996) show this by relying on the assumption that the score function satisfies pointwise the central limit theorem (CLT). However, when $k > k_0$ some of the model parameters are not identified and the general theory developed by Sin and White (1996) does not apply. Therefore, we apply the empirical process theory and provide a set of sufficient conditions which imply that the score function converges uniformly over the set of unidentified parameters to a Gaussian process and since the LR is a continuous functional over this process, it has an asymptotic distribution. This approach was used by Hansen (1992) to test the number of regimes in a Markov switching model proposed by Hamilton (1989), see also Andrews (1994), for a description of the econometric applications of empirical process theory.

Next, we need to introduce more notations and assumptions.[3]

Let $L_T(\theta_1(k), \theta_2(k)) = \sum_{t=1}^{T} \log f(y_t|w_t; \theta_1(k), \theta_2(k))$ denote the likelihood function given the parameters $\theta_1(k)$, $\theta_2(k)$ where we assume that the likelihood is twice continuously differentiable in $\theta_2(k)$ in the interior of $\Theta_1(k) \times \Theta_2(k)$, and $D_t(\theta_1(k), \theta_2(k))$ denotes the L_2-vector of partial derivatives of $\log f(y_t|w_t;\theta_1(k), \theta_2(k))$ with respect to $\theta_2(k)$, and $D_t^2(\theta_1(k), \theta_2(k))$denotes the $L_2 \times L_2$ matrix of second partial derivatives with respect to $\theta_2(k)$. Let $D_T(\theta_1(k), \theta_2(k)) = \sum_{t=1}^{T} D_t(\theta_1(k), \theta_2(k))$, $D_T^2(\theta_1(k), \theta_2(k)) = \sum_{t=1}^{T} D_t^2(\theta_1(k), \theta_2(k))$ and $\hat{\theta}_{2T}(k, \theta_1)$ be the maximum likelihood estimator of $\theta_2(k)$ for a fixed $\theta_1(k) \in \Theta_1(k)$, i.e. the estimator satisfies:

© Blackwell Publishing Ltd 2005

$$L_T(\theta_1(k), \hat{\theta}_{2T}(k, \theta_1)) = \sup_{\theta_2(k)\in\Theta_2(k)} L_T(\theta_1(k), \theta_2(k)) \quad \text{for} \quad \theta_1(k) \in \Theta_1(k). \quad (8)$$

We denote by $\theta_2^*(k)$ the true value of $\theta_2(k)$ in which some of the model parameters are not identified for $k > k_0$, i.e. $L_T(\theta_1(k), \theta_2^*(k)) = L_T(\theta^*(k_0))$ for all $\theta_1(k)$. Note that the likelihood function does not depend on the nuisance parameters under the true number of states. Therefore, when $k = k_0$ Assumptions 3′ and 8 are not relevant since $\Theta_1(k)$ is empty and Assumptions 6 and 7 refer only to $\theta_2(k) \in \Theta_2(k) \equiv \Theta(k_0)$.

The LR statistic is defined as

$$LR_T = Q_{T,k} - Q_{T,k_0} = \sup_{\theta_1(k)\in\Theta_1(k)} L_T(\theta_1(k), \hat{\theta}_{2T}(k, \theta_1)) - L_T(\hat{\theta}_T(k_0)) \quad (9)$$

Assumption 3′.

(a) *For every neighbourhood* $\bar{\Theta}_2(k) \subset \Theta_2(k)$ *of* $\theta_2^*(k)$,

$$\lim_{T\to\infty} \inf_{\theta_1(k)\in\Theta_1(k)} \big(E(\log f(y_t \mid w_t; \theta_1(k), \theta_2^*(k)) - \max_{\theta_2\in\Theta_2(k)\setminus\bar{\Theta}_2(k)} E\big(\log f(y_t \mid w_t; \theta_1(k), \theta_2(k))\big) > 0$$

(b) $\theta_2^*(k)$ *is an interior point of* $\Theta_2(k)$.

Assumption 6.

(a) $L_T(\theta_1(k), \theta_2(k))$ *is twice continuously partially differentiable in* $\theta_2(k)$ *for all* $\theta_2(k) \in \Theta_2(k)$ *and all* $\theta_1(k) \in \Theta_1(k)$.

(b) *The elements of* $|D_t(\theta_1(k), \theta_2(k)) \cdot D_t(\theta_1(k), \theta_2(k))'|$ *are dominated by* p_0*-integrable functions independent of* $\theta_2(k)$ *for all* $\theta_1(k) \in \Theta_1(k)$.

(c) *For all* $\theta_1(k) \in \Theta_1(k)$ *the elements of* $|\partial f(y_t|w_t, \theta_1(k), \theta_2(k))/\partial\theta_2(k)|$ *and* $|\partial^2 f(y_t|w_t, \theta_1(k), \theta_2(k))/\partial\theta_2(k) \cdot \partial\theta_2'(k)|$ *are dominated by* p_0*-integrable functions independent of* $\theta_2(k)$.

Assumption 7.

(a) $\frac{1}{T} D_T^2(\theta_1(k), \theta_2(k)) \xrightarrow{\text{a.s.}} E\big(D_t^2(\theta_1(k), \theta_2(k))\big)$, *uniformly over* $\theta_2(k) \in \Theta_2(k)$ *and* $\theta_1(k) \in \Theta_1(k)$.

(b) *The matrix* $E\big(D_t^2(\theta_1(k), \theta_2(k))\big)$ *is invertible, positive definite and continuous in* $(\theta_1(k), \theta_2(k))$ *uniformly over* $\Theta_1(k) \times \Theta_2(k)$.

Assumption 8. $\frac{1}{\sqrt{T}} D_T(\theta_1(k), \theta_2^*(k))$ *is asymptotically stochastically equicontinuous.*

Assumption 3′ modified Assumption 3 to allow for uniform identifiability on $\theta_1(k)$ and enables us to show in Lemma 2 uniform strong consistency of $\theta_2(k)$ over the set of the unidentified parameters. This result is useful in establishing that the

© Blackwell Publishing Ltd 2005

likelihood ratio is uniformly bounded in probability. Assumptions 6 and 7 are standard assumptions that impose moments and smoothness conditions that are commonly used. Condition 6(c) allows us to change the order of the integration and differentiation operator and to show that the gradient of the likelihood function is a martingale difference (under dynamic model misspecification this assumption can be omitted). This result underlines Condition 6(b), which is needed to apply the CLT for stationary, ergodic, martingale difference processes. For the general case, Assumption 6 can be replaced by a high-level assumption that $\frac{1}{\sqrt{T}} D_T(\theta_1(k), \theta_2^*(k))$ converges in distribution for all $\theta_1(k) \in \Theta_1(k)$. Under model dynamic misspecification we will have to impose stronger conditions on the memory of the process and strengthen our moment requirements. We can consider processes near epoch dependent (NED) on a strong mixing process[4] and provide sufficient conditions based on the CLT of Wooldridge (1986). Kapetanios (2001) and Davidson (2002) showed that under mild conditions the linear SR models are NED of any size. Condition 7(a) can be verified by applying the USLLN for stationary and ergodic sequences. However, this condition can be weakened similar to Assumption 4 in order to show the weak consistency of selection. Note that Assumption 7(b) does not hold even for a fixed value of $\theta_1(k)$ in switching regression models with constant probabilities as was mentioned above. To establish Assumption 8, we can modify Theorem 5 of Hansen (1996b) to hold under Assumption 1, see our example in Section 5. Hansen's results are particularly suited for Lipschitz smooth functions of the unidentified parameters, $\theta_1(k)$. In a similar way this assumption can be verified for NED processes. Further, the stochastic equicontinuity property is obtained under more general types of functions see e.g. the empirical process theory developed by Andrews (1993) and Andrews and Pollard (1994) and numerous references cited therein. These assumptions are used to prove the following lemmata.

LEMMA 2. *Given Assumptions* 1, 2, 3′ *and* 4 *for* $k > k_0, \hat{\theta}_{2T}(k, \theta_1) \to \theta_2^*(k)$ *almost surely*, *uniformly over* $\Theta_1(k)$.

PROOF. See Appendix. □

We will use Lemma 2, a Taylor expansion of the likelihood function in the neighbourhood of the identified parameters and the stochastic equicontinuity property of the normalized score function to show that the likelihood ratio is bounded in probability.

LEMMA 3. *Given Assumptions* 1–8, *for* $k > k_0$, $LR_T = O_P(1)$.

PROOF. See Appendix. □

Given the results of the lemmata above and the conditions on the penalty term in Assumption 5, Theorem 1 establishes the weak consistency (convergence in probability) of the penalized likelihood statistic. This assumption is similar to Sin

© Blackwell Publishing Ltd 2005

and White's (1996) conditions on the penalty term for strictly nested models and ensures us that, given Lemma 3, when $k > k_0$ the likelihood ratio will be dominated by the penalty terms in probability and hence, we will pick the most parsimonious model. For instance, we can choose $c_{T,k} = 0.5 \cdot \dim(\Theta(k)) \cdot (\log(T))^b$ for some $b > 0$, note that for $b = 1$, we use the Bayesian information criterion.

