A model selection method for S-estimation

Arie Preminger [a] and Shinichi Sakata [b] [*]

Abstract

In least squares, least absolute deviations, and even generalized M-estimation, outlying observations sometimes strongly influence the estimation result, masking an important and interesting relationship existing in the majority of observations. The S-estimators are a class of estimators that overcome this difficulty by smoothly downweighting outliers in fitting regression functions to data.

In this paper, we propose a method of model selection suitable in S-estimation. The proposed method chooses a model to minimize a criterion named the penalized S-scale criterion (PSC), which is decreasing in the sample S-scale of fitted residuals and increasing in the number of parameters. We study the large sample behavior of the PSC in nonlinear regression with dependent, heterogeneous data, to establish sets of conditions sufficient for the PSC to consistently select the model with the best fitting performance in terms of the population S-scale, and the one with the minimum number of parameters if there are multiple best performers. Our analysis allows for partial unidentifiability, which is often a practically important possibility when selecting one among nonlinear regression models. We offer two examples to demonstrate how our large sample results could be applied in practice. We also conduct Monte Carlo simulations to verify that the PSC performs as our large sample theory indicates, and assess the reliability of the PSC method in comparison with the familiar Akaike and Schwarz information criteria.

Keywords: Robust model selection; partial identification; law of the iterated logarithm

JEL classification: C22, C52

---

[a] Center of Econometrics and Operation Research, University de Catholique, LLN, Belgium.

[b] Department of Economics, University of British Columbia, Vancouver, BC, Canada.

[*] The authors thank Luc Bauwens, Sharon Rubin, and the members of the econometrics group in CORE, Universite' de Catholique as well as the participants of the 15th annual meeting of the Midwest Econometrics Group for their helpful comments. They also appreciate CORE for providing financial support for their collaboration. Preminger grateful for research support from the Ernest Foundation and Sakata acknowledges support from the SSHRC Standard Research Grant Program.

Introduction

In econometric analysis, a data set often contains several outlying observations. These outliers are sometimes grossly mismeasured observations. In other situations, they reflect some exceptional situations, events, or behavior, which the model of interest is not designed to explain. In these cases, a researcher would not prefer that the results of his analysis be driven by the outlying observations. Hence, he would naturally like to employ an estimation method that is sufficiently resistant to outliers. The S-estimators are a class of estimators that meet these needs. In this paper, we develop a method of model selection suitable in S-estimation. The proposed method of model selection itself also tolerates many outliers.

The high sensitivity of the least squares (LS) estimator to outliers is well known in linear regression. The least absolute deviations (LAD) estimator is subject to the same problem, if there are outlying observations of the regressors, which form leverage points. Even the generalized M-estimators designed to attain higher robustness to outliers can only tolerate a small number of outliers. Rousseeuw and Leroy (1987) offer a detailed discussion on this point. Stromberg and Ruppert (1992) and Sakata and White (1995) demonstrate that the LS, LAD, and some other familiar estimators are also sensitive to outliers in nonlinear regression as well as in linear regression. This raises a serious concern, especially in time series analysis, where each outlying observation may appear in the regressor vector multiple times due to the presence of lagged variables, as discussed in Sakata and White (2001).

Hampel (1975) offers a solution to the problem caused by outliers. He suggests that an estimator that minimizes a highly outlier-resistant scale measure of the fitted residuals would be less sensitive to outliers. The least median of squares (LMS) estimator (Rousseeuw 1984), the least trimmed sum of squares (LTS) estimators (Rousseeuw 1983), the S-estimators (Rousseeuw and Yohai 1984), and the τ-estimators (Yohai and Zamar 1988) are implementations of Hampel's idea. Stromberg and Ruppert (1992) and Sakata and White (1995) formally verify that these estimators have good resistance to many outliers in nonlinear regression as well as in linear regression.

Among the estimators with resistance to many outliers, the S-estimators are particularly attractive. Unlike the LTS and LMS estimators, the S-estimators smoothly downweigh outlying observations. This smooth handling of outliers is not only natural from practitioners' point of view, but also it tends to make the S-estimators efficient relative to the LMS estimator. Further, as shown in Sakata and White (2001), the S-estimators are more efficient than the LS estimator under various error distributions. In addition, they are computationally less demanding than the LTS and τ-estimators.

In this paper, we develop a method to select models in S-estimation. Model selection is an important step in S-estimation, as is true for other regression methods. For example, the number of variables that are potentially useful for explaining the dependent variable is often large, even after the irrelevant variables are carefully screened out. Given the sample size, however, the danger of the over-fitting problem usually precludes using all of the potentially relevant regressors. This common situation requires a method to choose variables for the regressors. Besides the variable selection, it is often necessary to choose a functional specification. This

includes selection of the number of regimes in the regime switching models or choice of the number of hidden units in the neural network models. Thus, availability of a reliable model selection method is essential for successful application of S-estimation.

A common approach taken in model selection is to choose the model that maximizes an information criterion (IC). In the LS regression, each selection method of this type chooses a model that minimizes the sum of the logarithm of the sample root mean square error (RMSE) of the fitted residuals and a monotonically increasing function of the number of parameters. The first term in the sum can be interpreted as a goodness-of-fit measure, while the second term penalizes richer models and encourages the use of parsimonious models. Leading examples of the IC's are the Akaike (1974) information criterion (AIC), the Schwarz (1978) information criterion (SIC), and the Hannan and Quinn (1979) information criterion (HQ).

The S-estimator minimizes the sample S-scale of the fitted residuals, while the LS estimator minimizes the sample RMSE. Motivated by this fact, we replace the sample RMSE in the IC's for LS estimation with the sample S-scale. We call the resulting criteria the penalized S-scale criteria (PSC). In the PSC model selection method, we choose the model whose PSC is the minimum among the competing models. In the context of M-estimation, similar approaches have been taken, and the properties of the proposed methods have been studied (Martin 1980, Ronchetti 1985, Tsai 1990, Machado 1993, Ronchetti and Staudte 1994, Qian and Kunsch 1996, Ronchetti et al. 1997). In our approach, each model is estimated by the S-estimator, which not only delivers an outlier-robust estimate of regression parameters, but also provides a scale measure of the fitted residuals that is resistant to many outliers. This feature makes the PSC less sensitive to outlying observations; so, it is natural to expect that the PSC method is outlier-robust. Our Monte Carlo simulations positively verify this conjecture.

The adaptation of IC to generate model selection criteria described above could be applied in other robust scale-minimizing regression such as LMS, LTS, and τ-estimation. One could also adapt the cross validation method, such as the one by Craven and Wahba (1979), in an analogous way, though the application of the cross validation approach in robust scale-minimizing regression is likely to be computationally challenging. In this paper, we focus on the PSC method, leaving the analysis of the other methods for future work.

Given the formulation of the PSC, one might think that its asymptotic properties are covered by the results readily available in the literature; Nishii (1988) Vuong (1989), Sin and White (1996), Kapetanious (2001) to mention a few. Sin and White (1996) establish the consistency of the IC based model selection methods and their variants. Nevertheless, even their general result is not applicable to our problem, due to the fact that the sample S-scale is only defined as an implicit function of the fitted residuals. We therefore study the behavior of the PSC method and establish useful sets of conditions sufficient for its weak and strong consistency.

In selecting one in a collection of nonlinear models, the parameters of the “large” models are often only partially identified. This occurs, for example, when the models have additive nonlinear components, threshold type nonlinearities, or switching parameters. The PSC method is useful in such situations as well. We formally verify that the PSC method consistently selects the best and parsimonious model, even when the parameters of the “larger” models are partially

unidentified. With the partial unidentifiability, regular statistical tests fail to work due to the "flatness" of the objective function. To handle this nonstandard setup, we adapt the framework of Davies (1977, 1987), which treats the statistic of interest as a function of unidentified parameters. A similar approach is also taken by Bierens (1990), Bierens and Ploberger (1997), Hansen (1996a), and Stinchcombe and White (1998) in the context of specification testing. A difficulty in applying a test developed in this framework is that the critical value of the test typically has to be determined through simulations in each application, reflecting the fact that the limiting distribution of the test statistic is nonstandard, and its percentiles cannot be tabulated. Fortunately, the PSC method is not subject to this difficulty, as it requires no critical values.

The rest of this paper is organized as follows. Section 2 introduces the basic setup of our model selection problem, discusses some aspects of the setup in detail, and formally defines the PSC. Section 3 then establishes sets of conditions sufficient for strong and weak consistency of the PSC method, covering a wide range of regression models and allowing for dependent, heterogeneous data generating processes. Section 4 refines the results of section 3, imposing additional conditions that typically hold in the nested model selection. The analysis of section 4 allows for partial identification of the larger model, which makes our result applicable in a wide range of situations. Section 5 offers two examples to demonstrate how the consistency results in sections 3 and 4 apply in practice. Section 6 conducts Monte Carlo simulations to verify that the PSC method performs as our large sample results indicate, and assess the reliability of the PSC method in comparison with the familiar IC method. Finally, section 7 concludes the paper with some remarks. The proofs of the theorems as well as the definitions of some technical terms are collected in the appendices.

2. General Framework

We consider the model selection problem in regression of a variable Y_t on the current and past variables. Our analysis focuses on the selection between two competing models. This is not restrictive, because when the choice is made by linearly ordering a finite number of models, as is the case in the PSC model selection as well as information criterion based selection, the consistency of the model selection method is equivalent to the consistency of the method in the pairwise selection between the best parsimonious model and each of the other models.

Suppose that model $k \in \{1, 2\}$ specifies a regression function $\mu_{k,t}(t = 1, 2, \ldots)$ with a parameter space Θ_k. For each metric space A, let $\mathrm{B}(A)$ denote the Borel σ-field on A and write $\mathrm{B}^v \equiv \mathrm{B}(\Re^v)$ for simplicity.

Assumption 1

a) The series $\{Z_t \equiv (Y_t, X_t')\}_{t \in N}$ is an $\Re^v$-valued stochastic process on a complete probability space $(\Omega, \Im, P)$, where Y_t is 1×1 and X_t is $(v-1) \times 1$.

b) For each $k = 1, 2$, Θ_k is a compact subset of the p_k-dimensional Euclidean space and $\{\mu_{k,t} : \Omega \times \Theta_k \to \Re\}_{t \in \mathrm{N}}$ a sequence of functions measurable-$\Im \otimes \mathrm{B}(\Theta_k)/\mathrm{B}$ such that for each

$t \in \mathrm{N}$ and P-almost all $\omega \in \Omega$, $\mu_{k,t}(\omega,\cdot) : \Theta_k \to \Re$ is continuous. Further, for each $k = 1,2$ and $\theta_k \in \Theta_k$, $\{\Im_{t-1} \equiv \sigma(\ddot{Z}^{t-1})\}_{t \in N}$ is adapted to $\{\mu_{k,t}(\cdot,\theta_k)\}_{t \in \mathrm{N}}$, where $\ddot{Z}^{t-1} \equiv (Z_1, \ldots Z'_{t-1}, X_t')$, $t \in \mathrm{N}$.

The compactness imposed on the parameter space Θ_k is standard and innocuous in many applications, as we can often take a large compact set for it. Because $\Im_{t-1}$ is the σ-field generated by the past variables and the current variables other than Y_t, $\mu_{k,t}(\cdot,\theta_k)$ can be written as a function of the current and past variables (but not of the future variables). It, of course, does not prevent $\mu_{k,t}(\cdot,\theta_k)$ from depending only on the current variables; so, our analysis covers the cross sectional setup as well.

Our goal is to select the model that fits Y_t better than the other. In case the two models fit Y_t equally well, we aim to choose the more parsimonious one. We ignore the possibility that the two models have the same number of parameters and fit Y_t equally well, as it does not matter which model is chosen in such situations.

When we analyze the case in which the two models share the same fitting performance, we assume that the first model is more parsimonious than the second ($p_1 < p_2$). We now introduce the precise notion of better and equal fits in a way suitable in our large sample analysis. Following Rivers and Vuong (2002), we consider two types of relationship between two models. Let $s^*_{k,T}$ denote the (population) measure to assess the goodness-of-fit of the model k. In LS regression, $s^*_{k,T}$ would be the minimum of the square root of the average mean squared residuals, where the minimum is taken over the parameter space, and the average is taken over the sample periods. If $\liminf_{T\to\infty}(s^*_{1,T} - s^*_{2,T}) > 0$, we say that the second model is *asymptotically better* than the first model. If instead $\lim_{T\to\infty}(s^*_{1,T} - s^*_{2,T}) = 0$, we say that the two models are *asymptotically equivalent.* When $(s^*_{1,T} - s^*_{2,T}) = 0$ for each sample size T, the two models are called *equivalent*. Note that when the process is strictly stationary, the notion of the asymptotic equivalence reduces to the equivalence, provided that the models are time homogenous and have no feedback mechanisms such as the one of the autoregressive moving average model, because the scale minimizing parameter values are independent from the sample size in such setup, and so are the population scales.

Our method of model selection in S-estimation, which is described later in this section, is a natural adaptation of the IC based approach in the L_q estimation. The L_q estimator is the regression estimator that picks the parameter value that minimizes the sample q th moment of the fitted residuals, where q is the real number no less than unity. The LS and LAD estimators are L_q estimators with $q = 2$ and $q = 1$, respectively. The q th root of the minimized sample q th moment can be interpreted as the sample scale of the fitted residuals or the goodness-of-fit of the estimated model. The IC based approach of model selection in the L_q estimation leads to the decision rule that picks the model with the smaller criterion value, where this criterion is

obtained by adding a penalty term, increasing in the number of parameters, to the product of the sample size and the logarithm of the minimized sample scale of the fitted residuals.

The scale measures employed in the L_q estimators are very sensitive to outliers. A measure of the sensitivity of estimators to outliers is the *breakdown point*, which is defined to be the minimum proportion of the data for which contamination by outliers can lead to completely non-informative estimation results. Several measures of the breakdown point have been suggested, notably by Hampel (1968, 1971), Donoho and Huber (1983), Stromberg and Ruppert (1992), and Sakata and White (1995). Under all measures, it has been shown that the L_q estimators for regression models tend to be easily affected by a small number of outliers. More specifically, these estimators are shown to have a breakdown point of $1/T$, where T is the sample size. This means that a single bad observation can already cause a breakdown. Furthermore, Sakata and White (2001) show that a small change in the underlying distribution can cause a large change in the L_q scale measure; so, it may also have a significant impact on the model selection in L_q regression.

