

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

Arie Preminger and David Wettstein

Discussion Paper No. 03-14

December 2003

Monaster Center for Economic Research
Ben-Gurion University of the Negev
P.O. Box 653
Beer Sheva, Israel

Fax: 972-8-6472941
Tel: 972-8-6472286

**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*

Department of Economics, Monaster Center for Economic Research

Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

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 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 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, Extended Switching Regression Models, Penalized Likelihood Method, Law of the Iterated Logarithm

JEL classification: C12, C32, C52

(*) The authors thank Uri Ben-Zion, Ezra Einy and Niklas Wagner for several insightful comments and suggestions and seminar participants of the FFM 2003 conference and the Technion – Israel institute of Technology for their comments, which improved this paper. Preminger gratefully acknowledges research support from the Kreitman Foundation.

1. Introduction

Hypothesis testing plays a crucial role 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 that analyze the problem of unidentified nuisance parameters. His work suggests viewing the test statistic as a function of the unidentified parameters, in order 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 (1996) proposed a similar approach in the context of testing nonlinear terms in linear 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 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 will 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 simulation is 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 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.

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 a generalized version of switching regression models. These models serve as tools to model data dynamics, and are applied in several areas of empirical work. This class of models has been proposed by Preminger et al. (2003a, 2003b) in which, a model with several latent state variables is considered. In such a model the parameters are partitioned into disjoint groups, each one of which is independently determined by a corresponding state variable. These models were called the Extended Switching Regression (ESR) models. The estimation of such models requires, in particular, the specification of the number of states assumed by the latent variables. An important hypothesis that we will address in this paper is the validity of the model structure i.e. the number of states for each state variable, given the number of state variables.

Determining the number of states in a switching regression model, which is equivalent to an ESR model with one state variable, 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-squared distribution, since under the null hypothesis the nuisance parameters that describe the N -th state are unidentified. That is, the likelihood function under the null hypothesis is non-quadratic and flat with respect to the nuisance parameters 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. We encounter the same singularity problem in the extended switching regression models because they are nested in the switching regression models. Therefore, we analyze in this paper, ESR models with time varying probabilities which we denote by TV-ESR.

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 latent structure of the TV-ESR model will be chosen.

This paper is organized as follows: in Section 2 we discuss the TV-ESR models and define the structure vector that represents the model's latent structure and the penalized likelihood statistic we use 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

In this section we describe the general set up and the TV-ESR model as well as the penalized likelihood statistic used for model selection. More specifically, the observed data is a realization of a stochastic process $\{Z_t : \Omega \rightarrow R^v, t = 1, 2, \dots\}$ on a complete probability space $(\Omega, \mathfrak{F}, P_0)$, where $\Omega = \times_{t=1}^{\infty} R^v$ and \mathfrak{F} is the Borel σ -field generated by measurable finite dimensional product cylinders, and P_0 is the probability measure governing the behavior of the data.

Assumption 1: The random vectors $\{Z_t\}_{t \in \mathbb{N}}$ are strictly stationary and ergodic.

Let \mathfrak{F}_t be the σ -field generated by current and past Z_t , i.e. $\mathfrak{F}_t = \sigma(\dots, Z_{t-1}, Z_t)$, where $\mathfrak{F}_{t-1} \subset \mathfrak{F}_t \dots \subset \mathfrak{F}$. 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 $\psi \in \Psi \subset R^d$, a compact set and conditioned on $\bar{\mathfrak{F}}_{t-1} \equiv \sigma(Z_{t-\tau}, \dots, Z_{t-1}, X_t)$, $\tau < \infty$, which is given by

$\{P_t^\psi(y_t | \bar{\mathfrak{F}}_{t-1}; \psi), \psi \in \Psi, \bar{\mathfrak{F}}_{t-1} \subset \mathfrak{F}_t\}$. These measures exist by Jirina's theorem (Bauer p.319, 1972) and our parametric models include explicitly a finite number of lags. These conditional distributions also have Radon-Nikodym densities with respect to the usual Lebesgue measure that is $f(y_t | w_t, \psi) \equiv dP_t^\psi(y_t | w_t; \psi) / d\nu$, 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).

Let us assume that the data generating process is according to the TV-ESR 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 there are several unobserved selection processes which are characterized by p independent selections

from the parameter sets $\{\Psi_i\}$ such that $\Psi = \times_{i=1}^p \Psi_i$. From each set Ψ_i we select one

element among k possible ones. The selection is random and dependent on the value

of the state variables $\{s_t^i\}$, which can assume only an integer value $\{1, \dots, k\}$. Let $\{\Pr(s_t^i = j \mid \tilde{z}_{it}, \gamma_i)\}_{j=1}^k$ be the probability distribution in the i -th selection procedure. The state probabilities are not constant where the vector \tilde{z}_{it} contains explanatory variables that affect the state probabilities and are made up of known measurable functions of w_t . $\gamma_i \in \Gamma_i$ is the parameter vector, where the set Γ_i is compact and restricted to allow only time varying probabilities and these functions satisfy the standard measurability and continuity requirements¹ on $\tilde{z}_{it} \times \Gamma_i$, for all i, t .

Our analysis can be applied to a switching regression model by assuming that $p = 1$. Let $\varphi_i \subset \Psi_i \subset R^{d_i}$ be a set of k distinct values chosen by s_t^i , where each element of this set is denoted by φ_{ij_i} with probability $\Pr(s_t^i = j_i \mid \tilde{z}_{it}, \gamma_i)$ where $j_i \in \{1, \dots, k\}$. The conditional density of y_t can be described by:

$$(1) \quad f(y_t \mid w_t, \theta) = \sum_{j_1=1}^k \cdots \sum_{j_p=1}^k \left[\left(\prod_{l=1}^p \Pr(s_t^l = j_l \mid \tilde{z}_{it}; \gamma_l) \right) \cdot f(y_t \mid w_t, \{s_t^l = j_l\}, \varphi_{1j_1}, \dots, \varphi_{pj_p}) \right]$$

$\theta = (\varphi_{11}, \dots, \varphi_{1k}, \dots, \varphi_{p1}, \dots, \varphi_{pk}, \gamma_1, \dots, \gamma_p) \in \Phi \subset R^L$ is the vector of the model parameters. We define the likelihood function of the sample as

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

Next we define the structure vector as $e = [k_1, \dots, k_p, d_1, \dots, d_p]$. The first p elements of this vector denote the number of values (states) in each φ_i and the last p elements denote its dimension ($\dim(\Psi_i)$).

For simplicity, we assume a priori that $k_i = k$, $1 \leq k \leq K_{MAX}$ and $d_1 = d_2 = \dots = d_p = 1$ that is, $p = \dim(\Psi)$. Thus, the structure vector can be written as $e = [k, \dots, k] \in Z_{++}^p$. The case where $k = 1$ corresponds to a scenario where the set of parameters governing the data generation process is constant over time. When $p = 1$

¹ The standard measurability and continuity requirements are defined as in Wooldridge (p. 2726, 1994)

and $k > 1$ we get mixture models, or more generally, switching regression models with k unobserved states. The TV-ESR model, previously described, with p state variables corresponds to the general case of a p -dimensional structure vector e . Let e^* be the true structure generating the data.