THEOREM 1. *Given Assumptions* 1, 2, 3′, *and* 4–8, $\hat{k}$ *converges to* k_0 *in probability*.

PROOF. See Appendix. □

4. ALMOST SURE CONSISTENCY OF THE PENALIZED LIKELIHOOD STATISTIC

We prove the strong consistency of the penalized maximum likelihood statistic. We start by discussing additional sufficient conditions which guarantee the selection of the true number of states, with probability one. By Lemma 1 we know that $\hat{k}$ does not asymptotically underestimate the value of k_0. Therefore, it remains to prove that $\hat{k}$ does not asymptotically overestimate this value. We will use the ULIL when $k > k_0$. The definition of the ULIL which will be used in this work is as follows.

DEFINITION. $\{u_t\colon \Omega \times \Phi \to R\}$ *is said to satisfy a ULIL on* $\Gamma \subseteq \Phi$ *if* $\sigma^2(\gamma) = \lim_{T\to\infty} \frac{1}{T} \operatorname{var}(\sum_{t=1}^{T} u_t(\gamma))$ *exists and is strictly positive and*

(a) $\limsup_{T\to\infty} \left| \sum_{t=1}^{T} \frac{(u_t(\gamma) - E(u_t(\gamma)))}{\sigma(\gamma)\sqrt{2T\log\log(T)}} \right| = 1$ *almost surely for all* $\gamma \in \Gamma$.

(b) $\left\{ \frac{1}{\sqrt{2T\log\log(T)}} \sum_{t=1}^{T} u_t(\gamma) \right\}_{t\in\mathbf{N}}$ *is strongly stochastic equicontinuous on* Γ.

In the context of model selection, the law of the iterated logarithm (LIL) was first applied by Nishii (1988). He showed that for nested but possibly misspecified models of i.i.d. processes, the use of the penalized likelihood statistic leads to the selection of the model with the lowest Kullback–Leibler (1951) divergence from the true data generating process as the sample size increases. Sin and White (1996) generalized Nishii's (1988) results to dependent and heterogeneous processes. In our case, we apply the ULIL on the score function in order to provide almost surely bounds for the likelihood ratio on $\Theta_1(k)$, e.g. over the set of parameters which is not identified when $k > k_0$, and use this result to obtain the strong consistency of our statistic. Therefore, we add Assumptions 9 and 10.

© Blackwell Publishing Ltd 2005

ASSUMPTION 9. *For $k > k_0$ and all $\theta_1(k)$ the elements of $\{D_t(\theta_1(k), \theta_2^*(k))\}$ satisfy the ULIL on $\Theta_1(k)$ and for $k = k_0$ the elements of $\{D_t(\theta^*(k))\}$ satisfy the LIL.*

ASSUMPTION 10. *The penalty term also satisfies*

$$j \cdot (\dim(\Theta_2(k)) - \dim(\Theta(k_0))) = \left(\frac{c_{T,k} - c_{T,k_0}}{\log\log(T)}\right), \quad j > 1.$$

Assumption 9 can be verified by first providing bounds for the score function via the LIL for stationary, ergodic, martingale difference sequences for all $\theta_1(k) \in \Theta_1(k)$. For example, we can use the results by Stout (1970) and Assumptions 1 and 6 to show that for each value of the nuisance parameters, the score function satisfies the LIL. Under model dynamic misspecification, we need to impose the additional condition that $\{D_t(\theta_1(k), \theta_2^*(k)), t = 1, 2, \ldots\}$ is NED of size $-1/2$ on a mixing process of appropriate size in order to ensure the same pointwise convergence. This follows from Corollary AIII.3 of Sin and White (1992) and Theorem 17.5 of Davidson (1994). The strong stochastic equicontinuity requirement will be met if we assume that the score function is differentiable almost surely at each point of $\Theta_1(k)$, and that the gradient vector of the score function with respect to $\theta_1(k)$ can be bounded uniformly over $\Theta_1(k)$ by the LIL (see Andrews, 1992; Altissimo and Corradi, 2002).

LEMMA 4. *Given Assumptions* 1–3, 3′, 4, 6, 7 *and* 9, *for* $k > k_0$, $\limsup_{T\to\infty} \frac{LR_T}{\log\log(T)} \leq \dim(\Theta_2(k)) - \dim(\Theta(k_0))$.

PROOF. See Appendix. □

Theorem 2 follows from Lemmata 1 and 4 and sufficient conditions on the penalty term. Assumption 10 ensures that for $k > k_0$ the LR is dominated by the penalty term with probability one and hence, guarantees that we will choose the most parsimonious model (the model with k_0 states). This assumption relaxes Nishii's (1988) conditions on $c_{T,k}$ by providing sharper bounds on the penalty term and modifies Sin and White's (1996) conditions; since the penalty term does not depend on the total number of the model parameters, and takes into account only the dimension of the identified parameter space.

THEOREM 2. *Given Assumptions* 1–3, 3′, 4–7, 9 *and* 10, $\hat{k}$ *converges to* k_0 *almost surely*.

PROOF. See Appendix. □

© Blackwell Publishing Ltd 2005

5. SIMULATION STUDY

We consider an autoregressive SR model with two states and time varying state probabilities, which we assume is the true data generating process. This model is widely used for modelling nonlinear time series with non-Gaussian features such as outliers, flat stretches and change points (Li and Wong, 2001; Lanne and Saikkonen, 2003). We verify that the key assumptions stated in Theorems 1 and 2 are satisfied in this case. We also examine the performance of the penalized likelihood statistic in small samples and its sensitivity to different penalty terms. The SR model is

$$y_t = \alpha_t + \beta_t y_{t-1} + \varepsilon_t \tag{10}$$

where $\varepsilon_t \sim$ i.i.N$(0, \sigma^2)$ and $s_t \in \{1, 2\}$ are unobserved independent state variables, which determine the value of coefficients α_t, β_t. By this we mean:

$$y_t|y_{t-1} \sim \begin{cases} f(y_t|y_{t-1}, \alpha_1, \beta_1) = f_1(\cdot) & \text{if } s_t = 1 \\ f(y_t|y_{t-1}, \alpha_2, \beta_2) = f_2(\cdot) & \text{if } s_t = 2 \end{cases}. \tag{11}$$

$f(y_t|y_{t-1}, \alpha_t, \beta_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\lfloor -(y_t - \alpha_t - \beta_t y_{t-1})^2 / 2\sigma^2 \rfloor$.

The state probabilities are not constant over time and are given by a logistic function.

$$p_1(\cdot) = Pr(s_t = 1|y_{t-1}; \gamma_1, \gamma_2) = \Lambda(H_{t-1}) = \frac{\exp(H_{t-1})}{1 + \exp(H_{t-1})} \tag{12}$$

where $H_{t-1} = \gamma_1 + \gamma_2 y_{t-1}$ and $[\gamma_1, \gamma_2]$ are unknown model parameters. For the true model, $k_0 = 2$ and for the set of models being considered $k \in \{1, 2, 3\}$. We will apply the penalized likelihood method to estimate the true model from this set. If $k = 1$ the model is a linear regression model and for $k = 3$ the model parameters can assume one of three values given the realization of the state variables. The state probabilities are given as follows:

$$\begin{aligned} p_1(\cdot) &= \Pr(s_t = 1|y_{t-1}; \gamma_1, \gamma_2) = \Lambda(H_{t-1}) \\ p_2(\cdot) &= \Pr(s_t = 2|y_{t-1}; \gamma_1, \gamma_2) = \Lambda(\mu + H_{t-1}) - \Lambda(H_{t-1}), \end{aligned} \qquad i = 1, 2, \quad \mu > 0. \tag{13}$$

The density function is given by

$$g(y_t|y_{t-1}, \theta(3)) = \sum_{i=1}^{3} Pr(s_t = i|y_{t-1}, \gamma_1, \gamma_2, \mu) \cdot f(\cdot)_i = \sum_{i=1}^{3} p_i(\cdot) \cdot f_i(\cdot) \tag{14}$$

where $f_i(\cdot) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_t - \alpha_i - \beta_i y_{t-1})^2}{2\sigma^2} \right)$, $\quad i = 1, 2, 3.$

We now proceed to show that the environment we analyze in this section satisfies our key assumptions. We show that the SR process is strictly stationary and ergodic, thus Assumption 1 is satisfied. In this model the normality of the error term also implies that the SR process has finite moments of any order. Using this result we derive Assumption 4 (for $k \leq 2$) that the likelihood function is

© Blackwell Publishing Ltd 2005

bounded by integrable function. Furthermore, based on Hansen's (1996b) results for martingale difference processes; we show that Assumption 8 can be verified. Finally, applying Stout's (1970) results, we show that the score function satisfies the ULIL (Assumption 9). The other Assumptions are technical in nature and we directly assume that they are valid in the switching regression setup.

