J. Japan Statist. Soc.
Vol. 38 No. 1 2008 75–86

# ON THE SELECTION OF IRREGULAR, MISSPECIFIED REGRESSION MODELS: A COMMENT ON FOLKLORE

D. S. Poskitt*

In this paper we will investigate the consequences of applying model selection methods under regularity conditions that are sufficiently general to encompass (i) stochastic models involving non-stationary processes and (ii) situations where the true structure of the process falls outside the class of models under consideration. The properties of selection criteria that use very general measures of model complexity are considered and the results are used to draw attention to the fallacy of traditional beliefs concerning commonly employed model selection criteria.

*Key words and phrases*: Consistency, misspecified models, model selection, regression.

## 1. Introduction

Since the introduction of model selection criteria of the type first introduced by Akaike (Akaike (1974)) an extensive literature has been built up concerning the empirical and theoretical behaviour of such criteria, see, *inter alia*, McQuarrie and Tsai (1998). Much of this literature is concerned with the ability of the selection criteria to select the true model and in such discussions consistency is often deemed to be of fundamental importance. Here consistent means that the frequency of correct model selection converges to one as sample size $T$ increases. The importance of consistency arises because a set of candidate models $\mathfrak{R}$ is available for analysis, there is uncertainty about which model should be used, and a data-based choice is made from among the models in $\mathfrak{R}$. Subsequent inference is then conducted using the selected model. Employing a consistent procedure is therefore seen as desirable since, asymptotically at least, the true model will have been chosen and the inference will be valid.

Basing inference on the selected model as if it were the only one considered and ignoring model selection uncertainty is clearly a flawed process. Aspects of this problem are addressed in Potscher (1991) and Shen *et al.* (2004). Equally inappropriate is the notion that the chosen model represents the truth. Most theorists and practitioners would surely agree that models are only approximations to reality. ("All models are wrong, but some are useful" is a well known statement accredited to G. E. P. Box.) Classical derivations of consistency typically assume, nevertheless, that a minimal true model exists, that this true model is included in $\mathfrak{R}$, and when taking the limit in $T$ when deriving the asymptotic properties of a selection criterion the true model is held fixed. If we drop the

Accepted October 2, 2007.

*Department of Econometrics and Business Statistics, Monash University, Victoria 3800, Australia. Email: Don.Poskitt@Buseco.monash.edu.au

pretence that a true model exists then the property of consistency in the classical sense is less relevant, or at least needs reinterpretation. If all models are approximations, then an appropriate model for the analysis of a particular data set may depend on sample size—models employed with large data sets can be more highly structured and parameterized compared to those that are used when $T$ is small, the question of parsimony notwithstanding. Given this latter viewpoint, what is at issue is whether the data can be characterized by a model that is either (a) relatively simple, containing a few large, clearly discernable features, or (b) more complex, involving several smaller, more subtle interrelated effects. Thus we may want to consider increasing the number and complexity of the models in $\mathfrak{R}$ as $T$ increases and our analysis might be aimed at a moving target.

This paper examines model selection criteria in situations where the data generating processes under study and the models being used to describe them are not synonymous, allowing for the possibility that all models under consideration are false. Working in the context of fully specified, but possibly incorrect, probability models, Nishii (1988) and Vuong (1989) have shown that penalized log-likelihood and quasi-likelihood ratio test methods will, under appropriate conditions including i.i.d. assumptions, select models that minimize the Kullback-Liebler divergence, see also Sin and White (1996). The results presented here are similar to those obtained by these authors, but they are at the same time both (a) more specific, in that they focus on models that have been fitted using least squares, and (b) less restrictive, in that they impose regularity conditions on the data generating processes and models that are minimal and very weak.

To the current author's knowledge the majority of papers in this area employ assumptions that amount to assuming homoscedastic, stationary structures. Two notable exceptions are Paulsen and Tjøstheim (1985) and Potscher (1989). This paper expands on the ideas presented in these latter two articles and considers the properties of penalized least squares model selection procedures under conditions on the data generating process that are sufficiently weak to permit the analysis of heterogeneous data structures, as well as stationary and non-stationary processes. Thus we will provide an analytical background that is sufficiently flexible to allow for various different modeling scenarios involving irregular situations. We will also consider the behaviour of selection criteria constructed using penalty terms that are very general functions of the data. Unlike more conventional criteria such as $AIC$ and $BIC$ where the penalty only depends on the model dimension, such criteria attempt to measure other features of the model. One such criterion function, based on entropic complexity, is presented in Poskitt (1987).