On the other hand, the S-scales are highly resistant to outlying observations, as mentioned above. The estimators based on these scale measures have a high breakdown point accordingly. We now formally introduce our method to select models in S-estimation. Our method employs the penalized S-scale criterion (PSC) defined by

$$PSC_{k,T} \equiv T \log \hat{S}_{k,T} + c_{k,T}, \tag{1}$$

where $\hat{S}_{k,T}$ denotes the S-estimation objective function of model k, evaluated with the S-estimator, namely, the sample S-scale of the fitted residuals; and $c_{k,T}$ is the penalty term for model k. We allow $c_{k,T}$ to be random in our analysis, though all applications we consider in this paper take nonstochastic penalty terms. This criterion takes into account the trade-off between the complexity and the fitting performance of each model. Our selection method chooses the model $\hat{k}_T$ with the lower PSC. The penalty term can be a sequence of numbers such as $c^{p}_{k,T} = p_k$, $c^{L}_{k,T} = 0.5 p_k \log T$, and $c^{LL}_{k,T} = d \cdot p_k \log\log T$ with $d > 1$, which are analogs of the familiar penalty terms of the AIC, SIC and HQ, respectively.

For later discussion, we now formally define the S-estimators. The construction of an S-estimator requires a function ρ from $\Re$ to $\Re_+$, the set of all nonnegative numbers, that satisfies a set of conditions stated in the next assumption.

Assumption 2

The function $\rho : \Re \to \Re_+$ is even, bounded, twice continuously differentiable on $\Re$, and its derivative ρ' is positive at every positive point where ρ has not achieved its supremum $\overline{\rho}$.
Let, $\Re_{++} \equiv (0,+\infty)$, and define $h^k_T : \Theta_k \times \Re_{++} \times \Omega \to \Re$ by

$$h_{k,T}(\theta_k, s, \omega) \equiv \tfrac{1}{T}\sum_{t=1}^{T} \rho\left(\frac{Y_t(\omega) - \mu_{k,t}(\omega, \theta_k)}{s}\right), \qquad (\theta_k, s, \omega) \in \Theta_k \times \Re_{++} \times \Omega$$

and $\bar{h}_{k,T} : \Theta_k \times \Re_{++} \to \Re$ by

$$\bar{h}_{k,T}(\theta_k, s) \equiv \tfrac{1}{T}\sum_{t=1}^{T} E\left[\rho\left(\frac{Y_t - \mu_{k,T}(\theta_k)}{s}\right)\right], \qquad (\theta_k, s) \in \Theta_k \times \Re_{++}.$$

The S-estimation objective function is then:

$$S_{k,T}(\theta_k, \omega) \equiv \inf\left\{s \in \Re_{++} : h_{k,T}(\theta_k, s, \omega) < M\right\}, \tag{2}$$

where $M \in (0, \tfrac{1}{2}\bar{\rho})$. The random vector $\hat{\theta}_{k,T}$ that minimizes this function over Θ_k is called the S-estimator associated with (ρ, M). The breakdown point of the S-estimator in cross sectional regression is approximately equal to $M/\bar{\rho} \in (0, \tfrac{1}{2})$; so, they can resist up to $T/2$ outliers. In the case with time series data, the regressors often include lagged variables, so that a single outlier may appear multiple times in the regressor vector. Suppose p is the maximum lag order of the regression function for each $t \in \mathrm{N}$, Sakata and White (1998, 2001) show that a lower bound for the breakdown point is approximately $M/(\bar{\rho}(1+p))$. The smaller the maximum lag order is, the higher this lower bound is.

The population counterpart of $S_{k,T}$ is $\bar{S}_{k,T} : \Theta_k \to \Re_+$ is defined by

$$\bar{S}_{k,T}(\theta_k) \equiv \inf\left\{s \in \Re_+ : \bar{h}_{k,T}(\theta_k, s) < M\right\}. \tag{3}$$

The function $\bar{S}_{k,T}$, being continuous on the compact space Θ_k, attains its minimum $s^*_{k,T}$ on Θ_k (see Lemma 5.1 of Sakata and White (2001)). In the following sections, we use the uniform laws of large numbers (ULLN), the law of iterated logarithm (LIL), and the concept of stochastic equicontinuity (SE). The precise definitions of these terms are found in appendix A.

3. The consistency of the PSC method

In this section, we establish useful sufficient conditions for the weak and strong consistency of our model selection method. Notice that

$$PSC_{1,T} - PSC_{2,T} = T(\log \hat{S}_{1,T} - \log \hat{S}_{2,T}) + (c_{1,T} - c_{2,T}) \tag{4}$$

$$= T[(\log s^*_{1,T} - \log s^*_{2,T}) + (\log \hat{S}_{1,T} - \log s^*_{1,T}) - (\log \hat{S}_{2,T} - \log s^*_{2,T})] + (c_{1,T} - c_{2,T})$$

Suppose that the second model is asymptotically better than the first. If the second and third terms inside the square brackets on the right-hand side converge to zero a.s.-P (prob-P), as one naturally expects, and if the penalty terms are chosen in such a way that the fourth term is $o_{a.s.}(T)$ ($o_P(T)$), the difference in the penalized S-scale criterions is dominated by $T(\log s^*_{1,T} - \log s^*_{2,T})$, so that the PSC should tend to pick the second model, once the sample size grows large. If, instead, the first model is asymptotically better than the second, the same argument suggests that the PSC would be likely to pick the first model, when the sample size is large. On the other hand, when the two models are asymptotically equivalent and $p_1 < p_2$ (so that $c_{1,T} < c_{2,T}$), and the PSC method should tend to pick the more parsimonious first model, as we desire, if the order of magnitude of the penalty term asymptotically dominates that of $T(\log \hat{S}_{1,T} - \log \hat{S}_{2,T})$.

To formally establish the consistency of the PSC method, we first make the following assumptions.

Assumption 3

For each $k = 1, 2$:

a) The set $\Theta^*_{k,T}$ of minimizers of $\bar{S}_{k,T}$ is identifiable, i.e. for all $\delta > 0$,

$\liminf_{T\to\infty} \inf_{\theta^k \in \eta^c_{k,T}(\delta)} [\bar{S}_{k,T}(\theta_k) - \bar{S}_{k,T}(\theta^*_{k,T})] > 0$, where

$\eta^c_{k,T}(\delta) \equiv \{\theta_k \in \Theta_k : \inf_{\theta^*_{k,T} \in \Theta^*_T(k)} \left\| \theta_k - \theta^*_{k,T} \right\| \geq \delta\}$ ($\|\cdot\|$ is the Euclidean norm).

b) $\lim_{\alpha\to 0} \sup_{t\in T} \sup_{\theta_k \in \Theta_k} P\left(|Y_t - \mu_t(\cdot, \theta_k)| \leq \alpha\right) < 1 - (M/\bar{\rho})$.

c) Y_t and $\mu_{k,t}(\cdot, \theta_k)$ are P-integrable uniformly in $t \in T$ and $\theta_k \in \Theta_k$.

Assumption 4

For each $k = 1, 2$ and each $\bar{s}_1, \bar{s}_2 \in \Re_{++}$ such that $\bar{s}_1 \leq \bar{s}_2$:

a) $\{\bar{h}_{k,T}\}_{T\in \mathrm{N}}$ is equicontinuous in $(\theta_k, s) \in \Theta_k \times [\bar{s}_1, \bar{s}_2]$.

b) $\left\{\rho\left(\frac{Y_t - \mu_{k,t}(\cdot, \theta_k)}{s}\right)\right\}_{t\in N}$ obeys the uniform weak law of large numbers (UWLLN) on $\Theta_k \times [\bar{s}_1, \bar{s}_2]$.

Assumption 3 allows $\Theta^*_{k,T}$ to contain more than one point. This generality is useful in certain applications, as discussed later. The identifiability of $\Theta^*_{k,T}$ is implied by the identifiability of the set of minimizers of $\bar{h}_{k,T}(\cdot, s^*_{k,T})$ (see Lemma 5.2 of Sakata and White (2001)). The second

condition in assumption 3 requires that the probability of good fit is not too high and does not have an upward trend in t, while the third condition requires that the distributions of Y_t should not become more and more dispersed as t grows larger. These conditions and assumption 2 above ensure that the function $\bar{S}_{k,T}$ is "identifiable". Application of a suitable uniform strong law of large numbers (USLLN) (Ranga Rao (1962) and Jennrich (1969)) typically shows that assumption 4 is automatically satisfied, when the DGP is i.i.d. or stationary ergodic, due to the continuity and boundedness of ρ. Under heterogeneous and possibly dependent DGP's, the UWLLN can be established, by using the results of Andrews (1987, 1992). Assumptions 3-4 guarantee that for each $k \in \{1,2\}$, $\{\hat{S}_{k,T} - s^*_{k,T}\}_{T\in\mathrm{N}}$ converges in probability to zero. This property gives us a way to establish that the PSC method is weakly consistent when one model is asymptotically better than the other.

To cover the case with two asymptotically equivalent models, we examine the stochastic order of $\{\log(\hat{S}_{1,T} / \hat{S}_{2,T})\}_{T\in\mathrm{N}}$ for each $k \in \{1,2\}$, which we use for finding the growth rate of the penalty terms that ensures the consistency of the PSC method. If the regression functions $\mu_{k,t}(\omega,\cdot) : \Theta_k \to \Re$ are differentiable for P-almost all $\omega \in \Omega$, for each $k \in \{1,2\}$, the S-estimator $\hat{\theta}_{k,T}$ and the minimized sample S-scale $\hat{S}_{k,T}$ satisfy that

$$\Psi_{k,T}(\hat{\theta}_{k,T}, \hat{S}_{k,T}, \cdot) \equiv \tfrac{1}{T}\sum_{t=1}^{T} \psi_{k,t}(\hat{\theta}_{k,T}, \hat{S}_{k,T}, \cdot) = 0, \tag{5}$$

as shown in Sakata and White (2001), where $\Psi_{k,T}$ is the generalized score function; and $\psi_{k,t} \equiv (\psi_{k,1,t}, \psi_{k,2,t})$,

$$\psi_{k,1,t}(\theta_k, s, \omega) \equiv \rho'\left(\frac{Y_t(\omega) - \mu_{k,t}(\omega,\theta_k)}{s}\right)\nabla_\theta \mu_{k,t}(\omega,\theta_k), \tag{6}$$

and

$$\psi_{k,2,t}(\theta_k, s, \omega) \equiv \rho\left(\frac{Y_t(\omega) - \mu_{k,t}(\omega,\theta_k)}{s}\right) - M \tag{7}$$

for each $\theta_k \in \Theta_k$, each $s \in \Re_{++}$, and each $\omega \in \Omega$, (here, we employ the rule that the derivative of a non-differentiable function is zero).

Assumption 5

For each $k = 1,2$ and each $\bar{s}_1, \bar{s}_2 \in \Re_{++}$ such that $\bar{s}_1 \le \bar{s}_2$:

a) For almost all $T \in \mathrm{N}$, $\Theta^*_{k,T}$ contains only one element $\theta^*_{k,T}$ that is interior to Θ_k uniformly in $T \in \mathrm{N}$.

b) For each $t \in \mathrm{N}$, $\mu_{k,t}(\omega,\cdot) : \Theta_k \to \Re$ is twice continuously differentiable on Θ_k a.s.- P.

c) $\{V_{k,T} \equiv \mathrm{var}(T^{1/2}\Psi_{k,T}(\theta^*_{k,T}, s^*_{k,T},\cdot))\}_{T\in\mathrm{N}}$ is uniformly positive definite, and $\left\{E(\nabla'(\Psi_{k,T}(\theta^*_{k,T}, s^*_{k,T},\cdot)))\right\}_{T\in\mathrm{N}}$ is uniformly nonsingular, where ∇' denotes the Jacobian operator with respect to (θ_k, s).

d) $V_{k,T}^{-1/2}T^{1/2}\Psi_{k,T}(\theta^*_{k,T}, s^*_{k,T},\cdot) \overset{A}{\sim} N(0, I)$.

e) The families of functions,

$$\left\{\frac{1}{T}\sum_{t=1}^{T} E\left[\rho''\left(\frac{Y_t - \mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot\nabla_{\theta_k}\mu_{k,t}(\cdot,\theta_k)\cdot\nabla_{\theta_k}\mu(\cdot,\theta_k)'\right]\right\}_{T\in\mathrm{N}},$$

$$\left\{\frac{1}{T}\sum_{t=1}^{T} E\left[\rho'\left(\frac{Y_t - \mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot\nabla_{\theta_k\theta_k}\mu_{k,t}(\cdot,\theta_k)\right]\right\}_{T\in\mathrm{N}},$$

$$\left\{\frac{1}{T}\sum_{t=1}^{T} E\left[\rho''\left(\frac{Y_t - \mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot(Y_t - \mu_{k,t}(\cdot,\theta_k))\nabla_{\theta_k}\mu_{k,t}(\cdot,\theta_k)\right]\right\}_{T\in\mathrm{N}},$$

$$\left\{\frac{1}{T}\sum_{t=1}^{T} E\left[\rho'\left(\frac{Y_t - \mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot\nabla_{\theta_k}\mu_{k,t}(\cdot,\theta_k)\right]\right\}_{T\in\mathrm{N}},$$

and

$$\left\{\frac{1}{T}\sum_{t=1}^{T} E\left[\rho'\left(\frac{Y_t - \mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot(Y_t - \mu_{k,t}(\cdot,\theta_k))\right]\right\}_{T\in\mathrm{N}}$$

are equicontinuous in $(\theta_k, s) \in \Theta_k \times [\bar{s}_1, \bar{s}_1]$.

f) The sequences of random functions of (θ_k, s),

$$\left\{\rho''\left(\frac{Y_t - \mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot\nabla_{\theta_k}\mu_{k,t}(\cdot,\theta_k)\cdot\nabla_{\theta_k}\mu(\cdot,\theta_k)'\right\}_{t\in\mathrm{N}},$$

$$\left\{\rho'\left(\frac{Y_t - \mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot\nabla_{\theta_k\theta_k}\mu_{k,t}(\cdot,\theta_k)\right\}_{t\in\mathrm{N}},$$

$$\left\{\rho''\left(\frac{Y_t-\mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot(Y_t-\mu_{k,t}(\cdot,\theta_k))\nabla_{\theta_k}\mu_{k,t}(\cdot,\theta_k)\right\}_{t\in\mathrm{N}},$$

$$\left\{\rho'\left(\frac{Y_t-\mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot\nabla_{\theta_k}\mu_{k,t}(\cdot,\theta_k)\right\}_{t\in\mathrm{N}},$$

and

$$\left\{\rho'\left(\frac{Y_t-\mu_{k,t}(\cdot,\theta_k)}{s}\right)\cdot(Y_t-\mu_{k,t}(\cdot,\theta_k))\right\}_{t\in\mathrm{N}},$$

obey the UWLLN on $\Theta_k\times[\bar{s}_1,\bar{s}_1]$.