To show the stationarity and ergodicity of $\{y_t\}$ (Assumption 1) we use results derived in the theory of continuous state Markov chains (Chan and Tong, 1985; Tjostheim, 1990; Chan, 1993). Let $\{y_t\}$ be a time homogenous Markov chain on $(\Omega, \Im)$ with transition kernel

$$P(y_t \in A | y_{t-1} = x) = \int_A g(y_t | y_{t-1} = x) \tag{15}$$

where $x \in \Re$, $A \in \Im$ and

$$g(y_t | y_{t-1}, \theta(2)) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{2} p_i(s_t = i | y_{t-1}, \gamma_1, \gamma_2) \exp\left(\frac{(y_t - \alpha_i - \beta_i y_{t-1})^2}{2\sigma^2} \right). \tag{16}$$

We assume the true parameters $\theta^*(2) = [\gamma_1^*, \gamma_2^*, \alpha_1^*, \alpha_2^*, \beta_1^*, \beta_2^*, \sigma^*]$, lie in a compact space given by $\alpha_i \in [M_1, M_2]$, $\gamma_i \in [M_2, M_3]$, $0 < \sigma_{\min} < \sigma < \sigma_{\max} < \infty$ and $\beta_1, \beta_2 \in [-1 + \varepsilon, 1 - \varepsilon]$, where ε denotes an arbitrary small positive number and the M_j's are arbitrary constants. The normality of the error term implies that $\{y_t\}$ is λ-irreducible (λ is the Lebesgue measure) and aperiodic (Chan, 1993). Thus, every nonnull set is reached in a finite number of steps with positive probability and it is impossible for the chain to return to a given set only at specific time points. In addition, we note that every λ-nonnull compact set is small (Chan and Tong, 1985, pp. 667–669). In order to derive geometric ergodicity we use Theorem 3.

THEOREM 3 (TJOSTHEIM, 1990). *Let $\{y_t\}$ be a time homogenous Markov chain on $(\Omega, \Im)$ which is irreducible, aperiodic and for which any compact set is small. If there exists a drift function V: $\Omega \to [1, \infty]$, a compact set $C \subset \Omega$ and constants $\eta \in (0, 1)$, $a > 0$, such that $E(V(y_{t+m}) | y_{t-1} = x) \leq \eta \cdot V(x) + a \cdot 1_C(x)$ for all $x \in \Omega$ and some $m \geqslant 0$. Then the process is geometrically ergodic and there exists an invariant measure for the process. If y_t is initiated at the invariant distribution, then the Markov chain is stationary and ergodic, and $E(V(y_t)) < \infty$.*

To apply Theorem 3 we show the existence of such a drift function. As mentioned, y_t is a Markov chain with a one-step transition density given in Eqn (15). We consider the function $V(x) = 1 + x^2$ and a compact set $C = \{x \in \Re \mid |x| \leq c\}$ for some $c < \infty$. Then we have

$$\begin{aligned} E(1 + y_t^2 | y_{t-1} = x) &= p_i(\cdot) \cdot \beta_1^2 x^2 + (1 - p_i(\cdot)) \cdot \beta_2^2 x^2 + \sigma^2 + E(\alpha_t^2) \\ &\quad + 1 \leq \eta(H_{t-1}) x^2 + 1 + \sigma^2 + E(\alpha_t^2) \end{aligned} \tag{17}$$

© Blackwell Publishing Ltd 2005

where $\eta(H_{t-1}) = \frac{\beta_2^2+\beta_1^2\exp(H_{t-1})}{1+\exp(H_{t-1})}$. Since $|\beta_1|, |\beta_2| < 1$ there exists $\eta < 1$ and c large enough, such that $\eta(H_{t-1}) < \eta$ and $E(V(y_t)|y_{t-1} = x) \leq \eta V(x)$ for $x \notin C$. In addition, because the second and the third terms are bounded, we can choose some $a < \infty$ such that $E(V(y_t)|y_{t-1} = x) \leq \eta V(x) + a \cdot 1_C(x))$ for $x \in C$. Hence $\{y_t\}$ is ergodic and asymptotically stationary. Assumption 1 is then satisfied if $\{y_t\}$ is started either with the invariant distribution or in the infinite past.

COROLLARY. *Let $\{y_t\}$ be a stationary switching regression process as described above then $E(y_t^d) < \infty \Leftrightarrow E(\varepsilon_t^d) < \infty$ and for a Gaussian white noise the process is d-integrable for all $d < \infty$.*

Next, we show Assumption 4 for $k = 2$. That is $\log g(y_t|y_{t-1}, \theta(2))$ can be bounded by an integrable function. Since $0 < \sigma_{\min} < \sigma$, the density function is bounded from above and it is sufficient to check the integrability of $\log(p_1(\cdot)f_1(\cdot)) = \log p_1(\cdot) + \log f_1(\cdot)$. From Eqn (11), we see that

$$|\log f_1(\cdot)| < \frac{(|y_t| + |y_{t-1}| + |\alpha_{\max}|)^2}{2\sigma_{\min}} + C_1$$

$$|\log p_1(\cdot)| = |\log(\exp(\gamma_1 + \gamma_2 y_{t-1})) - \log(1 + \exp(\gamma_1 + \gamma_2 y_{t-1}))| < 2|\gamma_1 + \gamma_2 y_{t-1}| + \log 2 < 2 \cdot C_2 + \log 2 + 2 \cdot C_2|y_{t-1}|$$

where C_1, C_2 are some positive constants, and the RHS is clearly integrable from the corollary. For $k = 1$ we obtain Assumption 4 in a similar way.

In the case $k = 3$, the density function is given by Eqn (14) and we need to verify a different set of assumptions. In this case the estimated model is equivalent to the true model for the following parameters:

$$\theta_2^*(3) = [\gamma_1^*, \gamma_2^*, \sigma^*, \alpha_1^*, \beta_1^*, \alpha_2^*, \beta_2^*, \alpha_2^*, \beta_2^*] \in \Theta_2(3) \text{ and } \theta_1(3) \in \Theta_1(3) = \{\mu \in \Re | \mu > 0\}.$$

Let $D_t^j(y_{t-1}, \theta_1(3), \theta_2^*(3)) = \frac{\partial \log g(y_t|y_{t-1};\theta_1(3),\theta_2^*(3))}{\partial \theta_{2j}(3)}$ be the j-element of the gradient of the likelihood function. From Assumption 1 and because our model is correctly specified, $\{D_t^j(y_{t-1}, \theta_1(3), \theta_2^*(3))\}_{t\in\mathrm{N}}$ is a martingale difference, stationary and ergodic sequence, hence we can use Theorem 4 to show that Assumption 8 is satisfied.

THEOREM 4 (HANSEN, 1996B). *Let $\{D_t^j(y_{t-1}, \theta_1(3), \theta_2^*(3)) : \Re \times \Theta_1(3) \to \Re\}_{t\in\mathrm{N}}$ be a parametric class of random functions which satisfy for $j = 1, 2, \ldots, 9$ that*

(1) *$|D_t^j(y_{t-1}, \theta_1^1(3), \theta_2^*(3)) - D_t^j(y_{t-1}, \theta_1^2(3), \theta_2^*(3))| \leq \mathrm{B}^j(y_{t-1}) \cdot \|\theta_1^1(3) - \theta_1^2(3)\|$ for some function $\mathrm{B}^j(\cdot):\Re \to \Re$ and all $\theta_1^1(3), \theta_1^2(3) \in \Theta_1(3)$.*
(2) *$||\mathrm{B}^j(y_{t-1})||_2 < \infty$ and $\|D_t(y_{t-1}, \theta_1(3), \theta_2^*(3))\|_2 < \infty$ for all $\theta_1(3) \in \Theta_1(3)$.*
(3) *$\{D_t^j(y_{t-1}, \theta_1(3), \theta_2^*(3))\}_{t\in\mathrm{N}}$ is a martingale difference, stationary and ergodic sequence.*

© Blackwell Publishing Ltd 2005

Then, $\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} D_t(y_{t-1}, \theta_1(3), \theta_2^*(3))\}_{t\in\mathrm{N}}$ *is asymptotically stochastically equicontinuous.*

To apply Theorem 4 we show first that $\|D_t^j(y_{t-1}, \theta_1(3), \theta_2^*(3))\|_2 < \infty$, though there are several derivatives to check, we present detailed analysis for only one; the others follow a similar pattern. For example we consider

$$\left|\frac{\partial \log g(y_t|y_{t-1}, \theta_1(3), \theta_2^*(3))}{\partial \beta_1}\right| = \left|\frac{p_1(\cdot)\cdot f_1(\cdot)\cdot \varepsilon_{1t}\cdot y_{t-1}}{g(\cdot)}\right| < y_{t-1}^2 + |y_t y_{t-1}| + |\alpha_{\max}|\cdot|y_{t-1}|. \tag{18}$$