Usually we do not have any information about the true structure vector and our goal is to estimate the correct one from the set of all possible structure vectors,

$E = \{e \mid e = (k, \dots, k) \in Z_{++}^p \mid 1 \leq k < K_{\max}\}$ which we assume, contains the true structure vector. The set E can be split into two sets. The first set consists of $e \leq e^*$, where the true model is not nested in the alternative models. When $e < e^*$, we minimize the Kullback-Leibler information criterion (White (1982, 1984, 1994) and Vuong (1983, 1989)). The second set consists of $e > e^*$ where 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 $\theta(e) = [\theta_1(e), \theta_2(e)]$. Let $\theta_1(e) \in \Theta_1(e) \subset R^{L_1}$ be the parameters which are not identified when we assume that the structure vector is “bigger” than the true one, $\theta_2(e) \in \Theta_2(e) \subset R^{L_2}$ are the other parameters ($L_1 + L_2 = L$), where $\Theta_1(e), \Theta_2(e)$ are compact sets.

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

$$(3) \quad \pi \cdot f_1 + (1 - \pi) \cdot f_1 = 1 \cdot f_1 + 0 \cdot f_2$$

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. More generally, when $e > e^*$, part of the model parameters will converge to values which will reduce the model's structure to a structure equivalent to e^* , and in this case, some of the model parameters are not identified as was shown in the example above.

For a given structure vector e , we consider a model with parameter space $\Theta(e)$ and quasi-likelihood functions $\{f(y_t | w_t; \theta(e)) : e \in E, \theta(e) \in \Theta(e), t = 1, 2, \dots\}$.

For any $e \in E$ define the likelihood function:

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

The maximum likelihood statistic is:

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

where $\hat{\theta}_T(e)$ is the Maximum Likelihood Estimator (MLE). The penalized likelihood is defined by subtracting a penalty term from the maximum likelihood statistic, yielding what is called information criteria:

$$(6) \quad IC_T(e) = Q_{T,e} - c_{T,e}$$

The penalized likelihood statistic \hat{e} which estimates the true structure vector e^* is the maximum of the penalized likelihood:

$$(7) \quad \hat{e} = \arg \max_{e \in E} (IC_T(e))$$

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 (1973), who defined the Akaike 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 Estimated Structure Vector

For our further discussions, we need the following assumptions:

Assumption 2: For all $e \in E$ the functions $f(y_t | w_t; \theta(e))$ are positive and measurable $\sigma(y_t, w_t) \subset \overline{\mathfrak{F}}_t$ for every $\theta(e)$ in $\Theta(e) \subset R^L$ a compact set, and are continuous on $\Theta(e)$ for each (y_t, w_t) a.s. P_0 , for all t

Assumption 3: For all $e \in \overline{E}$, where $\overline{E} = \{e \in E | e \leq e^*\}$, $E(\log f(y_t | w_t; \theta(e)))$ has a unique maximum at $\theta^*(e)$ in $\Theta(e)$.

Assumption 4: For all $e \in E$, $|\log f(y_t | w_t; \theta(e))| < m(y_t, w_t)$ for all $\theta(e) \in \Theta(e)$ 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 $c_{T, \tilde{e}} > c_{T, e} > 0$ for $\tilde{e} > e$,

$\lim_{T \rightarrow \infty} c_{T, e} = +\infty$, $c_{T, e} = o(T)$.

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 $\log f(y_t | w_t, \theta)$ that has a finite expectation. In the TV-ESR model, as in the SR model with time varying probabilities, 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, the assumptions we made in this work establish the existence and consistency of the penalized likelihood statistic in the quotient topology. This topology is defined relative to the equivalence relation under which two sets of parameters are equivalent if they define the same (penalized) likelihood function (see Leroux (1992), Redner (1981) for the case of mixture distributions). This framework allows us to identify the parameters which maximize

the likelihood function as a singleton set and simplifies our discussion. The next lemma establishes that asymptotically the estimator \hat{e} does not underestimate the number of states for each state variable.

Lemma 1: Given Assumptions 1-5 $\hat{e} \geq e^*$ almost surely.

Proof: see Appendix.

In order to prove that \hat{e} is weakly consistent, we will have also to establish that it cannot overestimate the latent structure vector. We proceed to show first that for any $e > e^*$ the likelihood ratio converges in distribution. However, when $e > e^*$ some of the model parameters are not identified and we face the problem of singular information matrix mentioned above. Our approach in this case would be to view the likelihood ratio as an empirical process over the non-identified parameters. This approach was used by Hansen (1992) to test the number of regimes in a Markov switching model proposed by Hamilton (1989), and see also Andrews (1994) for the econometric applications of empirical process theory.

Next, we need to introduce some more notations and assumptions².

Let $L_T(\theta_1(e), \theta_2(e)) = \sum_{t=1}^T \log f(y_t | w_t; \theta_1(e), \theta_2(e))$ denote the likelihood function

given the parameters $\theta_1(e), \theta_2(e)$ and $D_t(\theta_1(e), \theta_2(e))$ denote the L_2 -vector of partial derivatives of $\log f(y_t | w_t; \theta_1(e), \theta_2(e))$ with respect to $\theta_2(e)$, and

$D_t^2(\theta_1(e), \theta_2(e))$ denote the $L_2 \times L_2$ matrix of second partial derivatives with respect to

$\theta_2(e)$. Let $D_T(\theta_1(e), \theta_2(e)) = \sum_{t=1}^T D_t(\theta_1(e), \theta_2(e))$ and

$D_T^2(\theta_1(e), \theta_2(e)) = \sum_{t=1}^T D_t^2(\theta_1(e), \theta_2(e))$. Let $\hat{\theta}_{2T}(e, \theta_1)$ be the maximum likelihood

estimator of $\theta_2(e)$ for a fixed $\theta_1(e) \in \Theta_1(e)$, i.e. the estimator satisfies:

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

² Our next Assumptions will be assumed for all $e > e^*$.