Unlike assumption 3(a), assumption 5(a) requires that $\theta^*_{k,T}$ is identifiably unique. Nevertheless, this assumption is sometimes violated, and $\Theta^*_{k,T}$ may contain finitely many elements. For example, the smooth transition autoregressive (STAR), neural network, and switching regression models have this problem due to "label switching". Fortunately, this problem can be overcome by merging all parameter values that yield the same regression function into a single equivalence class (see e.g. Leroux (1992) and Redner (1981)) and taking the set of such equivalence classes for $\Theta_{k,T}$. 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 the details of this remedy. Assumption 5(c) imposes a mild smoothness condition on the regression function. Assumption 5(d) is a high-level assumption, which can be verified under more primitive assumptions by using a variety of central limit theorems (CLT), such as the Lindeberger-Feller CLT (White (2001), p.117) for the i.i.d. processes and the double array CLT for near epoch dependent (NED) processes on a strong mixing processes by Wooldridge (1986). Assumptions 5(e) and 5(f) can be verified by applying the UWLLN as mentioned in the paragraph following assumption 4. Sakata and White (2001) also provide a set of primitive conditions that ensures the CLT and UWLLN assumptions in the case with a process NED on a strong mixing process.

Let $\hat{k}_T$ denote the model selected by the PSC method , i.e., $\hat{k}_T=\arg\min_{k\in\{1,2\}}PSC_T(k)$.

Proposition 1:

Suppose assumptions 1–4 hold. Then:
(i) If model 2 is asymptotically better than model 1, and if $T^{-1}c_{k,T}\to_P 0$, *then* $\hat{k}_T\to_P 2$.

(ii) Suppose in addition that assumption 5 holds, and that $T^{-1/2}(c_{2,T}-c_{1,T})\to_P\infty$.
If $(s^*_{1,T}-s^*_{2,T})=O(T^{-1/2})$, *then* $\hat{k}_T\to_P 1$.

Proposition 1(i) shows that the PSC with penalty terms that grows more slowly than T tends to select the asymptotically better model, once the sample size becomes large. This result does not require that the set of minimizers $\Theta_{k,T}^{*}$ be a singleton. When the models are asymptotically equivalent, the conditions imposed in proposition 1(ii) ensure that $T\log(\hat{S}_{2,T} - \log\hat{S}_{1,T}) = O_p(T^{1/2})$, which is asymptotically dominated by the difference in the penalty terms growing faster than $T^{1/2}$, so that the more parsimonious model is selected with a probability approaching one as the sample size becomes large. The conditions imposed upon the penalty terms in proposition 1 are satisfied, if we pick $c_{k,T} = p_k(T\log\log T)^{1/2}$ or $c_{k,T} = p_k(T\log T)^{1/2}$, for example.

It is important to note that the requirement in proposition 1(ii) that $s_{1,T}^{*} - s_{2,T}^{*} = O(T^{-1/2})$ is in general a stronger condition than the asymptotic equivalence between the two models. If $s_{1,T}^{*} - s_{2,T}^{*}$ converges to zero more slowly than $T^{-1/2}$, the requirement in proposition 1(ii) is not met, despite that the two models are asymptotically equivalent. Depending on the growth rate of $c_{2,T} - c_{1,T}$, the PSC method may not be consistent in this case. When the data generating process (DGP) is stationary, however, this problem never happens, provided that the models are time-homogeneous and have no feedback mechanism, because the asymptotic equivalence reduces to the equivalence in such setup, as discussed in section 2.

We now turn to investigation of the strong consistency of the PSC method. If one model is asymptotically better than the other, the strong consistency holds as long as $\hat{S}_{k,T} - s_{k,T}^{*}$ converges to zero a.s. and $c_{k,T} = o_{a.s.}(T)$. For the case with two asymptotically equivalent models, we apply the LIL to obtain almost sure bounds for $T(\log\hat{S}_{1,T} - \log\hat{S}_{2,T})$ and find what growth rate for the difference between the penalty terms leads to the strong consistency of the PSC method. The next assumption strengthens the requirements in assumptions 4(b) and 5(f) along this line.

Assumption 6

For each $k = 1,2$ and each $\bar{s}_1, \bar{s}_2 \in \Re_{++}$ such that $\bar{s}_1 \le \bar{s}_2$:

a) $\{\psi_{k,t}(\theta_{k,T}^{*}, s_{k,T}^{*}, \cdot)\}_{t\in\mathrm{N}}$ obeys the LIL.

b) $\left\{\rho\left(\frac{Y_t(\omega) - \mu_{k,t}(\omega,\theta)}{s}\right)\right\}_{t\in\mathrm{N}}$ obeys the USLLN on $\Theta_k \times [\bar{s}_1, \bar{s}_2]$.

c) The sequences of random functions of (θ_k, s) in assumption 5(f) obeys the USLLN on $\Theta_k \times [\bar{s}_1, \bar{s}_2]$.

Assumption 6(a) could be verified by applying an appropriate LIL such as the Harman-Winter LIL for i.i.d. (Davidson (1994), p.408) or the LIL for martingale difference processes (Hall and Heyde (1980)). Assumptions 6(b) and 6(c) hold if the sequences in question obey the pointwise strong law of large numbers and are strongly stochastically equicontinuous. The strong

stochastic equicontinuity can be verified in various ways. See Andrews (1987) and Gallant and White (1988) for details.

Proposition 2:

Suppose assumptions 1–3, 4(a), and 6(b) hold.

(i) If model 2 is asymptotically better than model 1, and if $c_{k,T} = o_{a.s.}(T)$, *then* $\hat{k}_T \to_{a.s.} 2$.

(ii) Suppose in addition that assumptions 5(a, b, c, e) and 6(a, c) hold, and that $(T \log\log T)^{-1/2}(c_{2,T} - c_{1,T}) \to_{a.s.} \infty$. *If* $(s^*_{1,T} - s^*_{2,T}) = O((T^{-1}\log\log T)^{1/2})$, *then* $\hat{k}_T \to_{a.s.} 1$.

We can pick $c_{k,T} = p_k (T \log T)^{1/2}$, for example, to yield a strongly consistent PSC method. As is the case for the weak consistency of the PSC method, the strong consistency may not follow, if the models are asymptotically equivalent and $T^{1/2}(\log\log T)^{-1/2}(s^*_{1,T} - s^*_{2,T})$ diverges as the sample size grows to infinity, which might happen in the absence of strict stationarity.

4. Model selection when some of the regression parameters are unidentified

In the previous section, a key assumption in establishing the consistency of the PSC method in the case with two asymptotically equivalent models is that $\Theta^*_{1,T}$ and $\Theta^*_{2,T}$ are singletons. For certain types of models, this condition is necessarily violated. Suppose, for instance, that the first model is a linear regression model and the second model is a STAR model (Luukkonen et al. 1988) that nests the first model. When the two models can reach the same population fitting performance, the nonlinear part of the second model contains unidentified parameters. The exponential autoregressive models (Ozaki (1985)) and the neural network models (Kuan and White (1994)) may be also partially unidentified in similar ways.

.

In this section, we modify assumptions 5 and 6 to allow the second model to be partially unidentified. Because, propositions 1(i) and 2(i) do not require that $\Theta^*_{1,T}$ and $\Theta^*_{2,T}$ be singletons, this change of the assumptions does not affect the weak and strong consistency of the PSC method in the case in which one model is asymptotically better than the other. We therefore focus on the case with two asymptotically equivalent models. Suppose that the parameter vector of the second model is partitioned into two subvectors as $\theta_2 = (\theta_2^1, \theta_2^2) \in \Theta_2^1 \times \Theta_2^2 \equiv \Theta_2$, where θ_2^1 consists of the parameters that are not identified and θ_2^2 contains the other parameters.

Assumption 5'

a) The conditions imposed in assumption 5 hold for $k = 1$.

b) $\Theta_2 = \Theta_2^1 \times \Theta_2^2$ for some compact sets $\Theta_2^1 \subset \Re^{p_2^1}$ and $\Theta_2^2 \subset \Re^{p_2^2}$ (so that $p_2^1 + p_2^2 = p_2$).

c) For almost all $T \in \mathrm{N}$ and each $\theta_2^1 \in \Theta_2^1$, $\bar{h}_{2,T}(\theta_2^1, \cdot, s_{2,T}^*) : \Theta_2^2 \to \Re_+$ is uniquely minimized at a point $\theta_{2,T}^+(\theta_2^1)$ interior to Θ_2^2 uniformly in $\theta_2^1 \in \Theta_2^1$ and $T \in \mathrm{N}$. Also, $\{\theta_{2,T}^+(\theta_2^1)\}_{T \in \mathrm{N}}$ is identifiable uniformly in $\theta_2^1 \in \Theta_2^1$ in the sense that for each real number $\delta > 0$.

$$\liminf_{T \to \infty} \inf\{\bar{S}_{2,T}(\theta_2^1, \theta_2^2) - \bar{S}_{2,T}(\theta_2^1, \theta_2^+(\theta_2^1)) : \| \theta_2^2 - \theta_2^+(\theta_2^1) \| > \delta, \theta_2^1 \in \Theta_2^1\} > 0.$$

d) Assumptions 5(b, e, f) hold for $k = 2$.

e) $\left\{ V_{2,T}^2(\theta_2^1) \equiv \mathrm{var}[T^{1/2} \Psi_{2,T}^2(\theta_2^1, \theta_{2,T}^+(\theta_2^1), s_{2,T}^*, \cdot)] \right\}_{T \in N}$ is bounded and positive definite, where the function $\Psi_{2,T}^2$ is the vector consisting of the last $(p_2^2 + 1)$ elements of $\Psi_{2,T}$ (which corresponds to (θ_2^2, s)). In addition, $\left\{ E[\nabla' \Psi_{k,T}^2(\theta_2^1, \theta_{2,T}^+(\theta_2^1), s_{2,T}^*, \cdot)] \right\}_{T \in \mathrm{N}}$ is nonsingular uniformly in $T \in \mathrm{N}$ and $\theta_2^1 \in \Theta_2^1$, where ∇' denotes the Jacobian operator with respect to (θ_2^2, s).

f) For each $\theta_2^1 \in \Theta_2^1$, $V_{2,T}^2(\theta_2^1)^{-1/2} T^{1/2} \Psi_{2,T}^2(\theta_2^1, \theta_{2,T}^+(\theta_2^1), s_{2,T}^*, \cdot) \overset{A}{\sim} N(0, I)$ as $T \to \infty$.

g) $\left\{ T^{1/2} \Psi_{2,T}^2(\theta_2^1, \theta_{2,T}^+(\theta_2^1), s_{2,T}^*, \cdot) \right\}_{T \in \mathrm{N}}$ is stochastically equicontinuous in $\theta_2^1 \in \Theta_2^1$

h) For $k = 1,2$ $\bar{s}_1, \bar{s}_2 \in \Re_{++}$ such that $\bar{s}_1 \le \bar{s}_2$,

$$\left\{ \tfrac{1}{T} \sum_{t=1}^{T} E\left[\rho'\left(\frac{Y_t - \mu_{k,t}(\cdot, \theta_k)}{s} \right) \cdot (Y_t - \mu_{k,t}(\cdot, \theta_k))^2 \right] \right\}_{T \in \mathrm{N}} \text{ is equicontinuous in } (\theta_k, s) \in \Theta_k \times [\bar{s}_1, \bar{s}_1].$$

i) $\left\{ \rho'\left(\frac{Y_t - \mu_{k,t}(\cdot, \theta_k)}{s} \right) \cdot (Y_t - \mu_{k,t}(\cdot, \theta_k))^2 \right\}_{t \in \mathrm{N}}$ obeys the UWLLN on $\Theta_k \times [\bar{s}_1, \bar{s}_2]$.

Assumption 6'

a) The conditions imposed in assumption 6 hold for $k = 1$.

b) $\left\{ \psi_{2,t}^2(\theta_2^1, \theta_{2,T}^+(\theta_2^1), s_{2,T}^*, \cdot) \right\}_{t \in \mathrm{N}}$ satisfies the LIL for each $\theta_2^1 \in \Theta_2^1$.

c) $\left\{ (T^{-1} \log\log T)^{-1/2} \Psi_{2,T}^2(\theta_2^1, \theta_{2,T}^+(\theta_2^1), s_{2,T}^*, \cdot) \right\}_{T \in \mathrm{N}}$ is strongly stochastic equicontinuous in $\theta_2^1 \in \Theta_2^1$.

d) Assumptions 5(b, e) and 6(b, c) holds for $k = 2$.

e) For each $k = 1,2$ and each $\bar{s}_1, \bar{s}_2 \in \Re_{++}$ such that $\bar{s}_1 \le \bar{s}_2$,

$$\left\{ \tfrac{1}{T} \sum_{t=1}^{T} \rho'\left(\frac{Y_t - \mu_{k,t}(\cdot, \theta_k)}{s} \right) \cdot (Y_t - \mu_{k,t}(\cdot, \theta_k))^2 \right\}_{T \in \mathrm{N}} \text{ obeys the USLLN on } \Theta_k \times [\bar{s}_1, \bar{s}_2].$$

Assumption 5'(c) modifies assumption 5(a) to ensure the identifiability of the second set of parameters uniform in the first set of parameters. Assumptions 5'(f) and 6'(b) can be verified using an appropriate CLT and LIL, respectively. In order to establish assumption 5'(g), we can apply Hansen's (1996b) results for martingale difference, strong mixing, and NED processes. These results are in particular suited for Lipschitz smooth functions of the unidentified parameters. The stochastic equicontinuity property is also available for other types of functions; see Andrews (1993, 1994) and numerous references cited therein. The strong stochastic equicontinuity requirement stated in assumption 6'(c) will be met if the generalized score function is differentiable almost surely at each point of Θ_2^1, and the Jacobian vector of the score function with respect to θ_1^2 is bounded uniformly over Θ_2^1 by the LIL (see e.g. Andrews (1992) and Altissimo and Corradi (2002)). Assumptions 5'(h,i) and 6'(e) are high-level assumptions, which can be verified using a suitable ULLN, as discussed earlier.

In order to establish the consistency of the PSC method, we use assumption 5' to show that the sample S-scales employed in the S-estimation are bounded in probability. In addition, assumption 6' allows us to form almost sure bounds for these sample scales. These results, combined with suitable conditions on the penalty term, imply that model 1 tends to be selected, once the sample size becomes large.

Proposition 3:

Suppose that assumptions 1–3, 4(a), and 5'(b) hold . Also, assume that there exists $\{\theta_{2,T}^{\oplus} \in \Theta_2^2\}_{T\in \mathrm{N}}$ such that $\mu_{1,t}(\cdot,\theta_{1,T}^{*}) = \mu_{2,t}(\cdot,\theta_2^1,\theta_{2,T}^{\oplus})$ for each t, $T \in \mathrm{N}$ and each $\theta_2^1 \in \Theta_2^1$.

(i) Suppose in addition that assumptions 4(b) and 5'(a, c-i) hold, and that $\sup_{\theta_2^1\in\Theta_2^1} \| \theta_{2,T}^{+}(\theta_2^1) - \theta_{2,T}^{\oplus} \| = O(T^{-1/2})$. If $c_{2,T} - c_{1,T} \to_P \infty$, then $\hat{k}_T \to 1$ in probability.

(ii) Suppose instead that assumptions 5(a, b, c, e) hold for $k = 1$, that assumptions 5''(c, e, h) and 6' hold, and that $\sup_{\theta_2^1\in\Theta_2^1} \| \theta_{2,T}^{+}(\theta_2^1) - \theta_{2,T}^{\oplus} \| = O(T^{-1/2}(\log\log T)^{1/2})$. If $(\log\log T)^{-1}(c_{2,T} - c_{1,T}) \to_{a.s.} \infty$, then $\hat{k}_T \to 1$ a.s. - P.