By the corollary the right hand side is clearly r dominated[5] for all r. Since the score function is almost surely differentiable on the set of unidentified parameters, we can use the mean value theorem to show that it satisfies the Lipschitz condition mentioned above for $\mathrm{B}(y_{t-1}) = \sup_{\theta(3)\in\Theta_1(3)}\left|\frac{\partial D_t^j(\cdot)}{\partial \mu}\right|$. Also, we have that

$$\begin{aligned}
\mathrm{B}(y_{t-1}) &= \left|\frac{\partial \log g(y_t|y_{t-1}, \theta_1(3), \theta_2^*(3))}{\partial \beta_1 \partial \mu}\right| \\
&< \left|1 - p_1(\cdot) + \frac{p_1(\cdot)(1-p_1(\cdot))\cdot(f_1(\cdot) - f_2(\cdot))}{g(\cdot)}\right| \cdot \left|\frac{p_1(\cdot)\cdot f_1(\cdot)\cdot \varepsilon_{1t}\cdot y_{t-1}}{g(\cdot)}\right| \\
&\leq |2\varepsilon_{1t}\cdot y_{t-1}| \leq y_{t-1}^2 + |y_t y_{t-1}| + |\alpha_{\max}|\cdot|y_{t-1}|
\end{aligned} \tag{19}$$

where the RHS can be bounded in a similar manner, therefore $\|\mathrm{B}(y_{t-1})\|_2 < \infty$. As the other derivatives are handled in the same way we see that Assumption 8 is satisfied. Hence the penalized likelihood statistic in which the penalty term satisfies Assumption 5 is weakly consistent by the first theorem.

In order to show strong consistency of selection, we verify Assumption 9 (ULIL). Since we verified Assumption 1 and demonstrated that $\|D_t^j(y_{t-1}, \theta_1(3), \theta_2^*(3))\|_2 < \infty$ the LIL of Stout (1970) can be applied to establish the pointwise convergence of the score function over the non identifiable parameter set. Furthermore, using this and some tedious algebra, we can obtain LIL bounds for the derivatives of the score function with respect to $\theta_1(3)$. The strongly stochastic equicontinuouity is verified following Altissimo and Corradi's (2002) approach. That is, given the differentiability of the score with respect to $\theta_1(3)$ it is sufficient that the derivatives of the score function with respect to the nonidentifiable parameters are bounded almost surely as discussed above. Given Assumption 9, the strong consistency follows if the penalty term satisfies Assumptions 5 and 10.

We have shown that the penalized likelihood statistic converges in probability (almost surely) for the SR model. An obvious question is how to decide which information criterion to use. As suggested by Granger *et al.* (1995) we will perform a simulation study where the data is generated by the true model ($k_0 = 2$) and we choose among the set of models described above, that is $\hat{k} \in \{1, 2, 3\}$, using the penalized likelihood method. The maximum likelihood estimates of each model were obtained by the EM algorithm developed by Dempster *et al.* (1977). This algorithm is known to increase the likelihood at each step and reach a local

© Blackwell Publishing Ltd 2005

maximum of the likelihood function. Thus we start from a grid search of initial values and it remains to calculate the maximum likelihood and subtract the penalty term from it. Note that there is a wide range of penalty terms, which satisfy Assumptions 5 and 10. We consider the following function as a penalty term:

$$c_{T,k} = 0.5 \cdot L \cdot (\log(T))^b, \quad b = 0.5, 1, 1.5 \tag{20}$$

where L is the number of parameters of the model concerned. Note that when $b = 1$, we use the common BIC as our statistic.

The performance of the penalized likelihood method has been assessed by looking at several modifications of the model parameters, both due to changes of the distance between the two components of the intercept and the slope, and due to changes of the state probabilities. These modifications are based on the work done by Mendell *et al.* (1991) and more recently by Lo *et al.* (2001) who examine the empirical distribution of the likelihood ratio under a mixture distribution assumption. Their results indicate that this distribution depends on the spacing between the mixture components, sample size and the mixing proportions. Therefore, we consider the following parameterizations: $D1 = [\alpha_1 = 0.2, \alpha_2 = 0.6, \beta_1 = 0.5, \beta_2 = 0.7]$; and $D2 = [\alpha_1 = 0.2, \alpha_2 = 1.0, \beta_1 = 0.5, \beta_2 = 0.9]$; where in D2 we see that the distance between the parameters under different states is much greater than the distances in D1. Under each parameterization we will examine two configurations of the probabilities of the state variables. In the first configuration, the logistic regression parameters are $S1 = [\gamma_{1i} = -0.1, \gamma_{2i} = 0.1]$ which implies that 95% of the probabilities of state variables vary in the range 0.5 ± 0.034. In the second case, the logistic regression parameters are $S2 = [\gamma_{1i} = -2.2, \gamma_{2i} = 0.1]$ which implies that the range of these probabilities is 0.12 ± 0.015.

We examined samples of 250, 500 and 1000 observations and in each Monte Carlo exercise we used 30 replications. The results are reported in the following tables. For a given sample size, we estimated the number of times we chose each model in the set $\{1, 2, 3\}$ where the model selection procedure consisted of calculating the penalized likelihood statistic given the value of b.

The results of Table I were calculated when the data was simulated under the assumption that the probabilities of the state variables vary around half. The ability of our statistic to detect the true number of states seems to be low for $T = 250$ and 500 in the case where the slope and the intercept are not well separated, i.e. for the case of D1 and for $T = 1000$ and $b = 0.5$ the results are improving but are not satisfactory. The results improve substantially for D2, when the model parameters are further apart; the penalized likelihood statistic picks the correct model almost perfectly even for small samples. The performance of the statistic is sensitive to the choice of the penalty terms, with more significant penalty terms leading to more accurate estimation of the true value of states. In Table II, we consider the case when the state probabilities are around 12%. We note that even though these probabilities imply a low separation between the states, the results

© Blackwell Publishing Ltd 2005

TABLE I

SIMULATION RESULTS FOR THE CASE S1

	D1			D2		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
T = 250						
$b = 0.5$	21	8	1	0	27	3
$b = 1$	27	1	2	0	30	0
$b = 1.5$	29	0	1	0	30	0
T = 500						
$b = 0.5$	18	10	2	0	28	2
$b = 1$	29	0	1	0	30	0
$b = 1.5$	30	0	0	0	30	0
T = 1000						
$b = 0.5$	8	21	1	0	25	5
$b = 1$	26	4	0	0	30	0
$b = 1.5$	30	0	0	0	30	0

TABLE II

SIMULATION RESULTS FOR THE CASE S2

	D1			D2		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
T = 250						
$b = 0.5$	18	8	4	0	26	4
$b = 1$	25	0	5	0	29	1
$b = 1.5$	29	0	1	1	25	4
T = 500						
$b = 0.5$	16	8	6	0	25	5
$b = 1$	28	0	2	0	27	3
$b = 1.5$	26	0	4	0	29	1
T = 1000						
$b = 0.5$	9	15	6	0	21	9
$b = 1$	27	2	1	0	29	1
$b = 1.5$	27	0	3	0	29	1

continue to improve substantially when we increase the distance between the model parameters. In Tables I and II for D1, we observe that the use of the common BIC ($b = 1$) is not optimal, while the use of the penalty term when $b = 0.5$ is more adequate due to its robustness under different cases and sample sizes.

Note that when comparing between a switching regression model with two states and a linear regression ($k = 1$), since $k_0 = 2$, the Monte Carlo results imply the power of underlying likelihood ratio tests when the null hypothesis is $k_0 = 1$ (against the alternative $k \geq 2$) and the critical values are determined by the difference between the penalty terms. These values are equal to $2 \cdot (\log(T))^b$ and for $b = 0.5$, 1 and 1.5, we get (4.7, 5.0, 5.3), (11.0, 12.4, 13.8) and (25.9, 31.0, 36.3) respectively. The simulation results indicate a low power under D1 where the implied LR tests tend to underestimate the true number of states. For D2, the power of the test increases substantially and the implicit null hypothesis that the true model is $k = 1$ is always being rejected.

© Blackwell Publishing Ltd 2005

6. SUMMARY

This paper addresses the problem of selecting the correct number of states in switching regression models. We use the penalized likelihood method and derive conditions on the penalty term (with other regularity conditions), which ensure the weak as well as the strong consistency of the penalized likelihood statistic. We consider as an example the autoregressive switching regression model and establish the consistency properties of the penalized likelihood statistic for this model. The small sample behaviour of our statistic is analyzed via Monte Carlo simulations. The simulation results suggest our estimator converges to the true number of states as the sample grows, but the results are dependent on our selection of model parameters and are less sensitive to the range of the state variable probabilities, which are usually not known to the analyst. The penalty term $c_{T,k}$ can be interpreted as a critical value in an implicit test of the hypothesis about the choice of the model with the true number of states. Therefore, the choice of the value for the penalty term may play an important role in the performance of our statistic as was shown in our simulation study.