We denote by $\theta_2^*(e)$ the true value of $\theta_2(e)$ in which some of the model parameters are not identified for $e > e^*$, i.e. $L_T(\theta_1(e), \theta_2^*(e)) = L_T(\theta_2^*(e^*))$ for all $\theta_1(e)$, note that the likelihood function does not depend on the nuisance parameters under the true model structure. The likelihood ratio statistic is defined as:

$$(9) \quad LR_T = 2 \cdot (Q_{T,e} - Q_{T,e^*}) = 2 \cdot \left(\sup_{\theta_1(e) \in \Theta_1(e)} L_T(\theta_1(e), \hat{\theta}_{2T}(e, \theta_1)) - L_T(\hat{\theta}_T(e^*)) \right)$$

Assumption 3’:

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

$$\lim_{T \rightarrow \infty} \inf_{\theta_1(e) \in \Theta_1(e)} \left(E(y_t | w_t; \theta_1(e), \theta_2^*(e)) - \max_{\theta_2 \in \Theta_2(e) \setminus \bar{\Theta}_2(e)} E(y_t | w_t; \theta_1(e), \theta_2(e)) \right) > 0$$

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

Assumption 6:

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

(b) The elements of $|\partial \log f(y_t | w_t, \theta_1(e), \theta_2(e)) / \partial \theta_2(e) \cdot \partial \log f(y_t | w_t, \theta_1(e), \theta_2(e)) / \partial \theta_2(e)|$ are dominated by P_0 -integrable functions independent of $\theta_2(e)$ for all $\theta_1(e) \in \Theta_1(e)$

(c) For all $\theta_1(e) \in \Theta_1(e)$ the elements of $|\partial f(y_t | w_t, \theta_1(e), \theta_2(e)) / \partial \theta|$ and

$|\partial^2 f(y_t | x_t, \theta_1(e), \theta_2(e)) / \partial \theta_2(e) \cdot \partial \theta_2(e)|$ are dominated by P_0 -integrable functions independent of $\theta_2(e)$.

Assumption 7:

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

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

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

The conditions above are sufficient to establish that the likelihood ratio is uniformly bounded in probability. Assumption 3' modified Assumption 3 to allow for uniform identifiability on $\theta_1(e)$. Assumptions 6-7 are standard assumptions that impose moments and smoothness conditions that are commonly used to derive consistency and asymptotic normality for the MLE $\forall \theta_1(e) \in \Theta_1(e)$. For a fixed $\theta_1(e)$, Assumption 7(a) can be verified using the strong law of large numbers for stationary and ergodic sequences. Uniform convergence over $\theta_1(e) \in \Theta_1(e)$ can then be obtained, e.g. by using results in Davidson (1994, pp.327-344). Note that Assumption 7(b) does not hold even for a fixed value of $\theta_1(e)$ in switching regression models or more generally in the ESR models with constant state variables probabilities. The stochastic equicontinuity property, in Assumption 8, can be deduced by assuming $\frac{1}{\sqrt{T}} D_T(\theta_1(e), \theta_2^*(e))$ satisfies Lipschitz-type sufficient conditions, see, e.g. Andrews (1992). These assumptions are used to prove the following lemmata.

Lemma 2: Given Assumptions 1, 2, 3', 4 for $e > e^*$, $\hat{\theta}_{2T}(e, \theta_1) \rightarrow \theta_2^*(e)$ almost surely, uniformly over $\Theta_1(e)$.

Proof: see Appendix.

We will use lemma 2 and a Taylor expansion of the likelihood function in the neighborhood of the identified parameters and the equicontinuity of the score function to show that the likelihood ratio is bounded in probability

Lemma 3: Given Assumptions 1, 2, 3', 4, 6, 7 for $e > e^*$, $LR_T = O_p(1)$

Proof: see Appendix.

Given the results of the lemmata above, the following theorem establishes the weak consistency (convergence in probability) of the penalized likelihood statistic.

Theorem 1: Given Assumptions 1, 2, 3', 4-8, \hat{e} converges to e^* in probability.

Proof: see Appendix.

4. Almost Sure Consistency of the Estimated Structure Vector

In this section 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 structure vector, with probability one. By lemma 1 we know that \hat{e} does not asymptotically underestimate the elements of the true structure vector. Therefore, it remains to prove that \hat{e} does not asymptotically overestimate the elements of the true structure vector. We will use the Uniform Law of the Iterated Logarithm (ULIL) when $e > e^*$. The definition of the ULIL which will be used in this work is as follows:

Definition: $\{u_t : \Omega \times \Phi \rightarrow R\}$ is said to satisfy a ULIL on $\Gamma \subseteq \Phi$ if almost surely

$$\limsup_{T \rightarrow \infty} \sup_{\theta \in \Gamma} \left| \sum_{t=1}^T (u_t - E(u_t)) \right| = O\left(\sqrt{T \log \log(T)}\right)$$

In the context of model selection, the Law of the Iterated Logarithm (LIL) was first applied by Nishii (1988) who 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 a Kullback-Leibler (1951) optimal model, with probability one, as the sample size increases. Sin and White (1996) generalized Nishii's (1988) results to dependent and heterogeneous processes. In our case, we have to prove that the likelihood ratio is bounded almost surely uniformly on $\Theta_1(e)$, e.g. over the set of parameters which is not identified when $e > e^*$, to obtain the strong consistency of our statistic. Therefore we have to add the following assumptions:

Assumption 9: For all $\theta_1(e) \in \Theta_1(e)$ the elements of $D_T(\theta_1(e), \theta_2^*(e))$ satisfy the ULIL on $\Theta_1(e)$.

Assumption 10: The penalty term also satisfies $\lim_{T \rightarrow \infty} \frac{\log \log(T)}{c_{T,e}} = 0$.

Assumption 9 can be verified following Altissimo and Corradi's (2002) approach, i.e. first providing bounds valid pointwise for all $\theta_1(e) \in \Theta_1(e)$ for the gradient of the likelihood function via the LIL for martingale difference sequences, and then by showing via a strong stochastic equicontinuity argument, that such bounds are also valid uniformly over $\Theta_1(e)$. For example, we can use the results by Stout (1970) and Assumptions 1 and 6(b) to show that for each value of the nuisance parameters, the score $D_T(\theta_1(e), \theta_2^*(e))$ satisfies the LIL. In order to obtain a strong stochastic equicontinuity, we can assume that the score function will agree with the Lipschitz-type sufficient conditions (see, e.g. Andrews (1992)). These conditions will be met if we assume that the score function is differentiable almost surely at each point of $\Theta_1(e)$, and that the gradient vector of the score function with respect to $\theta_1(e)$ can be bounded uniformly over $\Theta_1(e)$ and satisfy the LIL.

Lemma 4: Given Assumptions 1, 2, 3', 4, 6(a), 7, 9, for $e > e^*$,

$$\limsup_{T \rightarrow \infty} \sup_{\theta_1(e) \in \Theta_1(e)} \frac{LR_T}{\log \log(T)} = O(1)$$

Proof: see Appendix.

Theorem 2: Given Assumptions 1, 2, 3', 4, 5, 6(a), 7, 9, 10, \hat{e} converges to e^* almost surely.

Proof: see Appendix.