The conditions in propositions 3(i) and 3(ii) imply that the $T\log(\hat{S}_{2,T} - \log\hat{S}_{1,T})$ is $O_P(1)$ and $O_{a.s.}(\log\log T)$, respectively, and it tends to be dominated by the difference between the penalty terms in large samples. The conditions imposed on the penalty terms in proposition 3 are milder than those in propositions 1 and 2 and satisfied by $c_{k,T}^{L}$ introduced above to obtain strong consistency of selection. In proposition 3, we require an additional condition on the regression functions. The refined conditions reflect the assumptions on the relationship between the two models imposed in proposition 3, which is typically satisfied when one model is strictly nested in the other.

The full identifiability of the first model in proposition 3 is employed for simplicity. It is possible to modify proposition 3 to cover the PSC selection between two partially unidentified models, if we impose an appropriate structure on the relationship between the two models, so that the two models approach each other in a suitable way under the asymptotic equivalence. We, however, do not pursue this extension in this paper. Proposition 3 readily covers the PCS selection between two *fully identified* models. To appreciate this point, take any fully identified model and add to it a parameter that does not appear anywhere in the model. This, of course, does not change the model itself at all. The added parameter is, however, unidentified by construction. With this modified parameterization, the model is now only partially identified. Thus, we can apply proposition 3 to the PSC model selection between two fully identified models, by using this modified parameterization to the second model. The result obtained in this way is formally stated in proposition 4 below.

Proposition 4:

Suppose that assumptions 1–3 and 4(a) hold. Also, assume that there exists $\{\theta_{2,T}^{\oplus} \in \Theta_2\}_{T\in\mathrm{N}}$ *such that* $\mu_{1,t}(\cdot,\theta_{1,T}^{*}) = \mu_{2,t}(\cdot,\theta_{2,T}^{\oplus})$ *for each* $t, T \in \mathrm{N}$.

(i) Suppose in addition that assumptions 4(b), 5, and 5'(h,i) hold, and that $c_{2,T} - c_{1,T} \to_P \infty$. *If* $\left\|\theta_{2,T}^{*} - \theta_{2,T}^{\oplus}\right\| = O(T^{-1/2})$, *then* $\hat{k}_T \to 1$ *in probability.*

(ii) Suppose instead that assumptions 5(a, b, c, e), 5'(h), 6, and 6'(e) hold, and that $\left\|\theta_{2,T}^{*} - \theta_{2,T}^{\oplus}\right\| = O(T^{-1/2}(\log\log T)^{1/2})$. *If* $(\log\log T)^{-1}(c_{2,T} - c_{1,T}) \to_{a.s.} \infty$, *then* $\hat{k}_T \to 1$ *a.s. - P*

5. Some Applications

In this section, we demonstrate how the large sample results established in Sections 3 and 4 can be applied in practice, by analyzing two examples of the model selection problem. In the first example, we consider the choice of the number of regimes in a STAR model, where some of the parameters are possibly unidentified under the true model. In the second example, we examine selection of the lag order in an autoregressive (AR) model.

The STAR model belongs to the class of regime switching models and it is widely used for modeling nonlinearity in time series (see, e.g., Granger and Terasvirta (1993) and Terasvirta (1994)). It has been, however, pointed out that the presence of outliers may lead to spurious detection of regimes (see e.g. Balke and Fomby (1994) and Dijk et al. (1999)). The PSC model selection method is useful for avoiding this problem. The STAR model with h regimes is given by the following equation:

$$y_t = [\beta_{11} y_{t-1} + \ldots \beta_{p,1} y_{t-p}] + \sum_{j=2}^{h} G(\gamma_{0j} + \gamma_{1j} s_t) \cdot [\beta_{1j} y_{t-1} + \ldots \beta_{p,j} y_{t-p}] + \varepsilon_t . \tag{8}$$

The variable s_t is called the transition variable. In our example, we set $s_t = y_{t-1}$. The set of

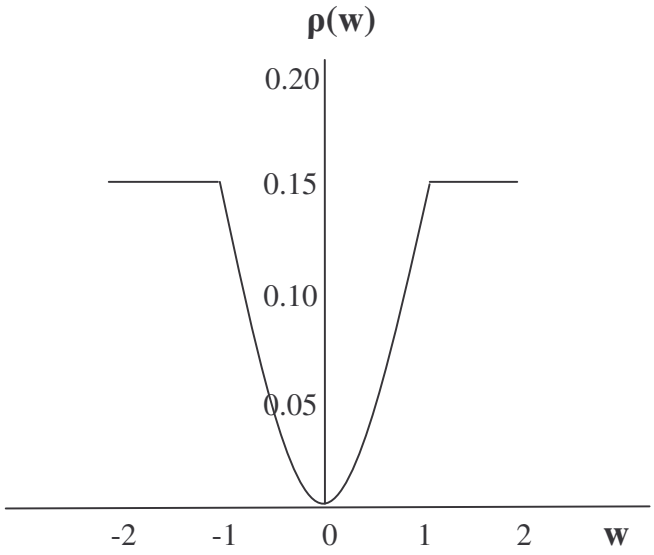

Figure 1: The 'Tukey' function ($\overline{\rho} = \frac{1}{6}, c = 1$).

models being considered is indexed by $h \in \{1,2,3\}$, while the true model has two regimes. We assume that these models are estimated by an S-estimator with the function ρ defined by

$$\rho(w) \equiv \begin{cases} \dfrac{w^2}{2} - \dfrac{w^4}{2c^2} + \dfrac{w^6}{6c^4} & for \quad |w| \le c, \\ \dfrac{c^2}{6} & for\ |w| > c, \end{cases} \tag{9}$$

which can be obtained by integrating Tukey's biweight function (Rousseeuw and Yohai (1984) and Beaton and Tukey (1974)). This function ρ is plotted in Figure 1, setting $c = 1$.

The model is the linear autoregression for $h = 1$. Since the true process has two regimes, our third model is only partially identified, having a redundant regime. We can view the parameters γ_{03}, γ_{13} as the unidentified parameters, while the other parameters are identified. Thus, the setup of section 4 applies to the current problem. The number of regimes is chosen by minimizing the PSC:

$$\hat{h}_T = \arg\min_{h \in \{1,2,3\}} \{PSC(h)\}. \tag{10}$$

We now state a set of conditions, under which we establish the strong consistency of the PSC method in selecting the number of regimes.

(A1) For $h = 1, 2$, assumptions 5(a,c) hold; and for $h = 3$ assumptions 5'(b,c,e) hold in the above-mentioned setup; and the parameter set of each model is compact.

(A2) The true DGP is a STAR process ($h_0 = 2$) initiated in the infinite past, and the error term ε_t is i.i.d. with a density continuous, symmetric about the origin, and positive on the entire real line. Assume that $E(|\varepsilon_t|^{\kappa}) < \infty$ for $\kappa \geq 6$, and $0 \leq G(u) \leq 1$ for each $u \in \Re$.

(A3) The function G is twice continuously differentiable on $\Re$, and for some real numbers $\overline{\alpha}_1$ and $\overline{\alpha}_2$, $\sup_{u\in\Re} |G'(u)| \leq \overline{\alpha}_1$ and $\sup_{u\in\Re} |G''(u)| \leq \overline{\alpha}_2$.

(A4) The ρ is as in (9).

(A5) All zeros of the polynomial $B(\lambda) = \lambda^p - \sum_{i=1}^{p} d_i \lambda^{p-i}$ are in the open unit disk, where $d_i = \sup_{z\in[0,1]} |\beta_{i1} + \beta_{i2} z|$, for $i = 1, \ldots, p$.

(A6) The penalty term satisfies that $c_{h,T} = o_{a.s}(T)$ and $\left(c_{3,T} - c_{2,T}\right) \big/ \log\log(T) \to_{a.s.} +\infty$.

Applying the standard results in the Markov chain theory with these assumptions proves that the STAR process with two regimes (the true model) is stationary, ergodic, and regular mixing with an exponential decay rate. Note that assumption (A.2) may be weakened to allow for an stationary error term that is a function of some i.i.d. variables. For example, we can consider an ARCH or GARCH error process (see e.g. Zhang et al. (2001)). Given these results, the smoothness of the regression function and the moment conditions imposed in (A2) allow us to establish the strong stochastic equicontinuity required for the generalized score function as well the LIL and other regularity conditions. Thus, we can apply propositions 2(i) and 3(ii) to show the strong consistency.

Proposition 5: *Suppose conditions A1-A6 hold. Then* $\hat{h}_T \to h_0$ *a.s.*

Next, we consider the lag order selection in an autoregressive (AR) model, which is given by

$$y_t = \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \varepsilon_t . \tag{11}$$

Various methods have been developed for selection of the lag order of the AR model; see, for example, Hannan and Quinn (1979), Lutkepohl (1985), and Koreisha and Pukkila (1993). It is known that the common selection criteria such as the SIC and HQ are not robust to outliers. For example, Basci and Zaman (1998) establishes that excess kurtosis, which can be interpreted as generating uncontaminated outlying innovations, may affect various model selection criteria. Raftery, and Martin (1996) also argue that additive outliers downward bias the standard selection criteria, showing that under an AR(1) process with heavy contamination, the SIC tends to select the AR(0) model instead of the true AR(1) model. Given these observations, a researcher may desire to use the PSC method, rather than the popular IC method.

The set of AR models being considered in our example have the lag orders $p \in \{1,2,3\}$ and the true lag order is $p_0 = 2$. A set of conditions sufficient for strong consistency of the correct lag order is provided below:

(B1) Assumptions 5(a,c) hold in the above-mentioned setup; and the parameter set of each model is compact.

(B2) The true DGP is a AR process ($p_0 = 2$) initiated in the infinite past and the error term ε_t is i.i.d. with a density continuous and symmetric about the origin and positive on the entire real line. Assume $E(|\varepsilon_t|^{\kappa}) < \infty$ for $\kappa \geq 2$.

(B3) The ρ is as in (9).

(B4) The polynomial $\mathrm{P}(\lambda) = \lambda^{p_0} - \sum_{i=1}^{p_0} \beta_i \lambda^{p-i}$ has all its zeros in the open unit disk.

(B5) The penalty term satisfies that $c_{p,T} = o_{a.s}(T)$ and $\left(c_{3,T} - c_{2,T}\right) / \log\log(T) \to_{a.s.} +\infty$.

As in the first example, we can show that the true process is stationary, ergodic, and β-mixing with an exponential decay rate. Using this result, we can also establish the other regularity conditions employed in propositions 2(i) and 4(ii) and prove the strong consistency.

Proposition 6: *Suppose conditions B1-B5 hold. Then the PSC-selected order* $\hat{p}_T \to p_0$ *a.s.*

We do not include the proof of proposition 6, because it is very similar to that of proposition 5. It is available from the authors upon request.

6. Monte Carlo simulations

In this section, we conduct Monte Carlo simulations to compare the PSC method and the information criteria combined with the LS estimator. Because the latter methods are familiar, this comparison should provide insights on the behaviour of the PSC method.

In each of our simulations, we consider a stationary AR process $\{y_t\}_{t\in \mathrm{N}}$ that satisfies the stochastic difference equation

$$y_t = \theta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + u_t \ , \tag{12}$$

or a stationary STAR process that satisfies

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 G(\gamma_0 + \gamma_1 y_{t-1}) y_{t-1} + u_t \ , \tag{13}$$

Table 1: The distributions of ε_t

Name	Distribution
Norm	Standard Normal
Laplace	Laplace distribution
T5	Student T-distribution with 5 degrees of freedom
MixedNorm2	The distribution obtained by mixing N(0, 1) and N(0, 9) with probabilities 0.8 and 0.2, respectively.
MixedNorm4	The distribution obtained by mixing N(0, 1) and N(0, 9) with probabilities 0.6 and 0.4, respectively.

where $\theta_0 = 0$, $\theta_1 = \theta_2 = 0.25$, $\beta_0 = 0$, $\beta_1 = 0.25$, $\beta_2 = 1$, $\gamma_0 = 1$, and $\gamma_1 = 1$; $\{u_t\}_{t\in\mathrm{N}}$ is a stationary and ergodic process described below; and $G : \Re \to \Re$ is a logistic distribution function shifted downwards by 0.5, that is $G(v) = -0.5 + (1 + \exp(-v))^{-1}$. The process $\{u_t\}_{t\in\mathrm{N}}$ is an ARCH process with a unit variance such that $u_t = (\alpha_0 + \alpha_1 u_{t-1}^2)^{1/2}\varepsilon_t$, where α_0 and α_1 are positive real constants, and $\{\varepsilon_t\}_{t\in\mathrm{N}}$ is an univariate i.i.d. process such that for each $t \in \mathrm{N}$, ε_t and $(y_{t-1}, y_{t-2}, \ldots)$ are independent. We set $\alpha_0 = 1$ and $\alpha_1 = 0$ (conditionally homoskedastic innovations) in some of our simulations and $\alpha_0 = 0.5$, $\alpha_1 = 0.5$ (conditionally heteroskedastic innovations) in others. The process $\{\varepsilon_t\}_{t\in\mathrm{N}}$ is generated by random draws from the distributions listed in Table 1.

The candidate models from which a model is selected are the AR(p) models with p ranging from one to five in the simulations with the AR DGP. In the simulations with the STAR DGP, we take for the candidate models the AR(1) model, the STAR model of (13), and the two models obtained by adding one and two more nonlinear terms, respectively, where the nonlinear terms are the function of y_{t-1} in the form of the third term on the right-hand side of (13). Note that the STAR experiments offer the situations in which the two largest models contain unidentified parameters.

The sample sizes used in our simulations are 200 and 1000. The former is of piratical relevance, being in the range of typical sample sizes in time series applications using monthly observations. The latter is of interest, because it not only can arise in using daily observations, but it also gives us a way to verify the large sample behaviour of the PSC method, which is studied in sections 3-4. The model selection is performed based on the LS estimator and the S-estimators with the ρ function of (9), taking 0.1, 0.3, and 0.5 for $M/\overline{\rho}$, which are respectively denoted S10, S30, and S50.

In each of our simulations, u_t is symmetrically distributed conditionally given $y_{t-1}, y_{t-2}, \ldots$. In such a setup, the population scales of the fitted residuals behind the S10, S30, S50, and LS estimators are minimized by the same parameter value. The optimal model with the minimum complexity is thus common across the four estimators and coincides with the simulation DGP (Sakata and White (2001), Theorem 2.1). This makes meaningful our comparison among the model selectors based on the S- and LS estimators.

Table 2: The standard error of the estimated probability of correction selection.