APPENDIX A

PROOF OF LEMMA 1. For all $1 \leq k \leq k_0$, given Assumptions 2 and 4 we can define the KLIC between the true model ($k = k_0$) and $k < k_0$ for some $\theta(k)$ as

$$I(\theta(k)) = E(\log f(y_t|w_t; \theta^*(k_0)) - E(\log f(y_t|w_t; \theta(k)).$$

Define $\tilde{I}(k) = \min_{\theta(k)\in\Theta(k)} I(\theta(k))$, Assumptions 2–4 ensure that $\tilde{I}(k)$ attains its minimum for the value $\theta^*(k)$ for $k < k_0$ and $\tilde{I}(k) > 0$ (see White, 1994, pp. 53–54). By definition of the maximum likelihood, we have

$$\frac{1}{T}Q_{T,k} - E(\log f(y_t|w_t; \theta^*(k_0)) \geq \frac{1}{T}L_T(\theta^*(k)) - E(\log f(y_t|w_t; \theta^*(k_0)). \tag{A.1}$$

We use Assumptions 1 and 4 to apply the strong law of large numbers, to get almost surely:

$$\liminf_{T\to\infty} \frac{1}{T}Q_{T,k} - E(\log f(y_t|w_t; \theta^*(k_0)) \geq -\tilde{I}(k). \tag{A.2}$$

Since the parameter set is compact, there exists a finite cover and we can divide $\Theta(k)$ into m closed balls $\Theta_1^m(k), \Theta_2^m(k), \ldots, \Theta_m^m(k)$, where $\theta_1^m(k), \theta_2^m(k), \ldots, \theta_m^m(k)$ will be an arbitrary sequence such that $\theta_i^m(k) \in \Theta_i^m(k) \cap \Theta(k)$. Let $\Theta(k, \delta, \theta')$ be a closed ball around $\theta'(k)$ where the distance between any two points in it does not exceed $\delta > 0$ and let $\bar{L}(\theta(k)) = E(\log f(y_t|w_t; \theta(k)))$. We see that

© Blackwell Publishing Ltd 2005

$$\frac{1}{T}Q_{T,k} - \bar{L}(\theta^*(k_0)) \leq \max_{1\leq i\leq m}\sup_{\theta(k)\in\Theta_i(k)}\left(\frac{1}{T}L_T(\theta(k) - \bar{L}(\theta^*(k_0))\right)$$

$$\leq \max_{1\leq i\leq m}\sup_{\theta(k)\in\Theta_i(k)}\left|\frac{1}{T}L_T(\theta(k)) - \bar{L}(\theta_i^m(k))\right| + \max_{1\leq i\leq m}\left(\bar{L}(\theta_i^m(k)) - \bar{L}(\theta^*(k_0))\right)$$

$$\leq \max_{1\leq i\leq m}\sup_{\theta(k)\in\Theta_i(k)}\left|\frac{1}{T}L_T(\theta(k)) - \frac{1}{T}L_T(\theta_i^m(k))\right|$$

$$+ \max_{1\leq i\leq m}\left|\frac{1}{T}L_T(\theta_i^m(k)) - \overline{L}(\theta_i^m(k))\right|$$

$$+ \max_{1\leq i\leq m}\left(\bar{L}(\theta_i^m(k)) - \bar{L}(\theta^*(k_0))\right)$$

$$\leq \sup_{\theta'(k)\in\Theta(k)}\sup_{\theta(k)\in\Theta(k,\delta,\theta')}\left|\frac{1}{T}L_T(\theta(k)) - \frac{1}{T}L_T(\theta'(k))\right|$$

$$+ \max_{1\leq i\leq m}\left|\frac{1}{T}L_T(\theta_i^m(k)) - \bar{L}(\theta_i^m(k))\right| + \max_{1\leq i\leq m}\left(\bar{L}(\theta_i^m(k) - \bar{L}(\theta^*(k_0))\right) \tag{A.3}$$

Under Assumptions 1–4 we can apply the uniform strong law of large numbers, and from Theorem 2 of Andrews (1992), we see that the likelihood function is strongly stochastically equicontinuous on $\Theta(k)$ and the likelihood function converges almost surely for all $\theta(k) \in \Theta(k)$. Therefore, by taking lim sup from both sides of this inequality and letting δ approach zero, the first two terms in the last inequality go to zero and we get

$$\begin{aligned}\limsup_{T\to\infty}\frac{1}{T}Q_{T,k} - \bar{L}(\theta^*(k_0)) &\leq \max_i(-I(\theta_i^m(k))) \\ &\leq -\min_i(I(\theta_i^m(k))) \leq -\tilde{I}(k_0)\end{aligned} \tag{A.4}$$

which implies that $\frac{1}{T}Q_{T,k} - \bar{L}(\theta^*(k_0)))$ converges a.s. ${}^{T_0}\tilde{I}(k)$. Using the same arguments and Assumptions 1–4 we can also show that $\frac{1}{T}Q_{T,k_0}$ converges almost surely to $\bar{L}(\theta^*(k_0))$. The rest of this proof is restricted to the event of probability one. According to Assumption 5 and the fact $\tilde{I}(k) > 0$, for all $k < k_0$ we get

$$Q_{T,k} - Q_{T,k_0} \leq c_{T,k} - c_{T,k_0} \Rightarrow IC_T(k) \leq IC_T(k_0). \tag{A.5}$$

We conclude that $\liminf_{T\to\infty}\hat{k} \geq k_0$. □

Proof of Lemma 2. The existence of a measurable maximum likelihood estimator $\hat{\theta}_{2T}(k,\theta_1)$ for each $\theta_1(k) \in \Theta_1(k)$ follows from Assumptions 1 and 2 and Theorem 2.12 of White (1994, p. 16). Let $\tilde{L}(\theta_1(k),\theta_2(k)) = E(\log(y_t|w_t;\theta_1(k),\theta_2(k)))$, using Assumption 3′ (a) given any neighbourhood $\bar{\Theta}_2(k)$ of $\theta_2^*(k)$, there exists an $\varepsilon > 0$ such that, $\lim_{T\to\infty}\inf_{\theta_1(k)\in\Theta_1(k)}\left(\tilde{L}(\theta_1(k),\theta_2^*(k)) - \max_{\theta_2\in\Theta_2(k)\setminus\bar{\Theta}_2(k)}\tilde{L}(\theta_1(e),\theta_2(k))\right) \geq \varepsilon > 0$.

Thus,

© Blackwell Publishing Ltd 2005

$$
\begin{aligned}
&\limsup_{T\to\infty} \Pr(\hat{\theta}_{2T}(k,\theta_1) \in \Theta_2(k)\backslash\bar{\Theta}_2(k) \text{ for some } \theta_1(k)) \\
&\quad \le \overline{\lim}_{T\to\infty} \Pr(\lim_{T\to\infty}\inf_{\theta_1(k)\in\Theta_1(k)}|\tilde{L}(\theta_1(k),\theta_2^*(k)) \\
&\qquad - \tilde{L}(\theta_1(k),\hat{\theta}_{2T}(k,\theta_1))| \ge \varepsilon \text{ for some } \theta_1(k)) \\
&\quad \le \overline{\lim}_{T\to\infty} \Pr(\lim_{T\to\infty}\sup_{\theta_1(k)\in\Theta_1(k)}|\tilde{L}(\theta_1(k),\theta_2^*(k)) \\
&\qquad - \tilde{L}(\theta_1(k),\hat{\theta}_{2T}(k,\theta_1))| \ge \varepsilon \text{ for some } \theta_1(k)) \\
&\quad \le \overline{\lim}_{T\to\infty} \Pr(\lim_{T\to\infty}\sup_{\theta_1(k)}|\tilde{L}(\theta_1(k),\theta_2^*(k)) \\
&\qquad - \frac{1}{T}(L_T(\theta_1(k),\hat{\theta}_{2T}(k,\theta_1))| > \varepsilon/2) \\
&\qquad + \overline{\lim}_{T\to\infty} \Pr(\lim_{T\to\infty}\sup_{\theta_1(k)}|\frac{1}{T}L(\theta_1(k),\hat{\theta}_{2T}(k,\theta_1)) \\
&\qquad - \tilde{L}_T(\theta_1(k),\hat{\theta}_{2T}(k,\theta_1))| > \varepsilon/2) \\
&\quad \le 2\cdot\overline{\lim}_{T\to\infty} \Pr(\lim_{T\to\infty}\sup_{\theta_1(k)\in\Theta_1(k),\theta_2\in\Theta_2(k)}|\tilde{L}(\theta_1(k),\theta_2(k)) \\
&\qquad - \frac{1}{T}L_T(\theta_1(k),\theta_2(k))| \ge \varepsilon/2) = 0. \qquad \text{(A.6)}
\end{aligned}
$$

The last inequality follows from the strong uniform convergence of the likelihood function, which is implied under Assumptions 1, 2 and 4, see, e.g. White (1994, Thm A.2.2). □