## 2. Modeling assumptions

We will suppose that we are interested in modeling a real valued stochastic process $y_t$, $t \in \mathbb{N}$, defined on a probability space $(\Omega, \mathfrak{F}, P)$, using a linear regression model where the regressors are chosen from the collection $\mathcal{R} = \{z_{tn} : n = 1, \ldots, N\}$, $N \in \mathbb{N}$, of real-valued processes defined on the same probability space as $y_t$. Here $\mathbb{N}$ denotes the natural numbers (positive integers) and the variables in $\mathcal{R}$ are those deemed to be appropriate for the analysis at hand. In this case the set of relevant models $\mathfrak{R}$ contains the $2^N$ different models $\mathcal{M}_J$, $J = 1, \ldots, 2^N$, where under model $\mathcal{M}_J$ the subset of regressors $\{z_{tn_{J(k)}} : k = 1, \ldots, K_J\}$ enter the regression equation. Let $\mathfrak{F}_s$, $s \in \mathbb{N} \cup \{0\}$, denote a filtration of the $\sigma$-field $\mathfrak{F}$. Then we will suppose that $\{z_{tn} : n = 1, \ldots, N\}$ are $\mathfrak{F}_{t-1}$-measurable for all $t \in \mathbb{N}$. Note that we are not assuming that the models exhibit any particular structure. They are not nested, for example, although situations where the potential regressors have a natural ordering and all or subsets of the models are nested are clearly encompassed within the framework that we envisage.

Assumption 1. There exists an $\mathfrak{F}_{t-1}$-measurable function $m_t$ such that $y_t$ can be decomposed almost surely (a.s.) into $y_t = m_t + u_t$ such that (i) $u_t$ is $\mathfrak{F}_t$ measurable, and (ii) $E[u_t \mid \mathfrak{F}_{t-1}] = 0$ for all $t \in \mathbb{N}$.

The notation employed in Assumption 1 is motivated by the observation that the stated decomposition always exists given appropriate integrability, with $m_t$ equal to the conditional mean of $y_t$. Thus we may loosely think of $m_t$ as the mean value of $y_t$. The conditions in Assumption 1 ensure that $m_t$ and $u_t$ are uniquely defined up to sets of measure zero and that the innovation process $u_t$ is a martingale difference sequence. More importantly, the processes $m_t$ and $u_t$ are assumed to exist independently of any of the models in $\mathfrak{R}$.

Before establishing the link between the decomposition of Assumption 1 and $\mathfrak{R}$ let us define some additional notation. Let $\boldsymbol{y} = (y_1, \ldots, y_T)'$ denote the $T \times 1$ vector of observations on the dependent variable, or regressand, and similarly set $\boldsymbol{m} = (m_1, \ldots, m_T)'$ and $\boldsymbol{u} = (u_1, \ldots, u_T)'$, the corresponding vectors of unobservable mean values and stochastic disturbances. Define $\boldsymbol{X}_J$ to be the $T \times K_J$ observation matrix for the regressor set for model $\mathcal{M}_J$ with rows $\boldsymbol{x}'_{Jt} = (z_{tn_{J(1)}}, \ldots, z_{tn_{J(K_J)}}\}$ for $t = 1, \ldots, T$. For any $T \times q$ matrix $\boldsymbol{A}$ with column rank $\rho \leq q$ we will use $\boldsymbol{P}_A$ to denote the idempotent, symmetric matrix $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{\dagger}\boldsymbol{A}'$ where $(\boldsymbol{A}'\boldsymbol{A})^{\dagger}$ denotes the Moore-Penrose generalized inverse of $\boldsymbol{A}'\boldsymbol{A}$. The $T \times T$ matrix $\boldsymbol{P}_A$ is the (prediction) operator of rank $\rho$ that projects on to the space spanned by the columns of $\boldsymbol{A}$ and $\boldsymbol{R}_A = \boldsymbol{I}_T - \boldsymbol{P}_A$ is the associated (residual) operator of rank $T - \rho$ which projects on to the orthogonal complement of that space. The residual sum of squares obtained from fitting model $\mathcal{M}_J$ to the data is $\boldsymbol{y}'\boldsymbol{R}_{X_J}\boldsymbol{y} = \|\boldsymbol{R}_{X_J}\boldsymbol{y}\|^2$, where for $\boldsymbol{x} \in \mathbb{R}^T$ $\|\boldsymbol{x}\|^2 = \boldsymbol{x}'\boldsymbol{x}$.

Definition 2.1. *A model $\mathcal{M}_J$ will be said to be true if $\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|^2 = 0$ with probability one. If, on the other hand, there exists a $\delta > 0$ such that for all $T > T'$ $P[\omega : T^{-1}\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|^2 > \delta] = 1$, then $\mathcal{M}_J$ will be said to be a false model with proximity bounded by $\delta$, or more simply a false model. A false model $\mathcal{M}_J$ will be called a pseudo true model of propinquity $\delta$ if there exists a $T'$ such that $P[\omega : T^{-1}\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|^2 < \delta] = 1$ for all $T > T'$. If $\lim T^{-1}\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|^2 = \delta$ a.s. then $\mathcal{M}_J$ will be called a $\delta$-neighbourhood model.*