True Probability of Correct Selection (%)	Standard Error
50	3.16
60	3.10
70	2.90
80	2.53
90	1.90
95	1.38

Since our simulations involve repeated S-estimation, it is computationally demanding. Reflecting this fact, we set the number of replications in each of our simulations equal to 250. We summarize the simulation results in tables of the probability of selecting the "true model". To inform the reader of the magnitude of the errors involved in the figures in the tables, Table 2 shows the standard error of the probability of selecting the "true model" *estimated* in the simulation under a few selected true probabilities of the correct selection.

Among various model selection methods using information criteria, the AIC and SIC are most widely used in practice. As described in Section 2, we can construct the corresponding PSC selectors by choosing p and $0.5p\log T$ for the penalty term, respectively, where p is the number of parameters in the model. Let PSC-A and PSC-S respectively denote them. It is well known that AIC is not a consistent model selector. Because the PSC-A violates the conditions imposed in proposition 3, it is not surprising to see that PSC-A also lacks consistency. Our simulation confirms this inconsistency of PSC-A in all experiments. As an example, we provide Table 3, which shows the probability of correct selection by AIC and PSC-A in the AR experiments with conditionally homoskedastic innovations. It is clear that the probability stays away from 100% regardless of the sample size. (The results for AIC and PSC-A in the other experiments are available from the authors upon request).

We now focus on SIC and PSC-S. Tables 4–7 show the probability that each of the four selectors chooses the true model, which provides very clear pictures on the behaviour of PSC-S. First, the probability of the correct selection becomes higher when the sample size goes up. This is true for each of the four selectors and in all experiments, as our large sample theory predicts. Second, the PSC-S methods with S10 and S30 perform equally well or better than the SIC based on LS (taking into account the standard error in Table 2), and the PSC-S sometimes wins by a large margin. Third, PSC-S with S30 does not necessarily outperform that with S10, even when the innovation distribution has fat tails (MixedNorm2 and MixedNorm4). This may reflect the trade-off between the efficiency and the outlier robustness. Fourth, the performance of PSC-S with S50 is very poor in some cases, in particular when the sample size is small.

Given the simulation results, it is fair to say that the PSC-S method in the S10 and S30 regression is at least as trustworthy as the SIC method in the LS regression. Also, stability of the PSC-S with S10 and S30 can be viewed as demonstrating the outlier-robustness of the PSC-S method. On the other hand, a large sample size may be required in order for the PSC-S method to

Table 3: The probabilities (%) that the AIC and PSC-A select the true model in the AR experiments with conditionally homoskedastic innovations

Distribution Sample Estimators of ε_t	Sample Size	Estimators			
		LS	S10	S30	S50
Norm	200	76.8	76.0	54.4	18.4
	1000	76.4	78.8	60.4	30.8
Laplace	200	73.6	76.4	60.8	30.4
	1000	75.6	75.6	63.6	40.4
T5	200	76.4	77.6	64.0	25.6
	1000	77.6	74.8	54.8	30.8
MixedNorm2	200	72.0	75.2	60.0	20.8
	1000	77.2	78.4	63.6	31.6
MixedNorm4	200	76.8	75.2	58.8	24.4
	1000	73.6	75.6	61.2	40.0

Note: The column labeled LS shows the probability that AIC selects the true model, while S10, S30, and S50 gives the probability that PSC-A selects the true model based on the S10, S30, and S50 estimates, respectively.

Table 4: The probabilities (%) that the SIC and PSC-S select the true model in the AR experiments with conditionally homoskedastic innovations.

Distribution Sample Estimators of ε_t	Sample Size	Estimators			
		LS	S10	S30	S50
Norm	200	86.8	86.0	75.6	39.6
	1000	98.4	98.0	95.2	77.6
Laplace	200	82.4	90.4	89.2	68.4
	1000	98.8	99.2	96.8	89.2
T5	200	86.0	91.2	86.8	64.4
	1000	98.4	98.8	96.4	76.8
MixedNorm2	200	82.0	93.6	94.0	62.4
	1000	99.6	99.6	97.2	82.8
MixedNorm4	200	82.4	90.4	92.8	66.8
	1000	99.2	99.2	96.8	83.6

Note: The column labeled LS shows the probability that SIC selects the true model, while S10, S30, and S50 gives the probability that PSC-S selects the true model based on the S10, S30, and S50 estimates, respectively.

Table 5: The probabilities (%) that the SIC and PSC-S select the true model in the AR experiments with conditionally heteroskedastic innovations.

Distribution Sample Estimators of ε_t	Sample Size	Estimators			
		LS	S10	S30	S50
Norm	200	78.4	80.8	70.8	44.4
	1000	96.8	98.0	94.4	75.6
Laplace	200	68.4	84.8	85.2	68.4
	1000	92.4	98.4	98.0	86.8
T5	200	74.0	85.2	80.0	54.4
	1000	92.8	98.0	92.8	76.0
MixedNorm2	200	72.8	88.0	89.2	65.6
	1000	94.4	99.2	95.6	76.4
MixedNorm4	200	72.4	90.0	91.2	64.4
	1000	91.6	98.8	95.2	80.0

Note: See the note of Table 4.

Table 6: The probabilities (%) that the SIC and PSC-S select the true model in the STAR experiments with conditionally homoskedastic innovations

Distribution Sample Estimators of ε_t	Sample Size	Estimators			
		LS	S10	S30	S50
Norm	200	53.2	49.2	50.4	47.6
	1000	100.0	100.0	100.0	94.8
Laplace	200	72.4	83.6	92.0	94.8
	1000	100.0	100.0	100.0	98.0
T5	200	63.6	74.0	81.6	78.0
	1000	100.0	100.0	100.0	98.8
MixedNorm2	200	83.2	94.4	95.6	89.2
	1000	99.6	100.0	100.0	99.6
MixedNorm4	200	75.2	85.2	97.6	97.2
	1000	100.0	100.0	100.0	99.6

Note: See the note of Table 4.

Table 7: The probabilities (%) that the SIC and PSC-S select the true model in the STAR experiments with conditionally heteroskedastic innovations

Distribution Sample Estimators of ε_t	Sample Size	Estimators			
		LS	S10	S30	S50
Norm	200	52.8	44.8	34.8	33.6
	1000	79.6	97.6	97.2	90.4
Laplace	200	59.2	68.8	73.2	77.2
	1000	74.0	98.4	99.2	96.0
T5	200	57.6	57.6	54.8	51.2
	1000	70.8	99.2	97.6	94.4
MixedNorm2	200	65.2	72.4	72.0	59.6
	1000	70.4	96.4	97.6	95.2
MixedNorm4	200	64.0	72.4	78.8	72.4
	1000	69.6	98.0	96.4	96.4

Note: See the note of Table 4.

perform reliably, when the S50 estimator is used. This again seems to reflect the well-known efficiency-robustness trade-off in S-estimation.

7. Conclusion

In this paper, we propose a method of model selection suitable in S-estimation. The proposed method chooses a model to minimize a criterion named the penalized S-scale criterion (PSC), which is decreasing in the sample S-scale of the fitted residuals and increasing in the number of parameters.

We study the large sample behaviour of the PSC in nonlinear regression with dependent, heterogeneous data, to establish sets of conditions sufficient for the PSC to consistently select the model with the best fitting performance in terms of the population S-scale, and the one with the minimum number of parameters if there are multiple best performers. Our large sample analysis allows the models to be partially unidentified. We offer two examples to demonstrate how our large sample results could be applied in practice. We also conduct Monte Carlo simulations to verify that the PSC performs as our large sample analysis predicts.

Our simulation results indicate that the PSC method has good resistance to outlying observations, in addition to its ability to select the simplest, best performing model in terms of the S-scale. This feature is very useful for a researcher who uses the S-estimation method to find an interesting relationship in the majority of observations, avoiding the strong effects of outliers. It is, however, important to note that, in general, the stochastic limit of the LS and S estimators are different. It is senseless to use the PSC method with the LS estimator of regression parameters, because the scales used in the PSC and the LS estimation are not compatible. We should first decide which estimation method is the most appropriate for the purpose of the research project and than apply a model selector suitable for it, rather than focus on which selection method performs better under special conditions.

The large sample results for the PSC method established in this paper are essentially the same as those for the familiar information criteria found in Sin and White (1996). This is not surprising, given the smoothness of the objective function in S-estimation and the similar appearances of the formulas of the PSC and information criteria. Nevertheless, the similarity between the PSC and information criteria is not an indication that our analysis is redundant. The objective function in S-estimation is implicitly defined through an equation dependent on data, unlike the objective functions in M-estimation, which are covered by Sin and White's framework. This nature of S-estimation adds substantial complexity and requires a separate treatment. Also, our analysis extends the problem setup of Sin and White, by allowing for partially identified models. This extension makes our results applicable to a substantially wider range of regression models.

The framework of our large sample analysis that allows for possibly partially unidentified models may be used to extend Sin and White's analysis of information criteria. Also, it is possible to use the framework of our analysis to study the behaviour of the PSC method with non-stationary or long memory data, by using recent developments in the large sample theory for S-estimators (see Lucas (1995) and Sibbertsen (1999, 2001)). These topics are beyond the scope of this paper and are left for future research.

Appendix A (Definitions):

Throughout the following definitions, assumption 1 is imposed, and Γ is a compact set

a) Uniform Weak Law of Large Numbers (UWLLN)

A sequence $\{q_t : \Omega \times \Gamma \to R^m\}$ is said to obey the UWLLN on Γ if

$$\sup_{\gamma \in \Gamma} \left\| \frac{1}{T} \sum_{t=1}^{T} \left(q_t(\cdot, \gamma) - E(q_t(\cdot, \gamma)) \right) \right\| = o_P(1).$$

b) Uniform Strong Law of Large Numbers (USLLN)

A sequence $\{q_t : \Omega \times \Gamma \to R^m\}$ is said to obey the USLLN on Γ if

$$\sup_{\gamma \in \Gamma} \left\| \frac{1}{T} \sum_{t=1}^{T} \left(q_t(\cdot, \gamma) - E(q_t(\cdot, \gamma)) \right) \right\| = o_{a.s.}(1).$$

c) Law of Iterated Logarithm (LIL)

A sequence $\{q_t : \Omega \times \Gamma \to R^m\}$ is said to satisfy the LIL for all $\gamma \in \Gamma$ if

$$\limsup_{T \to \infty} \left\| \sum_{t=1}^{T} \left(q_t(\cdot, \gamma) - E(q_t(\cdot, \gamma)) \right) \right\| = O_{a.s.}\left(\sqrt{T \log\log(T)} \right).$$

d) Stochastic Equicontinuity

A sequence $\{Q_T : \Omega \times \Gamma \to R^m\}$ is said to be stochastically equicontinuous on Γ

if for all $\varepsilon > 0$ and $\eta > 0$, there exists $\delta > 0$ such that

$$\limsup_{T \to \infty} P\left(\sup_{\gamma \in \Gamma} \sup_{\tilde{\gamma} \in O(\gamma, \delta)} \left\| Q_T(\cdot, \gamma) - Q_T(\cdot, \tilde{\gamma}) \right\| > \eta \right) < \varepsilon,$$

where $O(\gamma, \delta)$ is a closed sphere with a radius $\delta > 0$ centered at γ.

e) Strongly Stochastic Equicontinuity

A sequence $\{Q_T : \Omega \times \Gamma \to R^m\}$ is said to be strongly stochastically equicontinuous on Γ

if $\limsup_{T \to \infty} \sup_{\gamma \in \Gamma} \sup_{\tilde{\gamma} \in O(\gamma, \delta)} \left\| Q_T(\cdot, \gamma) - Q(\cdot, \tilde{\gamma}) \right\| \to 0$ as $\delta \to 0$ a.s.

Appendix B

Proof of Proposition 1

i) Given assumptions 1-4, $\{\hat{S}_{k,T} - s^*_{k,T}\}_{T\in\mathrm{N}}$ converges in probability to zero (Sakata and White (2001), Theorem 5.3). For each $k = 1,2$ there exists an interval $I = [\bar{s}_l, \bar{s}_u]$ where $\bar{s}_l$ and $\bar{s}_u$ are in $\Re_{++}$ such that $s^*_{k,T} \in I$ for any $T \in \mathrm{N}$ (Sakata and White (2001), Lemma 5.1(a)). Hence, there exists a real number $\Lambda_1 > 0$ such that

$$\text{(A.1)} \qquad \left|\log \hat{S}_{k,T} - \log s^*_{k,T}\right| \le \Lambda_1 \cdot \left|\hat{S}_{k,T} - s^*_{k,T}\right|$$

with a probability approaching to one as T grows to infinity. Since model 2, is asymptotically better than model 1, we also have that $\Delta = \liminf_{T\to\infty}(\log s^*_{1,T} - \log s^*_{2,T}) > 0$. It follows from (A.1) and equation (4) that

$$\text{(A.2)} \quad PSC_{1,T} - PSC_{2,T} \ge T\left(\Delta + (\log \hat{S}_{1,T} - \log s^*_{1,T}) - (\log \hat{S}_{2,T} - \log s^*_{2,T}) + T^{-1}(c_{1,T} - c_{2,T})\right)$$

$$\ge T(\Delta + o_P(1)).$$

The desired result therefore follows.

ii) Under Assumptions 1–5, $\{\hat{S}_{k,T}\}_{T\in\mathrm{N}}$ is $\sqrt{T}$ -consistent for $\{s^*_{k,T}\}_{T\in N}$ (Sakata and White (2001), Theorem 6.1), where $\{s^*_{k,T}\}_{T\in N}$ is bounded away from zero, as discussed above. It follows by (A.1) that $(\log \hat{S}_{k,T} - \log s^*_{k,T}) = O_P(T^{-1/2})$. Analogously, the assumption that $(s^*_{1,T} - s^*_{2,T}) = O(T^{-1/2})$ yields that $(\log s^*_{1,T} - \log s^*_{2,T}) = O_P(T^{-1/2})$. By applying these results in (4), we obtain that

$$\text{(A.3)} \qquad PSC_{2,T} - PSC_{1,T} = (c_{2,T} - c_{1,T}) + O_P(T^{1/2}).$$

If we dividing both sides of this equality by $T^{1/2}$ the first-term on the right hand side (RHS) diverges to infinity by hypothesis, while the second term is $O_P(1)$. Thus, $PSC_{2,T} - PSC_{1,T}$ is positive with a probability converging to one as $T \to \infty$. The desired result therefore follows.