Proof of Lemma 3. In order to show that $LR_T = O_P(1)$, note that the LR is equal to

$$
\begin{aligned}
LR_T &= \sup_{\theta_1(k)\in\Theta_1(k)} L_T(\theta_1(k),\hat{\theta}_{2T}(k,\theta_1)) - L_T(\hat{\theta}_T(k_0)) \\
&= \sup_{\theta_1(k)\in\Theta_1(k)} (L_T(\theta_1(k),\hat{\theta}_{2T}(k,\theta_1)) - L_T(\theta_1(k),\theta_2^*(k))) \\
&\quad - (L_T(\hat{\theta}_T(k_0)) - L_T(\theta^*(k_0))). \qquad \text{(A.7)}
\end{aligned}
$$

For simplicity we write $\hat{\theta}_{2T}(k)$ instead of $\hat{\theta}_{2T}(k,\theta_1)$. We first try to find the distribution $L_T(\theta_1(k),\hat{\theta}_{2T}(k)) - L_T(\hat{\theta}_T(k_0))$ for a given $\theta_1(k)$. From Assumptions 6(a) and 3′ (b) we can use the Taylor series expansion of $\frac{1}{T}D_T(\theta_1(k),\hat{\theta}_2(k))$ around $\theta_2^*(k)$ to get

$$
\begin{aligned}
0 &= \frac{1}{T}D_T(\theta_1(k),\theta_2^*(k)) + \frac{1}{T}D_T^2(\theta_1(k),\bar{\theta}_2(k))\cdot(\hat{\theta}_{2T}(k) - \theta_2^*(k))' \\
&= \frac{1}{T}D_T(\theta_1(k),\theta_2^*(k)) + E(D_t^2(\theta_1(k),\theta_2^*(k)))\cdot(\hat{\theta}_{2T}(k) - \theta_2^*(k))' + \xi_T \qquad \text{(A.8)}
\end{aligned}
$$

where

$$
\xi_T = \left[\frac{1}{T}D_T^2(\theta_1(k),\bar{\theta}_2(k)) - E(D_t^2(\theta_1(k),\theta_2^*(k)))\right]\cdot(\hat{\theta}_{2T}(k) - \theta_2^*(k))'
$$

and $\bar{\theta}_2(k)$ lies on the chord between $\hat{\theta}_{2T}(k)$ and $\theta_2^*(k)$, Assumption 7 and Lemma 2 imply that the ξ_T term is $o(1)$ uniformly over $\Theta_1(k)$. Since

© Blackwell Publishing Ltd 2005

$$
\begin{aligned}
&\sup_{\theta_1(k)}\|\xi_T\| \\
&= \sup_{\theta_1(k)}\left\|\left\{\frac{1}{T}D_T^2(\theta_1(k),\bar{\theta}_2(k)) - E\big(D_t^2(\theta_1(k),\theta_2^*(k))\big)\right\}\cdot(\hat{\theta}_{2T}(k)-\theta_2^*(k))'\right\| \\
&\le \sup_{\theta_1(k)}\left\|\frac{1}{T}D_T^2(\theta_1(k),\bar{\theta}_2(k)) - E\big(D_t^2(\theta_1(k),\theta_2^*(k))\big)\right\|\cdot\sup_{\theta_1(k)}\left\|\cdot(\hat{\theta}_{2T}(k)-\theta_2^*(k))'\right\| \\
&\le \sup_{\theta_1(k)}\left\|\frac{1}{T}D_T^2(\theta_1(k),\bar{\theta}_2(k)) - E\big(D_t^2(\theta_1(k),\bar{\theta}_2(k))\big)\right\|\cdot\sup_{\theta_1(k)}\left\|\cdot(\hat{\theta}_{2T}(k)-\theta_2^*(k))'\right\| \\
&\quad + \sup_{\theta_1(k)}\left\|E\big(D_t^2(\theta_1(k),\bar{\theta}_2(k))\big) - E\big(D_t^2(\theta_1(k),\theta_2^*(k))\big)\right\|\cdot\sup_{\theta_1(k)}\left\|\cdot(\hat{\theta}_{2T}(k)-\theta_2^*(k))'\right\| \\
&= o(1)\cdot o(1) + o(1)\cdot o(1) = o(1)\ \text{a.s.}
\end{aligned}
\tag{A.9}
$$

where $||\cdot||$ denotes the matrix Euclidean norm. From Assumption 7(b) we see that

$$
\sqrt{T}(\hat{\theta}_{2T}(k)-\theta_2^*(k)) = -\big[E\big(D_t^2(\theta_1(k),\theta_2^*(k))\big)\big]^{-1}\frac{1}{\sqrt{T}}D_T(\theta_1(k),\theta_2^*(k)). \tag{A.10}
$$

From a Taylor expansion of $L_T(\theta_1(k),\hat{\theta}_{2T}(k))$ around, $\theta_2^*(k)$ for a given $\theta_1(k)$, we obtain

$$
\begin{aligned}
&L_T(\theta_1(k),\hat{\theta}_{2T}(k)) - L_T(\theta_1(k),\theta_2^*(k)) \\
&= D_T(\theta_1(k),\theta_2^*(k))\cdot(\hat{\theta}_{2T}(k)-\theta_2^*(k))' \\
&\quad + \frac{1}{2}(\hat{\theta}_{2T}(k)-\theta_2^*(k))'\cdot D_T^2(\theta_1(k),\bar{\theta}_2(k))\cdot(\hat{\theta}_{2T}(k)-\theta_2^*(k)) \\
&= \frac{1}{\sqrt{T}}D_T(\theta_1(k),\theta_2^*(k))\cdot\sqrt{T}(\hat{\theta}_{2T}(k)-\theta_2^*(k))' \\
&\quad + \frac{1}{2}\sqrt{T}(\hat{\theta}_{2T}(k)-\theta_2^*(k))'\cdot\big[\big(E(D_t^2(\theta_1(k),\theta_2^*(k))\big)\big]\cdot\sqrt{T}(\hat{\theta}_{2T}(k)-\theta_2^*(k)) + \frac{1}{2}\varsigma_T
\end{aligned}
\tag{A.11}
$$

where

$$
\begin{aligned}
\varsigma_T = \sqrt{T}(\hat{\theta}_{2T}(k)-\theta_2^*(k))'\cdot&\left[\frac{1}{T}\ D_T^2(\theta_1(k),\bar{\theta}_2(k)) - E\big(D_t^2(\theta_1(k),\theta_2^*(k))\big)\right] \\
&\times\sqrt{T}(\hat{\theta}_{2T}(k)-\theta_2^*(k)).
\end{aligned}
$$

Upon substituting $\sqrt{T}(\hat{\theta}_{2T}(k)-\theta_2^*(k))$ and letting $W_T(\theta_1(k)) = \frac{1}{\sqrt{T}}\ D_T(\theta_1(k),\theta_2^*(k))$, we get

$$
2\cdot(L_T(\theta_1(k),\hat{\theta}_{2T}(k)) - L_T(\theta_1(k),\theta_2^*(k))) = \vartheta_T + \varsigma_T \tag{A.12}
$$

where

$$
\begin{aligned}
\varsigma_T &= W_T(\theta_1(k))\big[E\big(D_t^2(\theta_1(k),\theta_2^*(k))\big)\big]^{-1} \\
&\quad\times\left[\frac{1}{T}D_T^2(\theta_1(k),\bar{\theta}_2(k)) - E\big(D_t^2(\theta_1(k),\theta_2^*(k))\big)\right] \\
&\quad\times\big[E\big(D_t^2(\theta_1(k),\theta_2^*(k))\big)\big]^{-1}\cdot W_T(\theta_1(k));\vartheta_T \\
&= W_T(\theta_1(k))'\big[-E\big(D_t^2(\theta_1(k),\theta_2^*(k))\big)\big]^{-1}W_T(\theta_1(k)).
\end{aligned}
$$

© Blackwell Publishing Ltd 2005

In order to show that the LR converges in distribution, we need to establish that the empirical process $W_T(\theta_1(k))$ is uniformly bounded in probability. Assumptions 1, 2 and 6(c) imply that $D_t(\theta_1(k), \theta_2^*(k))$ is a stationary ergodic martingale difference, to which the central limit theorem is applied pointwise given Assumption 6(b). Using this result, the compactness of $\Theta_1(k)$and Assumption 8, we can conclude via the empirical processes theory (Andrews, 1994) that W_T converges weakly to a unique Gaussian process over $\Theta_1(k)$. The continuous mapping theorem and Lemma 4.5 of White (2001, p. 67) imply that $\sup_{\theta_1(k)} \in \Theta_1(k)\ W_T(\theta_1(k)) = O_P(1)$, this result and Assumption 7 imply that ϑ_T is $O_p(1)$ uniformly on $\Theta_1(k)$.