To illustrate Definition 2.1 suppose that $y_t$ is an autoregressive process of order $h$, so that $m_t = \alpha_1 y_{t-1} + \cdots + \alpha_h y_{t-h}$, and that $\mathcal{R} = \{y_{t-n+1} : n = 1, \ldots, N\}$, $N > h$. If the $J$'th model $\mathcal{M}_J$ corresponds to an autoregression with regressors $y_{t-1}, \ldots, y_{t-p}$, $\mathcal{M}_J = \mathcal{AR}_{\{p\}}$ say, then $\mathcal{M}_J = \mathcal{AR}_{\{p\}}$ will be be true if $p \geq h$, but false if $p < h$. If $y_t$ is a finite order moving-average process, however, all $\mathcal{AR}_{\{p\}}$ models will be false. The proximity of $\mathcal{M}_J = \mathcal{AR}_{\{p\}}$ will depend on how closely $y_t$ can be approximated in mean squared error by an autoregression of order $p$. For $p$ sufficiently large $\mathcal{M}_J = \mathcal{AR}_{\{p\}}$ will be a pseudo true model with propinquity that exceeds the almost sure limit of $\min_{\alpha_1,\ldots,\alpha_p} T^{-1} \sum_{t=1}^{T} [(m_t - \sum_{j=1}^{p} \alpha_j y_{t-j})^2]$.

Although Definition 2.1 allows for the possibility that there exists a member of $\mathfrak{R}$ that is congruent to the true data generating process, it does not of itself presuppose that the axiom of correct model specification holds. Assuming that $\mathfrak{R}$ contains a correct specification, i.e. a true model, will in general be implausible since the structure of most models is governed by questions other than realism, analytic tractability often being an important consideration. Indeed, the axiom of correct model specification can sometimes be precluded by virtue of the very nature of the models under investigation. We do not want to preclude the axiom of correct model specification in our analysis, but we do wish to allow for the rather more reasonable possibility that it is violated.

Assumption 2. The innovation process $u_t$ is such that (i) $\liminf_{t \geq 1} \sigma_t^2 > 0$ a.s., where $\sigma_t^2 = E[u_t^2 \mid \mathfrak{F}_{t-1}]$, and (ii) $\sup_{t \geq 1} E[u_t^{2+\gamma} \mid \mathfrak{F}_{t-1}] < \infty$ a.s. for some $\gamma > 0$.

The restrictions on $u_t$ in Assumption 2 mean, in essence, that the noise in the data generating mechanism does not die out and the observed process does not ultimately become deterministic. It is useful to note that under Assumption 2

$$\liminf_{T \to \infty} T^{-1} \|\boldsymbol{u}\| = \liminf_{t \geq 1} \sigma_t^2 > 0 \qquad \text{a.s.}$$

and that

$$T^{-1} \|\boldsymbol{u}\|^2 = T^{-1} \sum_{t=1}^{T} u_t^2 = \bar{\sigma}^2 + o(1) \qquad \text{a.s.}$$

where $\bar{\sigma}^2 = T^{-1} \sum_{t=1}^{T} \sigma_t^2$ (see Chow (1965)).

In Lai and Wei (1982) the eigenvalues of $\boldsymbol{X}_J' \boldsymbol{X}_J$ are used to characterize very weak regularity conditions for the strong consistency of least squares estimators in correctly specified stochastic regression models. If $\lambda_{\max}(\boldsymbol{X}_J' \boldsymbol{X}_J)$ and $\lambda_{\min}(\boldsymbol{X}_J' \boldsymbol{X}_J)$ are, respectively, the maximum and minimum of the non-zero eigenvalues of $\boldsymbol{X}_J' \boldsymbol{X}_J$, it is assumed that

$$\begin{aligned} &\lambda_{\min}(\boldsymbol{X}_J' \boldsymbol{X}_J) \to \infty \quad \text{a.s.} \qquad \text{and} \\ &\log \lambda_{\max}(\boldsymbol{X}_J' \boldsymbol{X}_J) = o(\lambda_{\min}(\boldsymbol{X}_J' \boldsymbol{X}_J)) \quad \text{a.s.} \end{aligned} \tag{2.1}$$

The restrictions on the regressors given in (2.1) are satisfied by several processes. If $T^{-1}(\boldsymbol{X}_J' \boldsymbol{X}_J)$ converges a.s., as would be the case if $\boldsymbol{x}_{Jt}$ were stationary and ergodic for example, then $\lambda_{\min}(\boldsymbol{X}_J'\boldsymbol{X}_J) = O(T)$ and $\log\lambda_{\max}(\boldsymbol{X}_J'\boldsymbol{X}_J) = O(\log T)$ a.s. A linear dynamic input-output system where the exogenous inputs $z_{tn} = O(t^\alpha)$, $\alpha > 0$, and the autoregressive operator has zeroes on the unit circle in the complex plane, but is otherwise stable, produces outputs $y_t = O(t^\beta)$, $\beta > 0$, (Lai and Wei (1982), Theorem 2) and $\log\lambda_{\max}(\boldsymbol{X}_J'\boldsymbol{X}_J) = O(\log T)$ a.s. for such a model. See Lai and Wei (1982) for further discussion. This motivates the following:

Assumption 3. The regressors in $\mathcal{R}$ are such that for all $\boldsymbol{X}_J$, $J = 1,\ldots, 2^N$, condition (2.1) holds and, furthermore, $\log\lambda_{\max}(\boldsymbol{X}_J'\boldsymbol{X}_J)/T \to 0$ a.s. as $T \to \infty$.