5. Simulation Results

We will illustrate the implications of our theoretical results and examine the performance of our statistic in small samples and its sensitivity to different penalty terms. We will consider the following time varying extended switching regression model (TV-ESR), which we assume is the true data generating process:

$$(10) \quad y_t = \alpha_t + \beta_t x_t + \varepsilon_t$$

where $\{(\varepsilon_t, x_t)\}_{t \in N}$ are distributed as standard normal variables and $s_t^1, s_t^2 \in \{1, 2\}$ are unobserved independent state variables, which determine the value of coefficients α_t, β_t . By this we mean:

$$(11) \quad y_t | x_t \sim \begin{cases} f(y_t | x_t, \alpha_1, \beta_1) & \text{if } s_t^1 = 1, s_t^2 = 1 \\ f(y_t | x_t, \alpha_1, \beta_2) & \text{if } s_t^1 = 1, s_t^2 = 2 \\ f(y_t | x_t, \alpha_2, \beta_1) & \text{if } s_t^1 = 2, s_t^2 = 1 \\ f(y_t | x_t, \alpha_2, \beta_2) & \text{if } s_t^1 = 2, s_t^2 = 2 \end{cases}$$

and $f(y_t | x_t, \alpha_t, \beta_t) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-(y_t - \alpha_t - \beta_t x_t)^2 / 2\right]$.

Given the state variables, the unknown parameters in the model are $[\alpha_t, \beta_t]$ where the state probabilities are not constant over time and are given by a logistic model:

$$(12) \quad \Pr(s_t^i = 1 | \tilde{z}_{it}; \gamma_{1i}, \gamma_{2i}) = \Lambda(\gamma_{1i} + \tilde{z}_{it} \gamma_{2i}) \quad , i = 1, 2$$

where for $i = 1, 2$, \tilde{z}_{it} is distributed uniformly on $[0, 1]$ and $[\gamma_{1i}, \gamma_{2i}]$ are unknown model parameters. The true structure vector for this model is $e^* = [2, 2]$, and we will apply the penalized likelihood method to estimate the true structure vector among the set $E = \{[1, 1], [2, 2], [3, 3]\}$. When $e = [1, 1]$ the model is a linear regression model and when $e = [3, 3]$ the model parameters are defined over two state variables where each variable can assume three possible states according to the following probabilities:

$$(13) \quad \begin{aligned} \Pr(s_t^i = 1 | \tilde{z}_{it}; \gamma_{1i}, \gamma_{2i}) &= \Lambda(\gamma_{1i} + \tilde{z}_{it} \gamma_{2i}) \\ \Pr(s_t^i = 2 | \tilde{z}_{it}; \gamma_{1i}, \gamma_{2i}) &= \Lambda(\mu + \gamma_{1i} + \tilde{z}_{it} \gamma_{2i}) - \Lambda(\gamma_{1i} + \tilde{z}_{it} \gamma_{2i}) \end{aligned} \quad i = 1, 2, \quad \mu > 0$$

The maximum likelihood estimates of the models parameters were obtained by the EM algorithm developed by Preminger et al. (2003b). This algorithm is known to increase the likelihood at each step and reach a local 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:

$$(14) \quad c_{T,e} = 0.5 \cdot L \cdot (\log T)^b \quad b = 0.5, 1, 1.5$$

where L is the number of parameters in 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 Mendel et al. (1991) and more recently by Lo et al. (2001), which examine the empirical distribution of the likelihood ratio under a mixture 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 $S1 = [\gamma_{1i} = -0.1, \gamma_{1i} = 0.2, \gamma_{21}]$ which implies that the probabilities of state variables vary in the range 0.5 ± 0.025 . In the second case, the logistic regression parameters are $S2 = [\gamma_{1i} = -2, \gamma_{1i} = 0.1, \gamma_{21}]$, which implies that the range of these probabilities is 0.125 ± 0.005 .

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 E, where the model selection procedure consisted of calculating the penalized likelihood statistic given the value of b .

Table 1: Simulation results for the case S1

		D1			D2		
		[1,1]	[2,2]	[3,3]	[1,1]	[2,2]	[3,3]
T=250	$b = 0.5$	22	5	3	0	29	1
	$b = 1$	27	0	3	4	23	3
	$b = 1.5$	30	0	0	28	2	0
T=500	$b = 0.5$	17	11	2	0	29	1
	$b = 1$	27	2	1	0	30	0
	$b = 1.5$	28	0	2	16	12	2
T=1000	$b = 0.5$	7	22	1	0	29	1
	$b = 1$	26	4	0	0	29	1
	$b = 1.5$	29	0	1	0	29	1

The results of table 1 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 structure model seems to be low for $T=250$ and $T=500$ in the case where the slope and the intercept are not well separated i.e. for the case of D1. The results improve for the case of D2 when the model parameters are further apart. The performance of the penalized likelihood statistic is very sensitive to the choice of the penalty terms, with more significant penalty terms leading to underestimation of the value of the structure vector. For $T=1000$, and D2, all the penalty terms lead to the true structure vector, while for D1, the results are improving but are not satisfactory.

Table 2: Simulation results for the case S2

		D1			D2		
		[1,1]	[2,2]	[3,3]	[1,1]	[2,2]	[3,3]
T=250	$b = 0.5$	23	4	3	6	23	1
	$b = 1$	24	0	6	19	8	3
	$b = 1.5$	25	0	5	22	0	8
T=500	$b = 0.5$	20	3	7	2	26	2
	$b = 1$	26	0	4	8	19	3
	$b = 1.5$	26	0	4	29	0	1
T=1000	$b = 0.5$	16	12	2	0	29	1
	$b = 1$	28	0	2	4	25	1
	$b = 1.5$	30	0	0	24	5	1

In table 2, we consider the case when the state probabilities are around 12%, which implies a very low separation between the states. Convergence of the penalized likelihood statistic is very slow, for D1 and the results are an underestimate for $b = 1.5$ and even for $T=1000$ it is very difficult for the statistic to separate the model states. However, for D2, the results are better for all sample sizes, especially for $b = 0.5$. This implies that the use of the common BIC criterion ($b = 1$) is not optimal under any case, while the use of the penalty term when $b = 0.5$ is more

adequate due to its robustness under different cases and sample sizes, as was shown in both tables.

6. Summary

This paper addresses the problem of selecting the correct structure of extended switching regression models among several competing 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. The small sample behavior of our statistic is analyzed via Monte Carlo simulations. The simulation results suggest our estimator converges to the true structure as the sample grows, but the results are dependent on our selection of model parameters and the range of the state variable probabilities, which are usually not known to the analyst. The penalty term $c_{T,e}$ can be interpreted as a critical value in an implicit test of the hypothesis about the choice of the model with the true structure vector. 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 simulations study.

Appendix

Proof of Lemma 1:

For all $e \in \bar{E}$, given Assumptions 2, 4 we can define the Kullback distance between the true model ($e = e^*$) and $e < e^*$ for some $\theta(e)$ as

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

Define $\tilde{I}(e) = \max_{\theta(e) \in \Theta(e)} I(\theta(e))$, Assumptions 2-4 ensure that $\tilde{I}(e)$ attains its minimum for the value $\theta^*(e)$ for $e < e^*$ and $\tilde{I}(e) > 0$, (see White (1994, pp. 53-54)).