We demonstrate this as follows:

$$\begin{aligned}\sup_{\theta_1(k)}\|\vartheta_T\| &\le \sup_{\theta_1(k)}\|W_T(\theta_1(k))\| \cdot \sup_{\theta_1(k)}\|[-E\left(D_t^2(\theta_1(k), \theta_2^*(k))\right)]^{-1}\| \cdot \sup_{\theta_1(k)}\|W_T(\theta_1(k))\| \\ &= O_p(1) \cdot O(1) \cdot O_p(1) = O_p(1). \end{aligned} \tag{A.13}$$

In a similar way we can show that $\sup_{\theta_1}(k) \in \Theta_1(k)|\varsigma_T| = o_p(1)$, hence

$$\sup_{\theta_1(k)\in\Theta_1(k)} (L_T(\theta_1(k), \hat{\theta}_{2T}(k)) - L_T(\theta_1(e), \theta_2^*(k))) = O_P(1) + o_p(1). \tag{A.14}$$

Under Assumptions 1–3, 6 and 7 (note $\theta_2^*(k_0) \equiv \theta^*(k_0)$ because for $k = k_0$ the model parameters are identified), it is obvious that

$$L_T(\hat{\theta}_T(k_0)) - L_T(\theta^*(k_0)) = O_P(1) \tag{A.15}$$

see e.g. White (1982, 1994). Therefore

$$LR_T = O_P(1) + O_p(1) + o_p(1) = O_P(1). \tag{A.16}$$

□

Proof of Theorem 1. Consider the probability of $\hat{k}$ being greater than k_0:

$$\Pr(\hat{k} > k_0) \le \sum_{k>k_0} \Pr(\hat{k} = k) \le \sum_{k>k_0} \Pr(IC_T(k) > IC_T(k_0)) \tag{A.17}$$

where for all $k > k_0$

$$\Pr(IC_T(k) > IC_T(k_0)) = \Pr\left(\frac{LR_T}{c_{T,k_0}} > \frac{c_{T,k}}{c_{T,k_0}} - 1\right). \tag{A.18}$$

From Assumption 5 we know that $\left(\frac{c_{T,k}}{c_{T,k_0}} - 1\right) > 0$ and $\frac{1}{c_{T,k_0}} \to 0$ and by Lemma 3 we have $LR_T = O_P(1)$. Therefore, for all $k > k_0, \Pr(\hat{k} = k) = 0$ in probability and using Lemma 1, we get that $\hat{k}$ converges to k_0 in probability. □

Proof of Lemma 4. In order to show that $\limsup_{T\to\infty} \frac{LR_T}{\log\log(T)} \dim(\Theta_2(k)) - \dim(\Theta(k_0))$, note that the LR is equal to

© Blackwell Publishing Ltd 2005

$$
\begin{aligned}
LR_T &= \sup_{\theta_1(k)\in\Theta_1(k)} L_T(\theta_1(k), \hat{\theta}_{2T}(k,\theta_1)) - L_T(\hat{\theta}_T(k_0)) \\
&= \sup_{\theta_1(k)\in\Theta_1(k)} (L_T(\theta_1(k), \hat{\theta}_{2T}(k,\theta_1)) - L_T(\theta_1(k), \theta_2^*(k))) \\
&\quad - (L_T(\hat{\theta}_T(k_0)) - L_T(\theta^*(k_0))).
\end{aligned} \tag{A.19}
$$

From Assumptions 3′(b), 6(a) and a Taylor's expansion of $D_T(\theta_1(k), \hat{\theta}_2(k))$ and $L_T(\theta_1(k), \hat{\theta}_{2T}(k))$ around $\theta_2^*(k)$ for a given $\theta_1(k)$, we get

$$
(\hat{\theta}_{2T}(k) - \theta_2^*(k)) = -\left[D_T^2(\theta_1(k), \bar{\theta}_2(k))\right]^{-1} D_T(\theta_1(k), \theta_2^*(k)) \tag{A.20}
$$

$$
\begin{aligned}
&L_T(\theta_1(k), \hat{\theta}_{2T}(k)) - L_T(\theta_1(k), \theta_2^*(k)) \\
&= (\hat{\theta}_{2T}(k) - \theta_2^*(k))' \cdot D_T(\theta_1(k), \theta_2^*(k)) \\
&\quad + \frac{1}{2}(\hat{\theta}_{2T}(k) - \theta_2^*(k))' \cdot D_T^2(\theta_1(k), \bar{\theta}_2(k)) \cdot (\hat{\theta}_{2T}(k) - \theta_2^*(k)).
\end{aligned} \tag{A.21}
$$

Let $V(\theta_1(k))$ be the variance covariance matrix of the score function. By substituting (A.20) into (A.21) and using Assumptions 1, 2 and 6(c) to establish the information matrix equality $(V(\theta_1(k)) = E(D_t(\cdot) \cdot D_t(\cdot)') = -E(D_t^2(\cdot)))$, we see that for all $\theta_1(k) \in \Theta_1(k)$:

$$
\begin{aligned}
&L_T(\theta_1(k), \hat{\theta}_{2T}(k)) - L_T(\theta_1(k), \theta_2^*(k)) \\
&= -\frac{1}{\sqrt{2T}} D_T(\theta_1(k), \theta_2^*(k))' \cdot \left[\frac{1}{T} D_T^2(\theta_1(k), \bar{\theta}_2(k))\right]^{-1} \frac{1}{\sqrt{2T}} D_T(\theta_1(k), \theta_2^*(k)) \\
&= \left(\frac{1}{\sqrt{2T}} V(\theta_1(k))^{-1/2} D_T(\theta_1(k), \theta_2^*(k))\right)' \left(\frac{1}{\sqrt{2T}} V(\theta_1(k))^{-1/2} D_T(\theta_1(k), \theta_2^*(k))\right) \\
&\quad + \frac{1}{\sqrt{2T}} D_T(\theta_1(k), \theta_2^*(k)) \cdot [E(D_t^2(\theta_1(k), \theta_2^*(k))^{-1} - E(D_t^2(\theta_1(k), \bar{\theta}_2(k))^{-1}] \\
&\quad \times \frac{1}{\sqrt{2T}} D_T(\theta_1(k), \theta_2^*(k)) - \frac{1}{\sqrt{2T}} D_T(\theta_1(k), \theta_2^*(k)) \\
&\quad \times \frac{1}{\sqrt{2T}} D_T(\theta_1(k), \theta_2^*(k)) - \frac{1}{\sqrt{2T}} D_T(\theta_1(k), \theta_2^*(k)) \\
&\quad \times \left[\frac{1}{T} D_T^2(\theta_1(k), \bar{\theta}_2(k))^{-1} - E(D_t^2(\theta_1(k), \bar{\theta}_2(k))^{-1}\right] \frac{1}{\sqrt{2T}} D_T(\theta_1(k), \theta_2^*(k)).
\end{aligned} \tag{A.22}
$$

The compactness of $\Theta_1(k)$ and Assumption 9 allow us to use Theorem 21.8 of Davidson (1994) to show that $\limsup_{T\to\infty} \sup_{\theta_1(k)\in\Theta_1(k)} \frac{1}{\sqrt{2T \log\log(T)}} D_T(\theta_1(k),\ \theta_2^*(k)) = O(1)$ almost surely. This result, Assumption 7 and Lemma 2 imply that the last two terms of the RHS of (A.22) are $o(\log\ \log\ (T))$. Let $a_T = \frac{1}{\sqrt{2T \log\log(T)}}$, since $\Theta_1(k)$ is compact there exists a finite cover $\{S(\theta_1^j(k), \delta)\, j = 1, \ldots, M)\}$, such that

© Blackwell Publishing Ltd 2005

$$\limsup_{T\to\infty}\sup_{\theta_1(k)\in\Theta_1(k)}\frac{1}{\log\log(T)}(L_T(\theta_1(k),\hat{\theta}_{2T}(k))-L_T(\theta_1(k),\theta_2^*(k)))$$

$$\leq \limsup_{T\to\infty}\max_{1\leq j\leq m}\{|a_T V(\theta_1^j(k))^{-1/2}D_T(\theta_1^j(k),\theta_2^*(k))'|$$

$$\times |a_T V(\theta_1^j(k))^{-1/2}D_T(\theta_1^j(k),\theta_2^*(k))|\}$$

$$+\limsup_{T\to\infty}\max_{1\leq i\leq m}\sup_{\theta_1'(k)\in S(\theta_1^i(k),\delta)}\{\|a_T V(\theta_1'(k))^{-1/2}D_T(\theta_1'(k),\theta_2^*(k))\|$$

$$\times \|a_T V(\theta_1'(k))^{-1/2}D_T(\theta_1'(k),\theta_2^*(k))|-|a_T V(\theta_1^j(k))^{-1/2}D_T(\theta_1^j(k),\theta_2^*(k))|$$

$$\times |a_T V(\theta_1^j(k))^{-1/2}D_T(\theta_1^j(k),\theta_2^*(k))|\}+o(1). \tag{A.23}$$

Assumption 9 implies that the first term on the RHS of the inequality obeys the pointwise LIL and converges almost surely to $\dim(\Theta_2(k))$, and the second term approaches zero almost surely as $\delta\to 0$, hence

$$\limsup_{T\to\infty}\sup_{\theta_1(k)\in\Theta_1(k)}\Big(L_T(\theta_1(k),\hat{\theta}_{2T}(k))-L_T(\theta_1(k),\theta_2^*(k))\Big)\leq \dim(\Theta_2(k))\cdot\log\log(T) \tag{A.24}$$

We can show in a similar way as in Sin and White (1996)) using Assumptions 1–3, 6, 7 and 9 (note $\theta_2^*(k_0)\equiv\theta^*(k_0)$) that

$$\limsup_{T\to\infty}\Big(L_T(\hat{\theta}_T(k_0))-L_T(\theta^*(k_0))\Big)\leq \dim(\Theta(k_0))\cdot\log\log(T) \tag{A.25}$$

hence, $\limsup_{T\to\infty} LR_T\leq(\dim(\Theta_2(k))-\dim(\Theta(k_0)))\cdot\log\log(T)$. □

Proof of Theorem 2. Given Lemma 1 it remains to prove that $\hat{k}$ does not overestimate k_0 almost surely. The event that $\hat{k}>k_0$ is given by $\Omega^*=\{\omega\in\Omega|\hat{k}(\omega)>k_0\}\subset\bigcup_{k>k_0}\{\omega\in\Omega|\hat{k}(\omega)=k)\}$.