## 3. Model selection

Let $\widehat{\sigma}^2_{TJ} = T^{-1}\|\boldsymbol{R}_{X_J}\boldsymbol{y}\|^2$. We will be concerned here with the problem of selecting a model (or models) from within $\mathfrak{R}$ using a model selection criterion of the form

$$SC_T(\mathcal{M}_J) = \log(\widehat{\sigma}^2_{TJ}) + \frac{C(\mathcal{M}_J,T)}{T},$$

where $C(\mathcal{M}_J,T)$ is a nonnegative, real-valued variable that is chosen so as to measure the complexity of the model $\mathcal{M}_J$. The complexity measure may be a function of the data and the order of magnitude of $C(\mathcal{M}_J,T)$ is assumed to be the same for all candidate models, so that the ratio of any two such functions is $O(1)$. Common choices are $C(\mathcal{M}_J,T) = 2K_J$ and $C(\mathcal{M}_J,T) = K_J\log T$, which give rise to *AIC* and *BIC* respectively. The operational characteristics of $SC_T(\mathcal{M}_J)$ can be determined from the following lemma, wherein an event $E_T \subseteq \Omega$ is said to occur eventually if $P[\Omega\backslash E_T] = 0$ for $T$ sufficiently large.

Lemma 3.1. *Suppose that Assumptions* 1, 2 *and* 3 *hold.*

(i) *Without loss of generality, let $\mathcal{M}_1$ be a true model or a pseudo true model of propinquity $\delta_1$, and $\mathcal{M}_2$ a false model with proximity bounded by $\delta_2 \geq \delta_1 > 0$. Then the event*

$$\log\left(\frac{\widehat{\sigma}^2_{T2}}{\widehat{\sigma}^2_{T1}}\right) > \frac{C(\mathcal{M}_1,T) - C(\mathcal{M}_2,T)}{T}$$

*will occur eventually if*

$$\frac{C(\mathcal{M}_1,T) - C(\mathcal{M}_2,T)}{T} \to 0 \qquad a.s.$$

*as $T \to \infty$.*

(ii) *Now let $\mathcal{M}_1$ and $\mathcal{M}_2$ denote two models for which*

$$\|(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{m}\| = O(\|(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{u}\|) \tag{3.1}$$

*and*

$$|\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2| = O(|\|\boldsymbol{R}_{X_2}\boldsymbol{u}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{u}\|^2|), \tag{3.2}$$

*set* $\xi_T = \max\{\|\boldsymbol{P}_{X_1}\boldsymbol{u}\|^2, \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|^2\}$, *and suppose that* $|C(\mathcal{M}_1,T) - C(\mathcal{M}_2,T)| \to \infty$ *as* $T \to \infty$ *such that*

$$\liminf_{T\to\infty} \frac{|C(\mathcal{M}_1,T) - C(\mathcal{M}_2,T)|}{\xi_T} \geq c > 0 \qquad a.s. \tag{3.3}$$

*Then if* $c$ *is sufficiently large, there exists a random variable* $c'_T(\omega)$ *such that for almost all* $\omega$ *and all* $T$ *sufficiently large,* $c'_T(\omega) \geq \gamma > 0$ *and*

$$\frac{T(SC_T(\mathcal{M}_1) - SC_T(\mathcal{M}_2))}{\xi_T} \to \pm c'_T(\omega) \qquad a.s.$$

*according to whether* $|C(\mathcal{M}_1,T) - C(\mathcal{M}_2,T)| = \pm(C(\mathcal{M}_1,T) - C(\mathcal{M}_2,T))$.

Note that for two true models conditions (3.1) and (3.2) hold trivially since $\|(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{m}\|^2 = 0$ and $\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|^2 = 0$. If $\mathcal{M}_J$ is a false model, however, the approximation error $\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|^2$ grows at least linearly with $T$, and it is possible for $\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|^2$ to have an order of magnitude greater than $T$. Thus, if $\mathcal{M}_1$ is true and $\mathcal{M}_2$ is false then (3.1) and (3.2) must be violated; since $\|(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{m}\|^2 = \|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2 = \|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 = O(T^\alpha)$ where $\alpha \geq 1$ but the right hand sides of (3.1) and (3.2) are both of order $\xi_T$ at most, and by Lemma 1 (iii) of Lai and Wei (1982) $\|\boldsymbol{P}_{X_J}\boldsymbol{u}\|^2 = O(\log \lambda_{\max}(\boldsymbol{X}'_J\boldsymbol{X}_J))$ a.s. and $\log \lambda_{\max}(\boldsymbol{X}'_J\boldsymbol{X}_J)/T \to 0$ a.s. by Assumption 3. Similarly, if $\mathcal{M}_1$ is a pseudo true model then $\|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2 = O(T)$, and if $\mathcal{M}_2$ is false and $\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 = O(T^\alpha)$ where $\alpha > 1$, then $\|(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{m}\|^2$ and $|\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2|$ will both be $O(T^\alpha)$ and again (3.1) and (3.2) will not hold. For two pseudo true models $\mathcal{M}_1$ and $\mathcal{M}_2$ whose regressors are not collinear, but which nevertheless satisfy (3.1) and (3.2), $T^{-1}|\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2| \to 0$ a.s. as $T \to \infty$, implying that the two models will eventually have the same proximity and propinquity. Conditions (3.1) and (3.2) will hold in a trivial sense if the regressors of two false models span the same space, of course, because then $\boldsymbol{R}_{X_2}\boldsymbol{m} = \boldsymbol{R}_{X_1}\boldsymbol{m}$.