By definition of the maximum likelihood, we have:

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

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

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

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

$$\begin{aligned}
\text{(A.3)} \quad & \frac{1}{T} Q_{T,e} - \bar{L}(\theta^*(e^*)) \leq \max_{1 \leq i \leq m} \sup_{\theta(e) \in \Theta_i^m(e)} \left(\frac{1}{T} L_T(\theta(e)) - \bar{L}(\theta^*(e^*)) \right) \\
& \leq \max_{1 \leq i \leq m} \sup_{\theta(e) \in \Theta_i^m(e)} \left| \frac{1}{T} L_T(\theta(e)) - \bar{L}(\theta_i^m(e)) \right| + \max_{1 \leq i \leq m} \left(\bar{L}(\theta_i^m(e)) - \bar{L}(\theta^*(e^*)) \right) \\
& \leq \max_{1 \leq i \leq m} \sup_{\theta(e) \in \Theta_i^m(e)} \left| \frac{1}{T} L_T(\theta(e)) - \frac{1}{T} L_T(\theta_i^m(e)) \right| + \max_{1 \leq i \leq m} \left| \frac{1}{T} L_T(\theta_i^m(e)) - \bar{L}(\theta_i^m(e)) \right| + \\
& \max_{1 \leq i \leq m} \left(\bar{L}(\theta_i^m(e)) - \bar{L}(\theta^*(e^*)) \right) \leq \sup_{\theta'(e) \in \Theta(e)} \sup_{\theta(e) \in \Theta(e, \delta, \theta')} \left| \frac{1}{T} L_T(\theta(e)) - \frac{1}{T} L_T(\theta'(e)) \right| + \\
& + \max_{1 \leq i \leq m} \left| \frac{1}{T} L_T(\theta_i^m(e)) - \bar{L}(\theta_i^m(e)) \right| + \max_{1 \leq i \leq m} \left(\bar{L}(\theta_i^m(e)) - \bar{L}(\theta^*(e^*)) \right)
\end{aligned}$$

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 $\Theta(e)$ and the likelihood function converges almost surely for all $\theta(e) \in \Theta(e)$. Therefore, by taking limsup 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:

$$\text{(A.4)} \quad \limsup_{T \rightarrow \infty} \frac{1}{T} Q_{T,e} - \bar{L}(\theta^*(e^*)) \leq \max_i (-I(\theta_i^m(e))) \leq -\min_i (I(\theta_i^m(e))) \leq -\tilde{I}(e^*)$$

which implies that $\frac{1}{T} Q_{T,e} - \bar{L}(\theta^*(e^*))$ converges a.s. $-\tilde{I}(e)$. Using the same arguments and Assumptions 1-4 we can also show that $\frac{1}{T} Q_{T,e^*}$ converges almost surely to $\bar{L}(\theta^*(e^*))$. The rest of this proof is restricted to the event of probability one.

According to Assumption 5 and the fact $\tilde{I}(e) > 0$, for all $e < e^*$ we get

$$\text{(A.5)} \quad Q_{T,e} - Q_{T,e^*} \leq c_{T,e} - c_{T,e^*} \Rightarrow IC_T(e) \leq IC_T(e^*)$$

We conclude that $\liminf_{T \rightarrow \infty} \hat{e} \geq e^*$ \square

Proof of Lemma 2:

The existence of a measurable maximum likelihood estimator $\hat{\theta}_{2T}(e, \theta_1)$ for each $\theta_1(e) \in \Theta_1(e)$ follows from Assumptions 1-2 and theorem 2.12 of White (1994, p.16).

Let $\tilde{L}(\theta_1(e), \theta_2(e)) = E(\log(y_t | w_t; \theta_1(e), \theta_2(e)))$, using Assumption 3'(a) given any neighborhood $\bar{\Theta}_2(e)$ of $\theta_2^*(e)$, there exists an $\varepsilon > 0$ such that,

$$\lim_{T \rightarrow \infty} \inf_{\theta_1(e) \in \Theta_1(e)} \left(\tilde{L}(\theta_1(e), \theta_2^*(e)) - \max_{\theta_2 \in \Theta_2(e) \setminus \bar{\Theta}_2(e)} \tilde{L}(\theta_1(e), \theta_2(e)) \right) \geq \varepsilon > 0$$

Thus,

$$\begin{aligned} \text{(A.6)} \quad & \Pr \left(\hat{\theta}_{2T}(e, \tilde{\theta}_1) \in \Theta_2(e) \setminus \bar{\Theta}_2(e), \text{ for some } \tilde{\theta}_1(e) \right) \\ & \leq \Pr \left(\lim_{T \rightarrow \infty} \inf_{\theta_1(e) \in \Theta_1(e)} \left| \tilde{L}(\theta_1(e), \theta_2^*(e)) - \tilde{L}(\theta_1(e), \hat{\theta}_{2T}(e, \tilde{\theta}_1)) \right| \geq \varepsilon, \text{ for some } \tilde{\theta}_1(e) \right) \\ & \leq \Pr \left(\lim_{T \rightarrow \infty} \sup_{\theta_1(e) \in \Theta_1(e)} \left| \tilde{L}(\theta_1(e), \theta_2^*(e)) - \tilde{L}(\theta_1(e), \hat{\theta}_{2T}(e, \tilde{\theta}_1)) \right| \geq \varepsilon, \text{ for some } \tilde{\theta}_1(e) \right) \end{aligned}$$

Since the event $\{x + y \geq \varepsilon\}$ is contained in $\{x \geq \varepsilon/2\} \cup \{y \geq \varepsilon/2\}$ we get

$$\begin{aligned} & \leq \Pr \left(\lim_{T \rightarrow \infty} \sup_{\theta_1(e)} \left| \tilde{L}(\theta_1(e), \theta_2^*(e)) - \frac{1}{T} L_T(\theta_1(e), \hat{\theta}_{2T}(e, \theta_1)) \right| > \varepsilon/2 \right) + \\ & \leq \Pr \left(\lim_{T \rightarrow \infty} \sup_{\theta_1(e)} \left| \frac{1}{T} L_T(\theta_1(e), \hat{\theta}_{2T}(e, \theta_1)) - \tilde{L}(\theta_1(e), \hat{\theta}_{2T}(e, \theta_1)) \right| > \varepsilon/2 \right) \\ & \leq 2 \cdot \Pr \left(\lim_{T \rightarrow \infty} \sup_{\theta_1(e) \in \Theta_1(e), \theta_2 \in \Theta_2(e)} \left| \tilde{L}(\theta_1(e), \theta_2(e)) - \frac{1}{T} L_T(\theta_1(e), \theta_2(e)) \right| \geq \varepsilon/2 \right) = 0 \end{aligned}$$

The last equality follows from the strong uniform convergence of the likelihood function, which is implied under Assumptions 1, 2, 4, see, e.g. Wooldridge (p. 2651, 1994)). \square