Proof of Proposition 2

i) The proof is the same as that of proposition 1(i), except that (A.1) holds for almost all $T \in \mathrm{N}$ a.s.-P and the remainder term in (A.2) is replaced with $o_{a.s.}(1)$ because under the given assumptions $\{\hat{S}_{k,T} - s^*_{k,T}\}_{T\in\mathrm{N}}$ converges to zero a.s.-P (Sakata and White (2001), Theorem 5.3).

ii) The standard linearization technique (e.g., White (1994), Theorem 6.10) used to establish the asymptotic normality of the S-estimator and the objective function in the S-estimation (i.e., $(\hat{\theta}_{k,T}, \hat{S}_{k,T})$) in our context) shows that there exists a real number $\Delta > 0$ such that $\left| \hat{S}_{k,T} - s^*_{k,T} \right| \leq \Delta \left| \Psi_{k,T}(\theta^*_{k,T}, s^*_{k,T}, \cdot) \right|$ a.s.-P (for example, we can set Δ equal to 2 times the operator norm of $E[\nabla \Psi_{k,T}(\theta^*_{k,T}, s^*_{k,T}, \cdot)]^{-1}$). We thus have by assumption 6(a) $| \hat{S}_{k,T} - s^*_{k,T} | = O_{a.s.}(T^{-1/2}(\log\log T)^{1/2})$, from which it further follows by (A.1) that $| \hat{S}_{k,T} - s^*_{k,T} | = O_{a.s.}(T^{-1/2}(\log\log T)^{1/2})$. Since, $(s^*_{1,T} - s^*_{2,T}) = O((T \log\log T)^{-1/2})$ we can also show analogously that $| \log s^*_{1,T} - \log s^*_{2,T} | = O((T \log\log T)^{-1/2})$. By applying these results in (4), we obtain that

$$PSC_{2,T} - PSC_{1,T} = (c_{2,T} - c_{1,T}) + O_{a.s.}((T \log\log T)^{1/2}). \tag{A.4}$$

If we dividing both sides of this equality by $(T \log\log T)^{1/2}$ the first-term on the RHS diverges to infinity by hypothesis, while the second term is $O_{a.s.}(1)$. Thus, $PSC_{2,T} - PSC_{1,T}$ is positive a.s.-P The desired result therefore follows.

Proof of Proposition 3

i) If $T(\log \hat{S}_{2,T} - \log \hat{S}_{1,T}) = O_P(1)$, then $PSC_{2,T} - PSC_{1,T} = (c_{2,T} - c_{1,T}) + O_P(1)$, so that $PSC_{2,T} - PSC_{1,T}$ is positive with a probability approaching to one, provided that $c_{2,T} - c_{1,T} \to \infty$ is assumed in this proposition. Thus, it suffices to show that $T(\log \hat{S}_{2,T} - \log \hat{S}_{1,T}) = O_P(1)$.

By the definition of $\{\theta^{\oplus}_{2,T}\}_{T \in \mathrm{N}}$ we have that for each $\theta^1_2 \in \Theta^1_2$, $S_{2,T}(\theta^1_2, \theta^{\oplus}_{2,T}) = S_{1,T}(\theta^*_{1,T})$; in particular $S_{2,T}(\hat{\theta}^1_2, \theta^{\oplus}_{2,T}) = S_{1,T}(\theta^*_{1,T})$, where $\hat{\theta}^1_{2,T}$ is the vector consisting of the first p_1 elements of $\hat{\theta}_{2,T}$. Using this fact we obtain that

$$T(\log \hat{S}_{2,T} - \log \hat{S}_{1,T}) = T(\log \hat{S}_{2,T} - \log S_{2,T}(\hat{\theta}^1_2, \theta^{\oplus}_{2,T})) - T(\log \hat{S}_{1,T} - \log S_{1,T}(\theta^*_{1,T})). \tag{A.5}$$

As Sakata and White (2001, Lemma 5.1(a, f)) state, for each $k = 1,2$ $\{S_{k,T}(\theta_k) - \overline{S}_{k,T}(\theta_k)\}_{T \in \mathrm{N}}$ is convergent in probability to zero uniformly in $\theta_k \in \Theta_k$, and $\overline{S}_{k,T}(\theta_k)$ is positive uniformly in $\theta_k \in \Theta_k$ and $T \in \mathrm{N}$. As the logarithmic function is Lipschitz on any closed interval in $\Re_{++}$, it follows that $T(\log \hat{S}_{2,T} - \log \hat{S}_{1,T}) = O_P(1)$ if both $T(\hat{S}_{2,T} - S_{2,T}(\hat{\theta}^1_2, \theta^{\oplus}_{2,T}))$ and $T(\hat{S}_{1,T} - S_{1,T}(\theta^*_{1,T}))$ are $O_P(1)$. In what follows, we only show that $T(\hat{S}_{2,T} - S_{2,T}(\hat{\theta}^1_2, \theta^{\oplus}_{2,T}))$ is $O_P(1)$ because the proof for $T(\hat{S}_{1,T} - S_{1,T}(\theta^*_{1,T}))$ is analogous and simpler.

For each $T \in \mathrm{N}$, each $\theta_2^1 \in \Theta_2^1$ and each $\omega \in \Omega$, let $\tilde{\theta}_{2,T}^2(\theta_2^1, \omega)$ be the random vector that minimizes $S_{2,T}(\theta_2^1, \theta_2^2, \cdot)$ in terms of θ_2^2 on Θ_2^2. For simplicity we write $\tilde{\theta}_{2,T}^2$ and $\theta_{2,T}^+$ instead of $\tilde{\theta}_{2,T}^2(\theta_2^1, \omega)$ and $\theta_{2,T}^+(\theta_2^1)$, respectively. Also let

$$\text{(A.6)} \qquad \tilde{S}_{2,T}^2(\theta_2^1, \omega) \equiv S_{2,T}(\theta_2^1, \tilde{\theta}_{2,T}^2, \omega), \qquad (\theta_2^1, \omega) \in \Theta_2^1 \times \Omega, \quad T \in \mathrm{N}.$$

Under assumptions 1-4 and 5'(b,c) and, repeating the proofs of Lemma 5.1 and 5.2 in Sakata and White (2001) establishes the identifiability of $\theta_{2,T}^+$ uniformly in $\theta_2^1 \in \Theta_2^1$, namely, for each $\delta > 0$

$$\text{(A.7)} \qquad \liminf_{T \to \infty} \inf_{\theta_2^1 \in \Theta_2^1} \inf_{\theta_2^2 \in \eta_T^c(\delta, \theta_2^2)} (\bar{S}_{2,T}(\theta_2^1, \theta_2^2) - s_{2,T}^*) > 0,$$

where $\eta_T^c(\delta, \theta_2^2) \equiv \left\{ \theta_2^2 \in \Theta_2^2 : \left\| \theta_2^2 - \theta_{2,T}^+ \right\| \right\}$. The convergence of $\{\hat{S}_{2,T}(\theta_2) - \bar{S}_{2,T}(\theta_2)\}_{T \in \mathrm{N}}$ in probability to zero uniformly in $\theta_2 \in \Theta_2$ (Sakata and White (2001), Lemma 5.1(f)) combined with this uniform identifiability yields the convergence of $\{\tilde{\theta}_{2,T}^2 - \theta_{2,T}^+\}_{T \in \mathrm{N}}$ and $\{\tilde{S}_{2,T}^2 - s_{2,T}^*\}_{T \in \mathrm{N}}$ in probability to zero uniformly over $\theta_2^1 \in \Theta_2^1$ (see also Andrews (1993)). Further, by hypothesis $\{\theta_{2,T}^+ - \theta_{2,T}^\oplus\}_{T \in \mathrm{N}}$ converges to zero uniformly in $\theta_2^1 \in \Theta_2$. It follows that $\{\tilde{\theta}_{2,T}^2 - \theta_{2,T}^\oplus\}_{T \in \mathrm{N}}$. converges to zero uniformly in $\theta_2^1 \in \Theta_2$, Given this fact, we take the first-order Taylor expansion of $S_{2,T}(\theta_2^1, \theta_2^2, \cdot)$ as a function of θ_2^2 about $\hat{\theta}_{2,T}^2$ with the Lagrange remainder, to obtain that

$$\text{(A.8)} \quad T(\hat{S}_{2,T}(\theta_2^1, \hat{\theta}_{2,T}^2) - S_{2,T}(\theta_2^1, \theta_{2,T}^\oplus)) = -\tfrac{1}{2} T(\hat{\theta}_{2,T}^2 - \theta_{2,T}^\oplus)' \nabla_{\theta_2^2 \theta_2^2} S_{2,T}(\theta_2^1, \bar{\theta}_{2,T})(\theta_{2,T}^2 - \theta_{2,T}^\oplus) + o_P(1),$$

where $\nabla_{\theta_2^2 \theta_2^2} S_{2,T}$ denotes the Hessian of $S_{2,T}$ with respect to the second parameters, $\bar{\theta}_{2,T}$ is $p_2 \times 1$ random vector on the chord between $\hat{\theta}_{2,T}^2$ and $\theta_{2,T}^\oplus$, and the remainder term reflects the fact that the gradient of $S_{2,T}$ at $\hat{\theta}_{2,T}^2$ may not vanish, when $\hat{\theta}_{2,T}^2$ is on the boundary of Θ_2^2. The remainder term converges to zero in probability uniformly in $\theta_2^1 \in \Theta_2^1$, because of the uniform convergence of $\{\tilde{\theta}_{2,T}^2 - \theta_{2,T}^+\}_{T \in \mathrm{N}}$ to zero discussed above. Because of this nature of the remainder term, we can now replace θ_2^1 with $\hat{\theta}_2^1$ to obtain that

$$\text{(A.9)} \quad T(\hat{S}_{2,T}(\hat{\theta}_2^1, \hat{\theta}_{2,T}^2) - S_{2,T}(\hat{\theta}_2^1, \theta_{2,T}^\oplus)) = -\tfrac{1}{2} T(\hat{\theta}_{2,T}^2 - \theta_{2,T}^\oplus)' \nabla_{\theta_2^2 \theta_2^2} S_{2,T}(\hat{\theta}_{2,T}^1, \bar{\theta}_{2,T})(\hat{\theta}_{2,T}^2 - \theta_{2,T}^\oplus) + o_P(1).$$

We below show that both $\{\nabla_{\theta_2^2 \theta_2^2} S_{2,T}(\hat{\theta}_{2,T}^1, \hat{\theta}_{2,T})\}_{T \in \mathrm{N}}$ and $\{(\hat{\theta}_{2,T}^2 - \theta_{2,T}^\oplus)\}_{T \in \mathrm{N}}$ are $O_P(1)$ to establish the desired result.

Since, $\nabla_{\theta_2^2\theta_2^2} S_{2,T}$ is a continuous function of $A_{2,T}, B_{2,T}, C_{2,T}$ and $D_{2,T}$ defined in appendix C and $\{A_{2,T} - \overline{A}_{2,T}\}_{T\in\mathrm{N}}, \{B_{2,T} - \overline{B}_{2,T}\}_{T\in\mathrm{N}}, \{C_{2,T} - \overline{C}_{2,T}\}_{T\in\mathrm{N}}$ and $\{D_{2,T} - \overline{D}\}_{T\in N}$ converges to zero in probability uniformly on Θ_2 under assumptions 1–4 and 5'(d, h), where $\overline{A}_{2,T}, \overline{B}_{2,T}, \overline{C}_{2,T}$ and $\overline{D}_{2,T}$ are defined in appendix C. It follows that $\{\nabla_{\theta_2^2\theta_2^2} S_{2,T} - \nabla_{\theta_2^2\theta_2^2} \overline{S}_{2,T}\}_{T\in\mathrm{N}}$ converges to zero in probability uniformly on Θ_2. Since, $\{\tilde{\theta}_{2,T}^2 - \theta_{2,T}^{\oplus}\}_{T\in\mathrm{N}}$ converges in probability to zero uniformly in $\theta_2^1 \in \Theta_2^1$, $\sup_{\theta_2^1\in\Theta_2^1} |\nabla_{\theta_2^2\theta_2^2} S_{2,T}(\cdot, \overline{\theta}_{2,T}) - \nabla_{\theta_2^2\theta_2^2} \overline{S}_{2,T}(\cdot, \theta_{2,T}^{\oplus})| = o_P(1)$. By assumptions 5'(d, i), we also have that $\sup_{\theta_2^1\in\Theta_2^1} |\nabla_{\theta_2^2\theta_2^2} \overline{S}_{2,T}(\theta_2^1, \theta_{2,T}^{+})| = O(1)$. Combining these facts proves that the middle factor on the RHS of (A.9) is $O_P(1)$.

We next show that $\{\sqrt{T}(\hat{\theta}_{2,T}^2 - \theta_{2,T}^{\oplus})\}_{T\in\mathrm{N}}$ is $O_P(1)$. Note that

$$\text{(A.10)} \quad \left|\sqrt{T}(\hat{\theta}_{2,T}^2 - \theta_{2,T}^{\oplus})\right| = \left|\sqrt{T}(\hat{\theta}_{2,T}^2 - \theta_{2,T}^{+}(\hat{\theta}_{2,T}^1))\right| + \left|\sqrt{T}(\theta_{2,T}^{+}(\hat{\theta}_{2,T}^1) - \theta_{2,T}^{\oplus})\right|$$

$$\leq \sup_{\theta_2^1\in\Theta_2^1}\left|\sqrt{T}(\tilde{\theta}_{2,T}^2 - \theta_{2,T}^{+})\right| + \sup_{\theta_2^1\in\Theta_2^1}\left|\sqrt{T}(\theta_{2,T}^{+}(\theta_2^1) - \theta_{2,T}^{\oplus})\right|.$$

Since the second term on the RHS is bounded by definition, it is sufficient to show that the first term is $O_P(1)$. In order to do so, we apply the standard linearization of the generalized score function, $\Psi_{2,T}^2$, to obtain the mean-value approximation to $(\tilde{\theta}_{2,T}^2, \tilde{S}_{2,T})$ for each $\theta_2^1 \in \Theta_2^1$. In the linear approximation, the consistency of $(\tilde{\theta}_{2,T}^2, \tilde{S}_{2,T})_{T\in\mathrm{N}}$ for $(\theta_{2,T}^{+}, s_{2,T}^{*})$ uniformly in $\theta_2^1 \in \Theta_2^1$ and assumptions 5'(d, e) imply that the difference between the component containing the inverse of the Jacobian of $\Psi_{2,T}^2$ and the inverse of $E[\nabla'\Psi_{2,T}^2(\cdot, \tilde{\theta}_{2,T}^2, s_{2,T}^{*}, \cdot)]$ converges to zero in probability uniformly in $\theta_2^1 \in \Theta_2^1$. Because $E[\nabla'\Psi_{2,T}^2(\cdot, \theta_{2,T}^{+}, s_{2,T}^{*}, \cdot)]$ is bounded and nonsingular uniformly in $\theta_2^1 \in \Theta_2^1$ and $T \in \mathrm{N}$, it follows that the operator norm of the inverse Jacobian part in the linear approximation is bounded. Therefore, the stochastic order of $\left\{\sup_{\theta_2^1\in\Theta_2^1}\left|\sqrt{T}(\tilde{\theta}_{2,T} - \theta_{2,T}^{+})\right|\right\}_{T\in N}$ is dominated by that of the supremum magnitude of the score, $\left\{\xi_T(\theta_2^1) \equiv \left|\Psi_{2,T}^2(\theta_2^1, \theta_{2,T}^{+}, s_{2,T}^{*}, \cdot)\right|\right\}_{T\in\mathrm{N}}$. Hence, the remaining task is to show that

$$\text{(A.11)} \qquad \sup_{\theta_2^1\in\Theta_2^1}\left|\sqrt{T}\xi_T(\theta_2^1)\right| = O_P(1).$$