Now let us suppose that model selection is based upon the minimization of $SC_T(\mathcal{M}_J)$ over the $2^N$ different models in $\mathfrak{R}$. The chosen model is

$$\widehat{\mathcal{M}} = \arg \min_{\mathcal{M}_J \in \mathfrak{R}} SC_T(\mathcal{M}_J),$$

or an arbitrary model that minimizes the criterion function if the minimum is not unique. Let $\mathfrak{R}(\delta)$ denote the subset of pseudo true models in $\mathfrak{R}$ of propinquity $\delta$, $\mathfrak{N}(\delta)$ the subset of $\delta$-neighbourhood models, and let $\mathfrak{T}$ denote the subset of models in $\mathfrak{R}$ that are true. Set $\mathfrak{R}(0) = \emptyset$ and $\mathfrak{T}(\delta) = \mathfrak{T} \cup \mathfrak{R}(\delta)$. Now let $\Delta$ denote the finite set of (at most $2^N$) different values of $\delta$ such that $\delta = \inf \delta' \in (0,\infty) : \mathcal{M}_J \in \{\mathfrak{T}(\delta') \cup \mathfrak{N}(\delta')\}$ for $J = 1, \ldots, 2^N$, and let $\delta^* \in \Delta$ be such that $\delta \in \Delta$ implies $\delta \geq \delta^* \geq 0$. Then the following result characterizes the behaviour of $\widehat{\mathcal{M}}$.

THEOREM 3.2. *Suppose that Assumptions* 1, 2 *and* 3 *hold. Then*:

(i) *The event* $\widehat{\mathcal{M}} \in \{\mathfrak{T}(\delta^*) \cup \mathfrak{N}(\delta^*)\}$ *occurs eventually if for all* $\mathcal{M}_J$ *the complexity measures satisfy* $C(\mathcal{M}_J,T)/T \to 0$ *a.s. as* $T \to \infty$.

(ii) *If for any two models* $\mathcal{M}_1, \mathcal{M}_2 \in \{\mathfrak{T}(\delta^*) \cup \mathfrak{N}(\delta^*)\}$ *the complexity measures satisfy* $C(\mathcal{M}_J, T) \to \infty$ *a.s. as* $T \to \infty$ *for* $J = 1, 2,,$ *and conditions* (3.1), (3.2) *and* (3.3) *of Lemma* 3.1 *obtain, then eventually* $\widehat{\mathcal{M}}$ *will be either* (i) *a true model of minimal complexity, or, if* $\mathfrak{T} = \emptyset$, *a pseudo true* $\delta^*$*-neighbourhood model of minimal complexity, or* (ii) *a false model with proximity bounded by* $\delta^*$.

## 4. Implications

Combining parts (i) and (ii) of Theorem 3.2 indicates that if the complexity measure is structured so that $C(\mathcal{M}_J, T) \to \infty$ as $T \to \infty$ such that $\liminf C(\mathcal{M}_J, T)/\|\boldsymbol{P}_{X_J}\boldsymbol{u}\|^2 \geq c > 0$ and $C(\mathcal{M}_J, T)/T \to 0$ a.s., then if $c$ is sufficiently large $SC_T(\mathcal{M}_J)$ will be consistent for the true model of minimal complexity, if such exists, and if no true model exists then $SC_T(\mathcal{M}_J)$ will eventually select the closest approximating, pseudo true model with minimal complexity.

Consider now the behaviour of

$$AIC = \log(\widehat{\sigma}^2_{TJ}) + \frac{2K_J}{T} \qquad \text{and} \qquad BIC_\beta = \log(\widehat{\sigma}^2_{TJ}) + \beta \cdot \frac{K_J \log T}{T}$$

where $\beta > 0$. *AIC* (Akaike (1974)) and $BIC = BIC_1$ (Schwarz (1978)) are arguably the two most popular model selection criteria in current use. Obviously *AIC* satisfies the first part of Theorem 3.2 but not the second. $BIC_\beta$ also satisfies Theorem 3.2 (i) for any $\beta > 0$. Moreover, $\|\boldsymbol{P}_{X_J}\boldsymbol{u}\|^2 = O(\log \lambda_{\max}(\boldsymbol{X}_J'\boldsymbol{X}_J))$ a.s. (Lai and Wei (1982), Lemma 1 (iii)) and if $\log \lambda_{\max}(\boldsymbol{X}_J'\boldsymbol{X}_J) = o(\log T)$ a.s. and $\beta$ is sufficiently large, $BIC_\beta$ will also satisfy Theorem 3.2 (ii) and will be consistent. Thus, under the current scenario we reproduce the "well known result" that *AIC* is not consistent in the classical sense. Perhaps rather surprisingly, we also find that consistency of *BIC* is *not* guaranteed.