Proof of Lemma 3:

In order to show that $LR_T = O_p(1)$, note that the likelihood ratio is equal to

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

For simplicity we write $\hat{\theta}_{2T}(e)$ instead of $\hat{\theta}_{2T}(e, \theta_1)$. We first try to find the distribution $L_T(\theta_1(e), \hat{\theta}_{2T}(e)) - L_T(\hat{\theta}_T(e^*))$ for a given $\theta_1(e)$. From Assumptions

6(a) and 3'(b) we can use the Taylor series expansion of $\frac{1}{T} D_T(\theta_1(e), \hat{\theta}_2(e))$ around $\theta_2^*(e)$ to get

$$(A.8) \quad 0 = \frac{1}{T} D_T(\theta_1(e), \theta_2^*(e)) + \frac{1}{T} D_T^2(\theta_1(e), \bar{\theta}_2(e)) \cdot (\hat{\theta}_{2T}(e) - \theta_2^*(e))' \\ = \frac{1}{T} D_T(\theta_1(e), \theta_2^*(e)) + E\left(D_T^2(\theta_1(e), \theta_2^*(e))\right) \cdot (\hat{\theta}_{2T}(e) - \theta_2^*(e))' + \xi_T$$

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

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

Since

$$(A.9) \quad \sup_{\theta_1(e)} \xi_T = \sup_{\theta_1(e)} \left\| \frac{1}{T} D_T^2(\theta_1(e), \bar{\theta}_2(e)) - E\left(D_T^2(\theta_1(e), \theta_2^*(e))\right) \cdot (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \right\| \\ \leq \sup_{\theta_1(e)} \left\| \frac{1}{T} D_T^2(\theta_1(e), \bar{\theta}_2(e)) - E\left(D_T^2(\theta_1(e), \theta_2^*(e))\right) \right\| \cdot \sup_{\theta_1(e)} \left\| (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \right\| \\ \leq \sup_{\theta_1(e)} \left\| \frac{1}{T} D_T^2(\theta_1(e), \bar{\theta}_2(e)) - E\left(D_T^2(\theta_1(e), \bar{\theta}_2(e))\right) \right\| \cdot \sup_{\theta_1(e)} \left\| (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \right\| + \\ \sup_{\theta_1(e)} \left\| E\left(D_T^2(\theta_1(e), \bar{\theta}_2(e))\right) - E\left(D_T^2(\theta_1(e), \theta_2^*(e))\right) \right\| \cdot \sup_{\theta_1(e)} \left\| (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \right\| \\ = o(1) \cdot o(1) + o(1) \cdot o(1) = o(1) \text{ a.s.}$$

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

$$(A.10) \quad \sqrt{T}(\hat{\theta}_{2T}(e) - \theta_2^*(e)) = -\left[E\left(D_T^2(\theta_1(e), \theta_2^*(e))\right) \right]^{-1} \frac{1}{\sqrt{T}} D_T(\theta_1(e), \theta_2^*(e))$$

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

$$(A.11) \quad L_T(\theta_1(e), \hat{\theta}_{2T}(e)) - L_T(\theta_1(e), \theta_2^*(e)) = D_T(\theta_1(e), \theta_2^*(e)) \cdot (\hat{\theta}_{2T}(e) - \theta_2^*(e))' + \\ \frac{1}{2} (\hat{\theta}_{2T}(e) - \theta_2^*(e))' \cdot D_T^2(\theta_1(e), \bar{\theta}_2(e)) \cdot (\hat{\theta}_{2T}(e) - \theta_2^*(e)) = \frac{1}{\sqrt{T}} D_T(\theta_1(e), \theta_2^*(e)) \cdot \sqrt{T}(\hat{\theta}_{2T}(e) - \theta_2^*(e))' + \\ \frac{1}{2} \sqrt{T}(\hat{\theta}_{2T}(e) - \theta_2^*(e))' \cdot \left[E\left(D_T^2(\theta_1(e), \theta_2^*(e))\right) \right] \cdot \sqrt{T}(\hat{\theta}_{2T}(e) - \theta_2^*(e)) + \zeta_T$$

where

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

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

get

$$(A.12) \quad 2 \cdot (L_T(\theta_1(e), \hat{\theta}_{2T}(e)) - L_T(\theta_1(e), \theta_2^*(e))) = \mathcal{G}_T + \zeta_T$$

where,

$$\begin{aligned} \zeta_T &= W_T(\theta_1(e)) \left[E(D_t^2(\theta_1(e), \theta_2^*(e))) \right]^{-1} \cdot \left[\frac{1}{T} D_T^2(\theta_1(e), \bar{\theta}_2(e)) - E(D_t^2(\theta_1(e), \theta_2^*(e))) \right] \\ &\cdot \left[E(D_t^2(\theta_1(e), \theta_2^*(e))) \right]^{-1} \cdot W_T(\theta_1(e)); \quad \mathcal{G}_T = W_T(\theta_1(e)) \left[- E(D_t^2(\theta_1(e), \theta_2^*(e))) \right]^{-1} W_T(\theta_1(e)); \end{aligned}$$

In order to show the likelihood ratio converges in distribution, we need to establish that the empirical process $W_T(\theta_1(e))$ is uniformly bounded in probability. Assumptions 1 and 6(c) imply that $D_t(\theta_1(e), \theta_2^*(e))$ is a stationary ergodic martingale difference, to which the central limit theorem is applied pointwise given Assumption 6(b). Hence, from lemma 4.5 of White (2001, p.67)), $W_T(\theta_1(e)) = O_p(1)$ for all $\theta_1(e)$. For $\eta > 0$ and given the compactness of $\Theta_1(e)$, let $\{S(\theta_1^j(e), \eta) \mid j=1, \dots, J\}$ be a finite cover of $\Theta_1(e)$ centered at $\theta_1^j(e)$. Then,

$$\begin{aligned} (A.13) \quad &P\left(\sup_{\theta_1(e) \in \Theta_1(e)} \|W_T(\theta_1(e))\| > \varepsilon\right) \\ &\leq P\left(\max_{j=1, \dots, J} \sup_{\theta_1(e) \in S(\theta_1^j(e))} \|W_T(\theta_1(e)) - W_T(\theta_1^j(e))\| > \varepsilon/2\right) + P\left(\max_{j=1, \dots, J} \|W_T(\theta_1^j(e))\| > \varepsilon/2\right) \\ &\leq P\left(\max_{j=1, \dots, J} \sup_{\theta_1(e) \in S(\theta_1^j(e))} \|W_T(\theta_1(e)) - W_T(\theta_1^j(e))\| > \varepsilon/2\right) + P\left(\bigcup_{j=1}^J \{ \|W_T(\theta_1^j(e))\| > \varepsilon/2 \}\right) \\ &\leq P\left(\max_{j=1, \dots, J} \sup_{\theta_1(e) \in S(\theta_1^j(e))} \|W_T(\theta_1(e)) - W_T(\theta_1^j(e))\| > \varepsilon/2\right) + \sum_{j=1}^J P\left(\|W_T(\theta_1^j(e))\| > \varepsilon/2\right) \end{aligned}$$

Assumption 8 and the pointwise convergence of $W_T(\theta_1(e))$ for each $\theta_1(e) \in \Theta_1(e)$ imply

$$(A.14) \quad \limsup_{T \rightarrow \infty} P\left(\sup_{\theta_1(e) \in \Theta_1(e)} \|W_T(\theta_1(e))\| > \varepsilon\right) < \varepsilon$$

This result and Assumption 7 imply that \mathcal{G}_T is $O_p(1)$ uniformly on $\Theta_1(e)$.