The event $\bigcup_{k>k_0}\{\omega\in\Omega|\hat{k}(\omega)=k)\}$ implies $\bigcup_{k>k_0}\{\omega\in\Omega|IC_T(k)>IC_T(k_0))\}$, thus for any $k>k_0$ we see that

$$Q_{T,k}-c_{T,k}>Q_{T,k_0}-c_{T,k_0}\Leftrightarrow\frac{LR_T}{\log\log(T)}-\frac{c_{T,k}-c_{T,k_0}}{\log\log(T)}>0. \tag{A.26}$$

From Assumptions 5, 10 and Lemma 4, $\limsup_{T\to\infty}\left[\frac{LR_T}{\log\log(T)}-\frac{c_{T,k}-c_{T,k_0}}{\log\log(T)}\right]$ tends almost surely to a strictly negative term. Therefore, for $k>k_0$, $\Pr(\limsup_{T\to\infty}IC_T(k)>IC_T(k_0))=0$, which implies that: $\Pr\{\Omega^*\}=0$. We conclude that $\hat{k}$ converges to k_0 almost surely. □

© Blackwell Publishing Ltd 2005

ACKNOWLEDGEMENT

The authors thank Uri Ben-Zion, Ezra Einy, Niklas Wagner and an anonymous referee for several insightful comments and suggestions and the seminar participants of the FFM 2004 conference and the Technion – Israel Institute of Technology for their comments, which improved this paper. Preminger gratefully acknowledges research support from the Kreitman Foundation.

NOTES

1. We should note that in our case, the models considered are strictly nested, and usually the more parsimonious model is selected as the null hypothesis. Hence the order of the tests does not pose any problem in hypothesis testing.
2. The standard measurability and continuity requirements are defined as in Wooldridge (1994, p. 2726).
3. Our next assumptions will be assumed for all $k \geq k_0$.
4. The stationary and ergodicity assumption may be too restrictive in certain time series applications and the NED process can be used instead, allowing for some degree of heterogeneity in the DGP.
5. r-domination is defined as in Gallant and White (1988, p.35).

REFERENCES

AKAIKE, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19, 716–23.

ALTISSIMO, F. and CORRADI, V. (2002) Bounds for inference with nuisance parameters present only under the alternatives. *Econometrics Journal* 5, 494–518.

ANDREWS, D. W. K. (1992) Generic uniform conditions. *Econometric Theory* 8, 241–57.

ANDREWS, D. W. K. (1993) An introduction to econometric applications of empirical process theory for dependent random variables. *Econometric Review* 12, 183–216.

ANDREWS, D. W. K. (1994) Empirical process methods in econometrics. In *Handbook of Econometrics 4*, ch. 37. New York, North-Holland, pp. 2248–92.

ANDREWS, D. W. K. and PLOBERGER, W. (1994) Optimal tests when the nuisance parameters are present only under the alternative. *Econometrica* 62, 1383–414.

ANDREWS, D. W. K. and POLLARD, D. (1994) An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review* 62, 119–32.

BAUER, H. (1972) *Probability Theory and Elements of Measure Theory*. New York, Holt, Rinehart and Winston.

CHAN, K. S. (1993) A review of some limit theorems of Markov chains and their applications, In *Dimension Estimation and Models* (ed. H. TONG), World Scientific Publishing. Singapore, New York.

CHAN, K. S. and TONG, H. (1985) On the use of deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Advances in Applied Probability* 17, 666–78.

DAVIDSON, J. (2002) Establishing conditions for the functional central limit theorem in nonlinear and semiparametric time series processes. *Journal of Econometrics* 106, 243–69.

DAVIDSON, J. (1994) *Stochastic Limit Theory*. New York, Oxford University Press.

DAVIES, R. B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–54.

© Blackwell Publishing Ltd 2005

DAVIES, R. B. (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74, 33–43.
DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society* B, 39, 1–38.
FENG, Z. D. and MCCULLOCH, C. E. (1996) Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society* B 58, 609–17.
GALLANT, A. R. and WHITE, H. (1988) *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Oxford, Basil Blackwell.
GRANGER, C. W. J., KING, M. L. and WHITE, H. (1995) Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics* 67, 173–87.
HAMILTON, J. D. (1989) A new approach to economic analysis of non-stationary time series and business cycles. *Econometrica* 57, 357–84.
HAMILTON, J. D. (1990) Analysis of time series subject to changes in regimes. *Journal of Econometrics* 45, 39–70.
HANSEN, B. E. (1992) The likelihood ratio test under nonstandard conditions: testing the Markov-switching model of GNP. *Journal of Applied Econometrics* 7, S61–82.
HANSEN, B. E. (1996a) Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64, 413–30.
HANSEN, B. E. (1996b) Stochastic equicontinuity for unbounded dependent heterogeneous arrays. *Econometric Theory* 12, 347–59.
JEFFRIES, O. N. (1998) Logistic mixture of generalized linear model times series. Ph.D. Dissertation, University of Maryland at College Park, Maryland.
KAPETANIOS, G. (2001) Model selection in threshold models. *Journal of Time Series Analysis* 22, 733–54.
KULLBACK, S. and LEIBLER, R. A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics* 22, 79–86.
LAM, P. S. (1990) The Hamilton model with general autoregressive component: estimation and comparison with other models of economic time series. *Journal of Monetary Economics* 26, 409–32.
LANNE, M. and SAIKKONEN, P. (2003) Modeling the U.S. short-term interest rate by mixture autoregressive processes. *Journal of Financial Econometrics* 1, 96–125.
LEROUX, B. G. (1992) Consistent estimation of a mixing distribution. *The Annals of Statistics* 20, 1350–60.
LI, W. K. and WONG, C. S. (2001) On a logistic mixture autoregressive model. *Biometrika* 88, 833–46.
LO, Y., MENDELL, N. R. and RUBIN, D. B. (2001) Testing the number of components in a normal mixture. *Biometrika* 88, 767–78.
MENDELL, N. R., THODE, H. C. and FINCH, S. J. (1991) The likelihood ratio test for the two components normal mixture problem: power and sample size analysis. *Biometrics* 47, 1143–48.
NISHII, R. (1988) Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis* 27, 392–403.
RAO, R. R. (1962) Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics* 33, 659–80.
REDNER, R. (1981) Note on the consistency of the maximum likelihood estimate for non-identifiable distributions. *Annals of Statistics* 9, 225–39.
SCHWARZ, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6, 461–4.
SIN, C. Y. and WHITE, H. (1992). Information criteria for selecting possibly misspecified parametric models, Department of Economics Discussion Paper 92–47. University of California, San Diego.
SIN, C. Y. and WHITE, H. (1996) Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics* 71, 207–25.
STOUT, W. F. (1970) The Hartman-Winter law of the iterated logarithm for martingale. *The Annals of Mathematical Statistics* 41, 2158–60.
TJOSTHEIM, D. (1990) Non-linear time series and Markov chain. *Advances in Applied Probability* 22, 587–611.
VUONG, H. Q. (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–33.
WHITE, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25.
WHITE, H. (1994) *Estimation, Inference and Specification Analysis*. New York, Cambridge University Press.
WHITE, H. (2001) *Asymptotic Theory for Econometricians (revised edition)*. New York, Academic Press.

© Blackwell Publishing Ltd 2005

WOOLDRIDGE, J. M. (1986) Asymptotic properties of econometric estimators. Ph.D. Dissertation, University of California, San Diego.

WOOLDRIDGE, J. M. (1994) Estimation and inference for dependent processes, In *Handbook of Econometrics 4* (eds R. F. ENGLE and D. L. MCFADDEN). Amsterdam, Elsevier Science B.V, pp. 2641–7000.

© Blackwell Publishing Ltd 2005