Knowledge that $BIC_\beta$ will be consistent if $\beta$ is chosen sufficiently large is not particularly useful. Even if it can be shown that $\log \lambda_{\max}(\boldsymbol{X}_J'\boldsymbol{X}_J) = o(\log T)$ a.s., the value of $\beta$ required to produce consistency depends on the models and the data generating process and will be unknown to the practitioner. An arbitrary choice of $\beta$ is unlikely to be innocuous, although there are situations where $BIC_\beta$ will be consistent for any choice of $\beta$. If $y_t$ is a stationary autoregressive process, for example, and $\boldsymbol{X}_J$ corresponds to an autoregressive model with regressors $y_{t-1}, \ldots, y_{t-J}$, then it can be shown that $\|\boldsymbol{P}_{X_J}\boldsymbol{u}\|^2 = O(\log \log T)$ and condition (3.3) can be replaced by $\liminf_{T\to\infty} |C(\mathcal{M}_1, T) - C(\mathcal{M}_2, T)|/\log \log T \geq c > 0$ (c.f. Hannan and Quinn (1979)). Thus $BIC_\beta$ will be consistent for any value of $\beta > 0$ in this case. Such examples lend support to the oft-stated comment that *BIC* is consistent, but examples of this type can also be used to show that, contrary to common belief, *AIC* can be consistent. Following Knight (1989), suppose that $y_t$ is a finite autoregression with innovations that constitute a simple random sample from a distribution that is in the domain of attraction of an $\alpha$-stable law where $\alpha \in (0, 2)$. Assumption 2 will now be violated. Nevertheless, it can be shown that $\|\boldsymbol{P}_{X_J}\boldsymbol{u}\|^2 = O(\log T/T)^{1/\alpha} = o(1)$ in probability (Theorem

2.1, Phillips (1990, p. 50)) and *AIC* will be (weakly) consistent (see Burridge and Hristova (2007) for further details). Indeed, from the inequality $\|\boldsymbol{P}_{X_J}\boldsymbol{u}\|^2 \leq \|\boldsymbol{X}_J\boldsymbol{u}\|^2/\lambda_{\min}(\boldsymbol{X}_J'\boldsymbol{X}_J)$ it follows from Assumption 3 that in situations where the assumptions on the innovation process $u_t$ are sufficient to ensure that $\|\boldsymbol{X}_J\boldsymbol{u}\|^2 = o(\lambda_{\min}(\boldsymbol{X}_J'\boldsymbol{X}_J))$, any criterion $SC_T(\mathcal{M}_J)$ where $C(\mathcal{M}_J, T)/T \to 0$ as $T \to \infty$, will be consistent.

If all the candidate models are false and $\delta^* > 0$ then $\widehat{\mathcal{M}}_{AIC}$ and $\widehat{\mathcal{M}}_{BIC_\beta}$, to use an obvious notation, will eventually enter $\mathfrak{N}(\delta^*)$ and we can think of *AIC* and $BIC_\beta$ as being *approximation*-consistent in that they will select the closest approximating model from within $\mathfrak{R}$. If $\mathfrak{N}(\delta^*)$ contains two or more models that are equally close to the data generating process then $T^{-1}\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 = T^{-1}\|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2 = \delta^*$. In this case *AIC* will not necessarily select the $\delta^*$-neighbourhood model of minimal dimension, nor need $BIC_\beta$. When $\mathfrak{T} = \emptyset$ the target model is, presumably, a model that minimizes the approximation error, but if more than one such model exists we may wish to distinguish between them on the basis of desiderata other than just parsimony. This suggests giving consideration to complexity measures $C(\mathcal{M}_J, T)$ that could be very different from the type of penalty terms that appear in *AIC* and *BIC*.

Finally, in discussions involving the most commonly used model selection devices it is often supposed that *AIC* is inconsistent and that *BIC* is consistent. From the previous analysis it is clear that a cavalier assumption that this folklore holds true is unlikely to be justified, even in the context of relatively simple models. Yang (2005) has shown that for the estimation of regression functions, consistency and minimax optimal convergence rates are incompatible. It seems therefore that we can conclude that there are cases where both *AIC* and *BIC* will be consistent but not optimal, and vice-versa! The ideas on model selection introduced by Akaike have spawned a vast literature, and Yang's seemingly paradoxical result suggests that they are likely to continue to do so.