Let ε be an arbitrary positive number. Then there exists a real number $\delta > 0$ such that

$$\text{(A.12)} \qquad \limsup_{T\to\infty} P\Big[\sup\Big\{\left|\sqrt{T}\xi_T(\theta_2^{1(1)}) - \sqrt{T}\xi_T(\theta_2^{1(2)})\right| : \left\|\theta_2^{1(1)} - \theta_2^{1(2)}\right\| < \delta,$$

$$, (\theta_2^1, \theta_2^1) \in \Theta_2^1 \times \Theta_2^1\Big\} > \varepsilon/2\Big] < \varepsilon/2,$$

since $\{\xi_T\}_{T\in\mathrm{N}}$ is stochastically equicontinuous by assumption 5'(g) . Since, Θ_2^1 is compact, we can cover it by a finite number, say m, of open balls with radius ε. Let $\theta_2^{1(i)}$ denote the center of the i th radius ($i = 1,2,\ldots,m$). Then $\{|\sqrt{T}\xi_T(\theta_2^{1(i)})|\}_{T\in\mathrm{N}}$ is $O_P(1)$ by assumptions 5'(e, f). Also, it follows from (A.12) that

$$\text{(A.13)} \qquad \sup_{\theta_2^1\in\Theta_2^1}\left|\sqrt{T}\xi_T(\theta_2^1)\right| \geq \max_{i\in\{1,2,\ldots,m\}}\left|\sqrt{T}\xi_T(\theta_2^{1(i)})\right| + (\varepsilon/2)$$

with a probability less than $\varepsilon/2$. Because the first term on the RHS of this inequality is an $O_P(1)$, we can choose a real number $\Delta > 0$, so that

$$\text{(A.14)} \qquad P\left[\max_{i\in\{1,2,\ldots,m\}}\left|\sqrt{T}\xi_T(\theta_2^{1(i)})\right| \geq \Delta\right] \leq (\varepsilon/2).$$

By the application of the implication rule, we obtain:

$$\text{(A.15)} \qquad P\left[\sup_{\theta_2^1\in\Theta_2^1}\left|\sqrt{T}\xi_T(\theta_2^1)\right| \geq \Delta^*\right] \leq \varepsilon,$$

where $\Delta^* = \Delta + (\varepsilon/2)$. Because $\varepsilon > 0$ was chosen arbitrarily, this proves (A.11). Hence, (A.9) is bounded in probability, and the desired result follows.

ii) If $T(\log\hat{S}_{2,T} - \log\hat{S}_{1,T}) = O_{a.s.}(\log\log T)$, than $PSC_{2,T} - PSC_{1,T}$ is positive a.s.-P given the assumption that $(\log\log T)^{-1}(c_{2,T} - c_{1,T}) \to_{a.s.} \infty$. Thus, it suffices to establish that $T(\log\hat{S}_{2,T} - \log\hat{S}_{1,T}) = O_{a.s.}(\log\log T)$. As in proposition 3(i), it sufficient to establish that the order of magnitude of $T(\hat{S}_{2,T} - S_{2,T}(\hat{\theta}_2^1, \theta_{2,T}^{\oplus}))$ is given by $O_{a.s.}(\log\log T)$.

The proof of proposition 3(ii) takes steps analogous to those taken in the proof of proposition 3(i). Replacing each convergence in probability with almost sure convergence and each stochastic equicontinuity with strong stochastic equicontinuity in the proof of proposition 3(i), we obtain that $|\nabla_{\theta_2^2\theta_2^2}\overline{S}_{2,T}(\theta_2^1, \theta_{2,T}^{+})| = O(1)$ and $|\nabla_{\theta_2^2\theta_2^2}S_{2,T}(\cdot, \overline{\theta}_{2,T}) - \nabla_{\theta_2^2\theta_2^2}\overline{S}_{2,T}(\cdot, \theta_{2,T}^{\oplus})| = o_{a.s.}(1)$ uniformly over Θ_2^1. In addition, we have that $\sup_{\theta_2^1\in\Theta_2^1}(\tilde{\theta}_{2,T} - \theta_{2,T}^{+}) = O_{a.s.}\left(\sup_{\theta_2^1\in\Theta_2^1}\left\|\xi_T(\theta_2^1)\right\|\right)$.

Now, given (A.9) with the $o_{a.s.}(1)$ remainder term, the desired result follows if $(\hat{\theta}_{2,T}^2 - \theta_{2,T}^{\oplus}) = O_{a.s.}(T^{-1/2}(\log\log T)^{1/2})$ uniformly over Θ_2^1. Let $a_T = T^{1/2}(\log\log T)^{-1/2}$ and note that

$$\text{(A.16)} \qquad \left|a_T(\hat{\theta}_{2,T}^2 - \theta_{2,T}^{\oplus})\right| = \left|a_T(\hat{\theta}_{2,T}^2 - \theta_{2,T}^{+}(\hat{\theta}_{2,T}^1))\right| + \left|a_T(\theta_{2,T}^{+}(\hat{\theta}_{2,T}^1) - \theta_{2,T}^{\oplus})\right|$$

$$\leq \sup_{\theta_2^1\in\Theta_2^1}\left|a_T(\tilde{\theta}_{2,T}^2 - \theta_{2,T}^{+})\right| + \sup_{\theta_2^1\in\Theta_2^1}\left|a_T(\theta_{2,T}^{+}(\theta_2^1) - \theta_{2,T}^{\oplus})\right|.$$

Since, the second term on the RHS is bounded by definition, it is sufficient to show that the first term is $O_{a.s.}(1)$. By assumption 6'(c), there exists a real number $\delta > 0$ such that

$$\text{(A.17)} \qquad \limsup_{T\to\infty} \sup_{\theta_2^1 \in \Theta_2^1} \sup_{\bar{\theta}_2^1 \in B(\theta_2^1,\delta)} \left\| \xi_T(\theta_2^1) - \xi_T(\bar{\theta}_2^1) \right\| \le 1,$$

where $B(\theta_2^1, \delta)$ denotes the closed sphere with radius δ centered at θ_2^1. Since Θ_2^1 is a compact set, it has a finite cover $\left\{ B(\theta_{2i}^1, \delta) : i = 1,2,\ldots,n \right\}$. It follows that

$$\text{(A.18)} \quad \limsup_{T\to\infty} \sup_{\theta_2^1 \in \Theta_2^1} \left\| \xi_T(\theta_2^1) \right\| \le \limsup_{T\to\infty} \max_{1\le i\le n} \left\| \xi_T(\theta_{2i}^1) \right\| + 1 \le \sum_{i=1}^{n} \limsup_{T\to\infty} \left\| \xi_T(\theta_{2i}^1) \right\| + 1,$$

Further, applying assumption 6'(b) to the first term on the RHS of this inequality, we obtain that $\sup_{\theta_2^1 \in \Theta_2^1} \left\| \xi_T(\theta_2^1) \right\| = O_{a.s.}(T^{-1/2}(\log\log(T))^{1/2})$ this complete the proof.

Proof of Proposition 5:

We begin by introducing some notations. Let Θ_h, $\bar{S}_{h,T}, \mu_{h,t}$ and $\psi_{h,t}$ be respectively the parameter space, the population S-scale measure, the regression function, and the summands in the generalized score function for the model with h- regimes. Denote by $\theta_{h,T}^*$ the parameter value that minimizes the population S-scale measure and $s_{h,T}^* \equiv \bar{S}_{h,T}(\theta_h^*)$. Also, for the partially unidentified model ($h = 3$) the parameter vector is partitioned as $\theta_3 = (\theta_3^1, \theta_3^2) \in \Theta_3^1 \times \Theta_3^2 \equiv \Theta_3$, where Θ_3^1, Θ_3^2, $\theta_{3,T}^+$, θ_3^1, and θ_3^2 are as in section 4 (for $k = 2$).

Conditions (A1)-(A4) clearly imply assumptions 1-2, 3(a,b) for all h; assumptions 5(a,b,c) for $h = 1,2$, and assumptions 5'(b,c',e) for $h = 3$ also follow from them. To establish the rest of the assumptions, we first show that the true STAR process ($h_0 = 2$) is a strictly stationary, ergodic and exponential β- mixing process.

In order to show the geometric ergodicity of $\{y_t\}$, we use results in the theory of continuous state Markov chains. We formulate the STAR process in its state space representation. Letting $x_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})'$ and $u_t = (\varepsilon_t, 0, \ldots, 0)'$, define matrix

$$\Phi = \begin{bmatrix} \beta_{11} + \beta_{12}G(\cdot) & \cdots & \beta_{p-1,1} + \beta_{p-1,2}G(\cdot) & \beta_{p-1,1} + \beta_{p,2}G(\cdot) \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix},$$

so that

(A.19) $$x_t = \Phi \cdot x_{t-1} + u_t .$$

The assumption on the density function of the error term implies that the STAR process is λ-irreducible (λ is the Lebesgue measure) and aperiodic (Chan, 1993). The boundedness of G implies that any λ non-null relatively compact sets, i.e., sets whose closure are compact, are small (Chan and Tong (1985), Chan (1993)). Given these results, 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

(A.20) $$E(V(x_{t+m}) \mid x_{t-1} = x) \le \eta \cdot V(x) + a \cdot 1_C(x) \text{ for all } x \in \Omega,$$

the process is geometrically ergodic and there exists an invariant measure for the process. If x_t is initiated at the invariant distribution or in the infinite past then the Markov chain is stationary, ergodic and β-mixing with exponential decay. Finally, $EV(x_t) < \infty$ (see Tjosthem (1990) and Meyen and Tweedie (1993), Ch.15-16). Let

$$\overline{\Phi} = \begin{bmatrix} \sup_z \mid \beta_{11} + \beta_{12} z \mid & \cdots & \sup_z \mid \beta_{p-1,1} + \beta_{p-1,2} z \mid & \sup_z \mid \beta_{p-1,1} + \beta_{p,2} z \mid \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix},$$

where $z \in [0,1]$. Note that $\overline{\Phi} \ge \Phi$, where the inequality means that it holds for each of the elements. Now, there exists a matrix Ξ such that $\overline{\Phi} = \Xi^{-1} \Lambda \Xi$ and Λ has its eigenvalues along its diagonal and arbitrarily small off-diagonal elements (Bellman (1960), pp.198-199). We consider the function $V(x) = 1 + \left\| \Xi \cdot x \right\|^k$ and a compact set $C = \left\{ x \in \Re^p : \left\| x \right\| \le c \right\}$ for some $c < \infty$, where $\left\| \cdot \right\|$ denotes the Euclidean norm for a vector or the spectral norm for a matrix. We have that

(A.21) $$V(x_t \mid x_{t-1} = x) \le 1 + \left\| \Xi \Phi x \right\|^k + \left\| \Xi u_t \right\|^k + O(x^{k-1}) \le 1 + \left\| \Xi \overline{\Phi} x \right\|^k + \left\| \Xi u_t \right\|^k + O(x^{k-1})$$

$$\le 1 + \left\| \Lambda \right\|^k \cdot \left\| \Xi x \right\|^k + O(x^{k-1}) .$$

The first inequality follows from the Minkowski inequality and the binomial formula, the second inequality is straightforward since $\Phi \le \overline{\Phi}$, and the last inequality results from condition (A2). By (A5), the largest eigenvalue of $\overline{\Phi}$ is less than one, and $\left\| diag(\Lambda) \right\|^k < 1$. Since the off diagonal elements of Λ can be made arbitrarily small, there exists $\eta \in (0,1)$ such that $\left\| \Lambda \right\|^k < \eta < 1$. Thus, we can choose $c < \infty$ such that

(A.22) $$E(V(x_t) \mid x_{t-1} = x) \le \eta \cdot V(x) + a \cdot 1_C(x) .$$

Hence $\{y_t\}$ is ergodic, asymptotically stationary and it satisfies that $E(y_t^{\kappa}) < \infty$ for $\kappa \geq 6$. Since the STAR process is started in the infinite past, it is strictly stationary, ergodic and β-mixing with an exponential decay.

These results, the compactness of the parameter set, and the boundedness of the function G imply the uniform integrability condition stated in assumption 3(c). Also, conditions (A2), (A3) and Theorem 3.35 of White (2001) imply that the sequences of random functions of (θ_h, s) defined in assumption 5(f) are stationary and ergodic (note that θ_k is replaced by θ_h). Furthermore, the existence of higher order moments, the compactness of the parameter set, the boundedness of ρ, ρ' and ρ'' and successive applications of the Holder inequality imply that these sequences are bounded uniformly by some integrable function. Hence, we can apply the USLLN for stationary ergodic processes (see, Ranga Rao (1962)) to show that these sequences obey the USLLN on $\Theta_h \times [s_1, s_2]$ and the family of functions defined in assumption 5(f) are continuous on $\Theta_h \times [s_1, s_2]$ where $s_1 \leq s_2$. Therefore, assumptions 3(c), 4(a), 5(e), 6(b,c) hold for each h and 5'(h), 6'(d,e) hold for $h = 3$. So, condition (A7) and proposition 2(i) imply that $\hat{h}_T \in \{2,3\}$ a.s. because these models are asymptotically better in comparison to the model with one regime.

Now, using conditions (A1)-(A3) and after some tedious calculations we have that the j-element of $\psi_{h,t}$ satisfies, $|\psi_{h,t}^{j}| \leq C_{h,j} | y_{t-i}^{\tau_1(j)} y_{t-1}^{\tau_2(j)} |$, where $\tau_1(j) = 0,1$, $\tau_2(j) = 0,1$ and $C_{h,j}$ are some positive constants; hence, Holder inequality and $E(y_t^{\kappa}) < \infty$ for $\kappa \geq 6$ imply that $E\left(|\psi_{h,t}^{j}|^2\right)$ is bounded uniformly over Θ_h for each h. Since the stationarity, ergodicity and the mixing properties of y_t are preserved by any continuous transformation, we can deduce that $\{\psi_{h,t}\}_{t \in \mathrm{N}}$ is strictly stationary, ergodic and β-mixing with an exponential decay process, which is bounded in the L_2-norm. So, the regular mixing zero-mean process $\{\psi_{h,t} - E(\psi_{h,t})\}_{t \in \mathrm{N}}$ is adapted L_2-mixingale of size –1/2 with respect to the subfields $\Im_t$ (see Davidson (1994), Theorem 14.2). Therefore, we can apply corollary AIII.3 of Sin and White (1992) to establish that this sequence satisfies the LIL for all $\theta_h \in \Theta_h$ for each h, so assumptions 6'(a, b) are satisfied.