We demonstrate this as follows (A.15)

$$\begin{aligned} \sup_{\theta_1(e)} \|\mathcal{G}_T\| &\leq \sup_{\theta_1(e)} \|W_T(\theta_1(e))\| \cdot \sup_{\theta_1(e)} \left\| \left[-E\left(D_t^2(\theta_1(e), \theta_2^*(e))\right) \right]^{-1} \right\| \cdot \sup_{\theta_1(e)} \|W_T(\theta_1(e))\| \\ &= O_p(1) \cdot O(1) \cdot O_p(1) = O_p(1) \end{aligned}$$

In a similar way we can show that $\sup_{\theta_1(e) \in \Theta_1(e)} \zeta_T = o_p(1)$, hence:

$$(A.16) \quad 2 \cdot \left(\sup_{\theta_1(e) \in \Theta_1(e)} L_T(\theta_1(e), \hat{\theta}_{2T}(e)) - L_T(\theta_1(e), \theta_2^*(e)) \right) = O_p(1) + o_p(1)$$

It's obvious that

$$(A.17) \quad 2 \cdot \left(L_T(\theta^*(e^*)) - L_T(\hat{\theta}_T(e^*)) \right) = O_p(1),$$

See e.g. White (1982, 1994)), because for $e = e^*$ the model parameters are identified.

Therefore,

$$\begin{aligned} (A.18) \quad LR_T &= 2 \cdot \left(\sup_{\theta_1(e) \in \Theta_1(e)} L_T(\theta_1(e), \hat{\theta}_{2T}(e)) - L_T(\hat{\theta}_T(e^*)) \right) \\ &= 2 \cdot \left(\sup_{\theta_1(e) \in \Theta_1(e)} L_T(\theta_1(e), \hat{\theta}_{2T}(e)) - L_T(\theta_1(e), \theta_2^*(e)) \right) + 2 \cdot \left(L_T(\theta^*(e^*)) - L_T(\hat{\theta}_T(e^*)) \right) \\ &= O_p(1) + O_p(1) + o_p(1) = O_p(1). \square \end{aligned}$$

Proof of Theorem 1:

Consider the probability of \hat{e} being greater than e^* :

$$(A.19) \quad \Pr(\hat{e} > e^*) \leq \sum_{e \in E \setminus \bar{E}} \Pr(\hat{e} = e) \leq \sum_{e \in E \setminus \bar{E}} \Pr(IC_T(e) > IC_T(e^*))$$

where for all $e \in E \setminus \bar{E}$

$$(A.20) \quad \Pr\left(IC_T(e) > IC_T(e^*)\right) = \Pr\left(\frac{LR_T}{c_{T,e^*}} > \frac{c_{T,e}}{c_{T,e^*}} - 1\right)$$

From Assumption 5 we know that $\left(\frac{c_{T,e}}{c_{T,e^*}} - 1\right) > 0$ and $\frac{1}{c_{T,e^*}} \rightarrow 0$ and by lemma 3

we have $LR_T = O_p(1)$. Therefore, for all $e \in E \setminus \bar{E}$, $\Pr(\hat{e} = e) = 0$ in probability and using lemma 1, we get that \hat{e} converges to e^* in probability. \square

Proof of Lemma 4:

In order to show that $\limsup_{T \rightarrow \infty} \sup_{\theta_1(e) \in \Theta_1(e)} \frac{LR_T}{\log \log(T)} = O(1)$, note that the

likelihood ratio is equal to:

$$(A.21) \quad LR_T = 2 \cdot \left(\sup_{\theta_1(e) \in \Theta_1(e)} L_T(\theta_1(e), \hat{\theta}_{2T}(e, \theta_1)) - L_T(\hat{\theta}_T(e^*)) \right) \\ = 2 \cdot \left(\sup_{\theta_1(e) \in \Theta_1(e)} L_T(\theta_1(e), \hat{\theta}_{2T}(e, \theta_1)) - L_T(\theta_1(e), \theta_2^*(e)) \right) + 2 \cdot \left(L_T(\theta^*(e^*)) - L_T(\hat{\theta}_T(e^*)) \right)$$

From Assumptions 3'(b), 7(a), and by Taylor's expansion of $D_T(\theta_1(e), \theta_2(e))$ around $\theta_2^*(e)$, we see that:

$$(A.22) \quad \left(\sqrt{T^{-1} \log \log T} \cdot \right)^{-1} (\hat{\theta}_{2T}(e) - \theta_2^*(e)) = \\ - \left[E(D_t^2(\theta_1(e), \theta_2^*(e))) \right]^{-1} \left(\sqrt{T \log \log T} \cdot \right)^{-1} D_T(\theta_1(e), \theta_2^*(e)) \\ + \left(\left[E(D_t^2(\theta_1(e), \theta_2^*(e))) \right]^{-1} - \left[E(D_t^2(\theta_1(e), \bar{\theta}_2(e))) \right]^{-1} \right) \cdot \left(\sqrt{T \log \log T} \cdot \right)^{-1} D_T(\theta_1(e), \theta_2^*(e)) \\ + \left(\left[E(D_t^2(\theta_1(e), \bar{\theta}_2(e))) \right]^{-1} - \left[D_t^2(\theta_1(e), \bar{\theta}_2(e)) \right]^{-1} \right) \cdot \left(\sqrt{T \log \log T} \cdot \right)^{-1} D_T(\theta_1(e), \theta_2^*(e))$$

Taking limsup of both sides, we see that lemma 2, the continuity of the matrix inverse and Assumptions 7 and 9, imply that the second and the third terms converge almost surely to zero, whereas Assumptions 7 and 9 imply that the first term is almost surely bounded and we get:

(A.23)

$$\limsup_{T \rightarrow \infty} \sup_{\theta_1 \in \Theta_1(e)} \sqrt{T^{-1} \log \log T} \cdot (\hat{\theta}_{2T}(e) - \theta_2^*(e)) = O(1)O(1) + o(1)O(1) + o(1)O(1) = O(1)$$