## 5. Proofs

Proof of Lemma 3.1. Part (i). Consider first the difference

$$T(\widehat{\sigma}_{T2}^2 - \widehat{\sigma}_{T1}^2) = \|\boldsymbol{R}_{X_2}\boldsymbol{y}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{y}\|^2. \tag{5.1}$$

Substituting $\boldsymbol{y} = \boldsymbol{m} + \boldsymbol{u}$ and expanding the right hand side we obtain

$$\begin{aligned}\|\boldsymbol{R}_{X_2}\boldsymbol{y}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{y}\|^2 = {} & (\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2) \\ & + 2(\boldsymbol{m}'\boldsymbol{R}_{X_2}\boldsymbol{u} - \boldsymbol{m}'\boldsymbol{R}_{X_1}\boldsymbol{u}) \\ & + (\|\boldsymbol{P}_{X_1}\boldsymbol{u}\|^2 - \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|^2).\end{aligned} \tag{5.2}$$

By Lemma 1 (iii) of Lai and Wei (1982)

$$\|\boldsymbol{P}_{X_J}\boldsymbol{u}\|^2 = O(\log\lambda_{\max}(\boldsymbol{X}_J'\boldsymbol{X}_J)) \quad \text{a.s.}, \quad J = 1, 2, \tag{5.3}$$

and by Corollary 2 to the same lemma (Lai and Wei (1982, p. 159))

$$\boldsymbol{m}'\boldsymbol{R}_{X_J}\boldsymbol{u} = O(\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|[\log\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|]^{1/2}) \quad \text{a.s.}, \tag{5.4}$$

for $J = 1, 2$. By assumption, however, $T^{-1}\|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2 < \delta_1$ and $T^{-1}\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 \geq \delta_2 > \delta_1$ a.s. It follows that for all $T$ sufficiently large the right hand side of (5.2) can be bounded below by

$$\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 \left[\left(1 - \frac{\delta_1}{\delta_2}\right)(1 + o(1)) + \frac{o(1)}{\delta_2}\right] \qquad \text{a.s.}$$

Now, $\|\boldsymbol{R}_{X_1}\boldsymbol{y}\|^2 \leq \|\boldsymbol{y}\|^2$ and applying Minkowski's inequality to $\|\boldsymbol{y}\| = \|\boldsymbol{m} + \boldsymbol{u}\|$ we find that

$$T\widehat{\sigma}^2_{T1} = \|\boldsymbol{R}_{X_1}\boldsymbol{y}\|^2 \leq \|\boldsymbol{u}\|^2 \left(\frac{\|\boldsymbol{m}\|}{\|\boldsymbol{u}\|} + 1\right)^2.$$

Thus we can conclude that the event

$$\frac{\widehat{\sigma}^2_{T2} - \widehat{\sigma}^2_{T1}}{\widehat{\sigma}^2_{T1}} \geq \frac{\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2}{\|\boldsymbol{u}\|^2} \left(\frac{\|\boldsymbol{u}\|}{\|\boldsymbol{m}\| + \|\boldsymbol{u}\|}\right)^2 \left[\left(1 - \frac{\delta_1}{\delta_2}\right)(1 + o(1)) + \frac{o(1)}{\delta_2}\right]$$

will occur eventually.

It follows that

$$\begin{aligned}
&\log\left(\frac{\widehat{\sigma}^2_{T2}}{\widehat{\sigma}^2_{T1}}\right) \\
&\quad = \log\left(1 + \frac{\widehat{\sigma}^2_{T2} - \widehat{\sigma}^2_{T1}}{\widehat{\sigma}^2_{T1}}\right) \\
&\quad \geq \log\left(1 + \frac{1}{(\bar{\sigma}^2 + o(1))}\left(\frac{\|\boldsymbol{u}\|}{\|\boldsymbol{m}\| + \|\boldsymbol{u}\|}\right)^2 [(\delta_2 - \delta_1)(1 + o(1)) + o(1)]\right) \\
&\quad > 0 \qquad \text{a.s.}
\end{aligned}$$

and hence that the event

$$\log\left(\frac{\widehat{\sigma}^2_{T2}}{\widehat{\sigma}^2_{T1}}\right) > \frac{C(\mathcal{M}_1, T) - C(\mathcal{M}_2, T)}{T}$$

will occur eventually if $\{C(\mathcal{M}_1, T) - C(\mathcal{M}_2, T)\}/T \to 0$ a.s.

Part (ii). By definition

$$\left|\log\left(\frac{\widehat{\sigma}^2_{T2}}{\widehat{\sigma}^2_{T1}}\right)\right| = \log\left(1 + \frac{|\widehat{\sigma}^2_{T2} - \widehat{\sigma}^2_{T1}|}{\min\{\widehat{\sigma}^2_{T2}, \widehat{\sigma}^2_{T1}\}}\right).$$

Let

$$\zeta_T = \frac{|\widehat{\sigma}^2_{T2} - \widehat{\sigma}^2_{T1}|}{\min\{\widehat{\sigma}^2_{T2}, \widehat{\sigma}^2_{T1}\}}.$$

By the Mean Value Theorem of calculus (Apostol (1960, Theorem 5-10))

$$\left|\log\left(\frac{\widehat{\sigma}^2_{T2}}{\widehat{\sigma}^2_{T1}}\right)\right| = \frac{\zeta_T}{1 + \lambda\zeta_T} \tag{5.5}$$

where $0 \leq \lambda \leq 1$ for $\zeta_T$ in a neighbourhood of the origin. We will show below that $\zeta_T \to 0$ a.s. as $T \to \infty$ and thus it follows that (5.5) will hold with probability one as $T \to \infty$.