Let $b_T = (T \log\log T)^{-1/2}$, we have to show that $b_T \sum_{t=1}^{T} \psi_{3,t}(\theta_3^1, \theta_{3,T}^{+}, s_{3,T}^{*}, \cdot)$ is strongly asymptotically equicontinuous on Θ_3^1 that is

$$\limsup_{T \to \infty} \sup_{\theta_3^1 \in \Theta_3^1} \sup_{\ddot{\theta}_3^1 \in B(\theta_3^1, \delta)} \left\| b_T \sum_{t=1}^{T} \psi_{h,t}(\theta_3^1, \theta_{3,T}^{+}, s_{3,T}^{*}, \cdot) - b_T \sum_{t=1}^{T} \psi_t^h(\ddot{\theta}_3^1, \theta_{3,T}^{+}, s_{3,T}^{*}, \cdot) \right\| \to 0$$

as $\delta \to 0$ a.s., where $B(\theta_3^1, \delta)$ denote the closed sphere with radius $\delta > 0$ centered at θ_3^1.
By using the smoothness and boundness of the first and second derivatives of ρ and G, the mean-value expansion around θ_3^1 and Holder inequality and noting that θ_3^{+} is independent of θ_3^1

and sample size, we can show that the difference between the j-elements of $b_T \sum_{t=1}^{T} \psi_{3,t}(\theta_3^1, \theta_{3,T}^+, s_{3,T}^*, \cdot)$ evaluated at θ_3^1 and $\ddot{\theta}_3^1$ satisfies that

$$\text{(A.23)} \quad b_T \left\| \sum_{t=1}^{T} \psi_{3,t}^{j}(\theta_3^1, \theta_{3,T}^+, s_{3,T}^*, \cdot) - \sum_{t=1}^{T} \psi_{3,t}^{j}(\ddot{\theta}_3^1, \theta_{3,T}^+, s_{3,T}^*, \cdot) \right\|$$

$$\leq \tilde{C}_j b_T \sum_{t=1}^{T} \left| y_{t-i}^{\tau_1(j)} y_{t-1}^{\tau_2(j)} \right| \cdot \left\| \theta_3^1 - \ddot{\theta}_3^1 \right\|$$

where $\tau_1(j) = 0,1$, $\tau_2(j) = 0,1,2$ and $\tilde{C}_j$ is some positive constant. Using similar arguments as mentioned above, we have that $\limsup_{T\to\infty} b_T \sum_{t=1}^{T} | y_{t-i}^{\tau_1(j)} y_{t-1}^{\tau_2(j)} | = O_{a.s.}(1)$. Therefore, the RHS of the inequality approaches zero a.s. as $\delta \to 0$ and assumption 6'(c) is verified. Since all assumptions imposed in 3(ii) are satisfied and the models are nested; condition (A7) on the penalty term implies that $\hat{h}_T \to 2$ a.s.

Appendix C (the derivatives of the scale function)

When $T \in N$ and $\omega \in \Omega$ satisfy that $0 < S_{k,T}(\theta_k, \omega) < \infty$ then for each $\theta_k \in \Theta_k$

$$\text{(C.1)} \quad h_{k,T}(\theta_k, S_{k,T}(\theta_k, \omega)) = M \quad \text{and} \quad \bar{h}_{k,T}(\theta_k, \bar{S}_{k,T}(\theta_k)) = M ,$$

as Lemma 3.1(c) in Sakata and White (2001) shows. By the implicit function theorem, we obtain that

$$\text{(C.2)} \quad \nabla_{\theta_k} S_{k,T} = -\frac{\nabla_{\theta_k} h_{k,T}}{\partial h_{k,T} / \partial s} = -A_{k,T} / B_{k,T} ,$$

where

$$A_{k,T}(\theta_k) = -\tfrac{1}{T} \sum_{t=1}^{T} \left[\rho' \left(\frac{Y_t - \mu_{k,t}(\theta_k)}{S_{k,T}(\theta_k)} \right) \nabla_{\theta_k} \mu_{k,t}(\theta_k) \right],$$

$$B_{k,T}(\theta_k) = -\tfrac{1}{T} \sum_{t=1}^{T} \left[\rho' \left(\frac{Y_t - \mu_{k,t}(\theta_k))}{S_{k,T}(\theta_k)} \right) \right] \cdot (Y_t - \mu_{k,t}(\theta_k)) S_{k,T}^{-2}(\theta_k) ,$$

and $B_T > 0$, since $S_{k,T}$ is a positive bounded number and ρ' is an odd function. Taking the further derivatives of $\nabla_\theta S_{k,T}$ delivers that

(C.3) $$\nabla_{\theta_k\theta_k} S_{k,T} = -\frac{[C_{k,T} \cdot B_{k,T} - A_{k,T} \cdot D_{k,T}']}{B_{k,T}^2},$$

where,

$$C_{k,T}(\theta_k) = \tfrac{1}{T}\sum_{t=1}^{T}\left[E_{k,t}(\theta_k) \cdot \nabla_{\theta_k} \mu_{k,t}(\theta_k)'\right] - \tfrac{1}{T}\sum_{t=1}^{T}\left[\rho'\left(\frac{Y_t - \mu_{k,t}(\theta_k)}{S_{k,T}(\theta_k)}\right)\nabla_{\theta_k\theta_k}\mu_{k,t}(\theta_k)\right],$$

$$D_{k,T}(\theta_k) = \tfrac{1}{T}\sum_{t=1}^{T}\left[E_{k,t}(\theta_k) \cdot S_{k,T}^{-2}(\theta_k) \cdot (Y_t - \mu_{k,t}(\theta_k)) + \rho'\left(\frac{Y_t - \mu_{k,t}(\theta_k)}{S_{k,T}(\theta_k)}\right) \cdot F_{k,t}(\theta_k)\right],$$

$$E_{k,t}(\theta_k) = \rho''\left(\frac{Y_t - \mu_{k,t}(\theta_k)}{S_{k,T}(\theta_k)}\right) \cdot \left(\frac{\nabla_{\theta_k}\mu_{k,t}(\theta_k)}{S_{k,T}(\theta_k)} - \frac{(Y_t - \mu_{k,t}(\theta_k)) \cdot A_{k,T}(\theta_k)}{S_{k,T}^2(\theta_k) \cdot B_{k,T}(\theta_k)}\right),$$

$$F_{k,t}(\theta_k) = S_{k,T}^{-2}(\theta_k) \cdot \left[\nabla_{\theta_k}\mu_{k,t}(\theta_k) - 2S_{k,T}^{-1}(\theta_k)(Y_t - \mu_{k,t}(\theta_k)) \cdot (A_{k,T}(\theta_k)/B_{k,T}(\theta_k))\right].$$

Let $\overline{A}_{k,T}, \overline{B}_{k,T}, \overline{C}_{k,T}$ and $\overline{D}_{k,T}$ be the population counterparts of $A_{k,T}, B_{k,T}, C_{k,T}$, and $D_{k,T}$, respectively (in which $S_{k,T}$ is replaced by $\overline{S}_{k,T}$). The first and second derivatives of $\overline{S}_{k,T}$, which can be calculated analogously, are:

(C.4) $$\nabla_{\theta_k}\overline{S}_{k,T} = -\frac{\nabla_{\theta_k}\overline{h}_{k,T}}{\partial\overline{h}_{k,T}/\partial s} = -\overline{A}_{k,T}/\overline{B}_{k,T}, \qquad \nabla_{\theta_k\theta_k}\overline{S}_{k,T} = -\frac{[\overline{C}_{k,T} \cdot \overline{B}_{k,T} - \overline{A}_{k,T} \cdot \overline{D}_{k,T}']}{\overline{B}_{k,T}^2}.$$

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. (1987). Consistency in nonlinear econometric models. A generic uniform law of large numbers. Econometrica, 55, 1465-1471.

Andrews D.W.K. (1992). Generic uniform convergence. 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, 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.

Balke, N.S. and Fomby. T.B. (1994). Large shocks, small shocks, and economic fluctuations: outliers in macroeconomic time series. Journal of Applied Econometrics, 9, 181-200.

Basci S. and Zaman A. (1998). Effects of skewness and kurtosis on model selection criteria. Economic Letters, 59, 17–22.

Beaton A.E. and Tukey J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics, 16, 147-192

Bellman R. (1960). Introduction to matrix analysis. New York, McGraw-Hill.

Bierens H. B. (1990). A conditional moment test of functional form. Econometrica, 58, 1443-1458.

Bierens H.J. and Ploberger W. (1997). Asymptotic theory of integrated conditional moments tests. Econometrica, 65, 1129-1152.

Chan K.S. and Tong H. (1985). On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equation. Advances in Applied Probability, 17, 666-678.

Chan K.S. (1993). A review of some limit theorems of Markov chains and their applications, in Dimension, Estimation and Models H. Tong editor, Singapore, New York, World Scientific Publishing.

Craven P. and Wahba G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation. Numerical Mathematics, 31, 377-403.

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.

Dijk V. D., Franses, P.H. and Lucas A. (1999). Testing for smooth transition nonlinearity in the presence of outliers. Journal of Business and Economic Statistics, 17, 217-235.

Donoho D.L. and Huber P.J. (1983). The notion of breakdown point. In A Festschrift for Erich L. Lehmann in Honor of his 65th Birthday, Bickel, P.J, Doksum K.A.. Hodges J.L. Jr (Eds), Belmont, CA, Wadsworth International Group, 157-184.

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. and Terasvirta T. (1993). Modelling nonlinear economic relationships. New York, Oxford University Press.

Hall P. and Heyde C.C. (1980). Martingale limit theory and its application. New York and London, Academic Press.

Hampel F.R. (1968). Contribution to the theory of robust estimation, Ph.D thesis, Berkeley, University of California.

Hampel F.R. (1971). A general qualitative definition of robustness, Annals of Mathematical Statistics, 42, 1887-1896.

Hampel F.R. (1975). Beyond location parameters: robust concepts and methods. In International Statistical Institute, Proceedings of the 40th Session, 46, 375-391.

Hannan E.J. and B.J. Quinn (1979). The determination of the order of autoregression. Journal of Royal Statistical Society Series B, 41, 190-195.

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

Hansen B.E. (1996b). Stochastic equicontinuity for unbounded dependent heterogeneous arrays. Econometric Theory 12, 347-359.

Jennrich R.I. (1969). Asymptotic properties of non-linear least squares estimators. Annals of Mathematical Statistics, 40, 633-643.

Kapetanios G. (2001). Model selection in threshold models. Journal of Time Series Analysis, 22, 733-754.

Koreisha S.G. and Pukkila T. (1993). Determining the order of a vector autoregression when the number of component series is large. Journal of Time Series Analysis, 14, 47–69.

Kuan C.M. and White H. (1994). Artificial neural networks: An econometric perspective. Econometric Reviews, 13, 1-91.

Le, N. D., Raftery A.E. and Martin R.D. (1996). Robust Bayesian model selection for autoregressive processes with additive outliers. Journal of the American Statistical Association, 91, 123–131.

Leroux B.G. (1992). Consistent estimation of a mixing distribution. The Annals of Statistics, 20, 1350-1360.

Lucas A. (1995). An outlier robust unit root test with an application to the extended Nelson-Plosser data, Journal of Econometrics, 66, 153-173.

Lutkepohl H. (1985). Comparison of criteria for estimating the order of a vector autoregressive process. Journal of Time Series Analysis, 6 , 35–52.

Luukkonen R., Saikkonen P. and Terasvirta T. (1988). Testing linearity against smooth transition autoregressive models. Biometrika, 75, 491–499.

Machado J.A.F. (1993). Robust model selection and M-estimation. Econometric Theory, 9, 478-493.

Martin R.D. (1980). Robust estimation of autoregressive models. In Direction in Time Series, D.R. Brillinger and G.C. Tiao (eds.), Hayward, CA: Institute of Mathematical Statistics, 228-262.

Nishii R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. Journal of Multivariate Analysis 27, 392-403.

Ozaki T. (1985). Non-linear time series models and dynamic systems. In Handbok of statistics, edited by E.J.Hannan, P.R. Krishnaiah and M.M. Rao. Vol 5. Amsterdam, New York, Elsevier Scientific Publishing.

Qian G. and Kunsch H.R. (1996). On model selection in robust linear regression. Research Report, 80, ETH Zurich.

Ranga Rao R. (1962). Relations between weak and uniform convergence of measures with applications. Annals of Mathematical Statistics 33, 659-680.

Redner R. (1981). Note on the consistency of the maximum likelihood estimate for non-identifiable distributions. Annals of Statistics 9, 225-239.

Rivers D. and Vuong Q. H. (2002). Model selection tests for nonlinear dynamic models. The Econometrics Journal, 5, 1-39

Ronchetti E. (1985). Robust model selection in regression, Statistics & Probability Letters, 3, 21-23.

Ronchetti E. and Staudte R.G. (1994). A Robust version of Mallow's C_p. Journal of the American Statistical Association, 89, 550-559.

Ronchetti E., Field C. and Blanchard W. (1997). Robust linear model selection by cross-validation. Journal of the American Statistical Association, 92, 1017-1023.

Rousseeuw P.J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79, 871-880.

Rousseeuw P.J. and Yohai V.J. (1984). Robust regression by means of S-estimators, in Robust and Nonlinear Time Series Analysis (Eds) W. H. Franke and R.D. Martin New York. Springer Verlag, 256-272.

Sakata, S. and White H. (1995). An alternative definition of finite sample breakdown point with applications to regression model estimation. Journal of the American Statistical Association, 90, 1099-1106.

Sakata, S. and White H. (1998). High-breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. Econometrica, 66, 529-567.

Sakata, S. and White H. (2001). S-estimation of nonlinear regression models with dependent and heterogeneous observations. Journal of Econometrics, 103, 5-72.

Schwarz G. (1978). Estimating the dimension of a model. Annals of Statistics 6, 461-464.

Sibbertsen P. (1999). S-estimation in the nonlinear regression model with long-memory error terms. Technical report 36/99, SFB 475, Dortmund, University of Dortmund.

Sibbertsen P. (2001). S-estimation in the linear regression model with long-memory error terms under trend. Journal of Time Series Analysis, 21, 353-363.

Sin C.Y. and White H. (1992). Information criteria for selecting possibly misspecified parametric models, Department of Economics Discussion Paper 92-47, San Diego, University of California.

Sin C.Y. and White H. (1996). Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71, 207-225.

Stinchcombe and White (1998). Consistent specification testing with nuisance parameters present only under the alternative. Econometric Theory, 14, 295-325.

Stromberg A.J. and Ruppert D. (1992). Breakdown in nonlinear regression. Journal of the American Statistical Association, 87, 991-997.

Tjostheim D. (1990). Non-linear time series and Markov chain. Advances in Applied Probability 22, 587-611

Terasvirta T. (1994). Specification, estimation and evaluation of smooth transition autoregressive models. Journal of the American Statistical Association, 89, 208-218.

Vuong H.Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 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. (1986). Asymptotic properties of econometric estimators. Ph.D. dissertation. San Diego, University of California.

Yohai V.J. and Zamer R.H. (1988). High breakdown point estimates of regressions by means of the minimizations of the efficient scale. Journal of the American Statistical Association, 83, 406-413.

Zhang M. Y., Russell J.R. and Tsay R.S. (2001). A nonlinear autoregressive conditional duration model with applications to financial transaction data. Journal of Econometrics, 104, 179-207.