Next, from a Taylor expansion of $L_T(\theta_1(e), \hat{\theta}_{2T}(e))$ around, $\theta_2^*(e)$ for

a given $\theta_1(e)$, we get:

$$(A.24) \quad L_T(\theta_1(e), \hat{\theta}_{2T}(e)) - L_T(\theta_1(e), \theta_2^*(e)) = (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \cdot D_T(\theta_1(e), \theta_2^*(e)) \\ + \frac{1}{2} (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \cdot D_T^2(\theta_1(e), \bar{\theta}_2(e)) \cdot (\hat{\theta}_{2T}(e) - \theta_2^*(e)) = \\ (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \cdot D_T(\theta_1(e), \theta_2^*(e))$$

$$\begin{aligned}
& + \frac{1}{2} \sqrt{T} (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \cdot \left[\frac{1}{T} D_T^2(\theta_1(e), \bar{\theta}_2(e)) - E(D_T^2(\theta_1(e), \theta_2^*(e))) \right] \cdot \sqrt{T} (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \\
& + \frac{1}{2} \sqrt{T} (\hat{\theta}_{2T}(e) - \theta_2^*(e)) \cdot E(D_T^2(\theta_1(e), \theta_2^*(e))) \cdot \sqrt{T} (\hat{\theta}_{2T}(e) - \theta_2^*(e))
\end{aligned}$$

By Assumptions 7 and 9 and lemma 2 and since $(\hat{\theta}_{2T}(e) - \theta_2^*(e)) = O(\sqrt{T^{-1} \log \log T})$ almost surely uniformly in $\Theta_1(e)$, we get:

$$(A.25) \quad \limsup_{T \rightarrow \infty} \sup_{\theta_1(e) \in \Theta_1(e)} (L_T(\theta_1(e), \hat{\theta}_{2T}(e)) - L_T(\theta_1(e), \theta_2^*(e))) = O(\log \log T)$$

We can show in a similar way as in Nishii (1988) or Sin and White (1996) that

$$(A.26) \quad 2 \cdot (L_T(\theta^*(e^*)) - L_T(\hat{\theta}_T(e^*))) = O(\log \log T)$$

hence, $LR_T = O(\log \log T)$. \square

Proof of Theorem 2:

Given lemma 1 it remains to prove that \hat{e} does not overestimate e^* almost surely.

The event that $\hat{e} > e^*$ is given by $\Omega^* = \{\omega \in \Omega \mid \hat{e}(\omega) > e^*\} \subset \bigcup_{e > e^*} \{\omega \in \Omega \mid \hat{e}(\omega) = e\}$.

The event $\bigcup_{e > e^*} \{\omega \in \Omega \mid \hat{e}(\omega) = e\}$ implies $\bigcup_{e > e^*} \{\omega \in \Omega \mid IC_T(e) > IC_T(e^*)\}$, thus for any $e > e^*$ we see that

$$(A.27) \quad Q_{T,e} - c_{T,e} > Q_{T,e^*} - c_{T,e^*} \Leftrightarrow \frac{LR_T}{\log \log(T)} \cdot \frac{\log \log(T)}{c_{T,e^*}} - \frac{c_{T,e}}{c_{T,e^*}} + 1 > 0$$

From Assumptions 5, 10 and lemma 5, $\limsup_{T \rightarrow \infty} \left[\frac{LR_T}{\log \log(T)} \cdot \frac{\log \log(T)}{c_{T,e^*}} - \frac{c_{T,e}}{c_{T,e^*}} + 1 \right]$ tends

almost surely to a strictly negative term.

Therefore, for $e > e^*$, $\Pr(\limsup_{T \rightarrow \infty} IC_T(e) > IC_T(e^*)) = 0$ which implies that

$\Pr\{\Omega^*\} = 0$. We conclude that \hat{e} converges to e^* almost surely. \square

References

Akaike H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* AC-19, 716-723.

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, 241-257.

Andrews D.W.K. (1993). An introduction to econometric applications of empirical process theory for dependent random variables, *Econometric Review*, 12(2), 183-216.

Andrews D.W.K. (1994). *Empirical Process Methods in Econometrics*, In Handbook of Econometrics 4, chapter 37, 2248-2292, New York, North-Holland.

Andrews D.W.K. and Ploberger W. (1994). Optimal tests when the nuisance parameters are present only under the alternative, *Econometrica*, 62, 1383-1414.

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, 247-254.

Davies R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika*, 74, 33-43.

Domowitz I. and White H. (1984). Non-linear regression with dependent observations, *Econometrica*, 52, 143-161.

Feng Z.D. and McCulloch C.E. (1996). Using bootstrap likelihood ratios in finite mixture models, *Journal of the Royal Statistical Society B*, 58(3), 609-617.

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, 173-187.

Hamilton J.D. (1989). A new approach to economic analysis of non-stationary time series and business cycles, *Econometrica*, 57, 357-384.

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-S82.

Hansen B.E. (1996). Inference when the nuisance parameter is not identified under the null hypothesis, *Econometrica*, 64, 413-430.

- Jeffries O.N. (1998). Logistic Mixture of Generalized Linear Model Times Series Ph.D. Dissertation, University of Maryland at College Park, Maryland.
- 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, 409-432.
- Lo Y., Mendell N.R. and Rubin D.B. (2001). Testing the number of components in a normal mixture, *Biometrika*, 88(3), 767-778
- Leroux B.G. (1992). Consistent estimation of a mixing distribution, *The Annals of Statistics*, 20(3), 1350-1360.
- 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, 1143-1148.
- Nishii R. (1988). Maximum likelihood principle and model selection when the true model is unspecified, *Journal of Multivariate analysis*, 27, 392-403.
- Preminger A., Ben-Zion U. and Wettstein D. (2003a). Extended switching regression models: allowing for multiple latent state variables, Working Paper 03-08, Monaster Center for Economic Research, Ben-Gurion University of the Negev.
- Preminger A., Ben-Zion U. and Wettstein D. (2003b). Extended switching regression models with time varying probabilities for combining forecasts, Working Paper 03-13, Monaster Center for Economic Research, Ben-Gurion University of the Negev.
- Rao C.R.(1973). *Linear Statistical inference and its Applications*, New York: John Wiley and Sons.
- Redner R. (1981). Note on the consistency of the maximum likelihood estimate for non-identifiable distributions, *Annals of Statistics*, 9, 225-239.
- Schwarz G. (1978). Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464.
- Sin C.Y. and White H. (1996). Information criteria for selecting possibly misspecified Parametric Models, *Journal of Econometrics*, 71, 207-225.
- Stout W.F. (1970). The Hartman-Winter law of the iterated logarithm for martingale, *The Annals of Mathematical Statistics*, 41(6), 2158-2160.
- Vuong H.Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, 57(2), 307-333.

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.

Wooldridge J.M. (1994). Estimation and inference for dependent processes, in *Handbook of Econometrics 4*, pp. 2641-2700, edited by Engle R.F. and McFadden D.L. Elsevier Science B.V., Amsterdam.