To verify that $\zeta_T \to 0$ a.s. as $T \to \infty$ note from (5.1) and (5.2) that

$$\begin{aligned} T|\widehat{\sigma}_{T2}^2 - \widehat{\sigma}_{T1}^2| \leq{}& |\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2| + 2|\boldsymbol{m}'(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{u}| \\ &+ |\|\boldsymbol{P}_{X_1}\boldsymbol{u}\|^2 - \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|^2|. \end{aligned} \tag{5.6}$$

Now, by Corollary 2 to Lemma 1 of Lai and Wei (1982)

$$|\boldsymbol{m}'(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{u}| = O(\|(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{m}\|[\log\|(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{m}\|]^{1/2}) \qquad \text{a.s.},$$

and for any two models that satisfy conditions (3.1) and (3.2)

$$\|(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{m}\| = O(\|\boldsymbol{P}_{X_1}\boldsymbol{u}\| + \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|)$$

and

$$|\|\boldsymbol{R}_{X_2}\boldsymbol{m}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{m}\|^2| = O(\|\boldsymbol{P}_{X_1}\boldsymbol{u}\|^2 + \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|^2),$$

since

$$\|(\boldsymbol{R}_{X_2} - \boldsymbol{R}_{X_1})\boldsymbol{u}\| = \|(\boldsymbol{P}_{X_2} - \boldsymbol{P}_{X_1})\boldsymbol{u}\| \leq \|\boldsymbol{P}_{X_1}\boldsymbol{u}\| + \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|$$

and

$$|\|\boldsymbol{R}_{X_2}\boldsymbol{u}\|^2 - \|\boldsymbol{R}_{X_1}\boldsymbol{u}\|^2| = |\|\boldsymbol{P}_{X_2}\boldsymbol{u}\|^2 - \|\boldsymbol{P}_{X_1}\boldsymbol{u}\|^2| \leq \|\boldsymbol{P}_{X_1}\boldsymbol{u}\|^2 + \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|^2.$$

We can therefore conclude from (5.6) that $T|\widehat{\sigma}_{T2}^2 - \widehat{\sigma}_{T1}^2|$ is majorized by

$$4\xi_T\left[O(1) + O\left(\frac{(\|\boldsymbol{P}_{X_1}\boldsymbol{u}\| + \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|)[\log(\|\boldsymbol{P}_{X_1}\boldsymbol{u}\| + \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|)]^{1/2}}{\|\boldsymbol{P}_{X_1}\boldsymbol{u}\|^2 + \|\boldsymbol{P}_{X_2}\boldsymbol{u}\|^2}\right)\right] \tag{5.7}$$

and hence, by virtue of (5.3) and Assumption 3, that $|\widehat{\sigma}_{T2}^2 - \widehat{\sigma}_{T1}^2| \to 0$ with probability one as $T \to \infty$.

Similarly,

$$\begin{aligned} &\min\{\widehat{\sigma}_{T2}^2, \widehat{\sigma}_{T1}^2\} \\ &\quad = T^{-1}\min_{J=1,2}\{\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|^2 + 2\boldsymbol{m}'\boldsymbol{R}_{X_J}\boldsymbol{u} + \|\boldsymbol{R}_{X_J}\boldsymbol{u}\|^2\} \\ &\quad = \min_{J=1,2}\left\{\frac{\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|^2}{T}\left[1 + 2O\left(\frac{[\log(\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|)]^{1/2}}{\|\boldsymbol{R}_{X_J}\boldsymbol{m}\|}\right)\right] + \frac{\|\boldsymbol{R}_{X_J}\boldsymbol{u}\|^2}{T}\right\} \\ &\quad \geq \frac{\{\|\boldsymbol{u}\|^2 - \max_{J=1,2}\|\boldsymbol{P}_{X_J}\boldsymbol{u}\|^2\}}{T} = \bar{\sigma}^2 + o(1) \qquad \text{a.s.} \end{aligned}$$

and it follows that $\zeta_T \to 0$ a.s. as $T \to \infty$, as required.

Thus, from (5.5) we can deduce that

$$\frac{T\{SC_T(\mathcal{M}_1) - SC_T(\mathcal{M}_2)\}}{\xi_T} = \zeta_T' \pm \frac{|C(\mathcal{M}_1, T) - C(\mathcal{M}_2, T)|}{\xi_T} \qquad \text{a.s.} \tag{5.8}$$

according to whether $C(\mathcal{M}_1,T)-C(\mathcal{M}_2,T)=\pm|C(\mathcal{M}_1,T)-C(\mathcal{M}_2,T)|$, where

$$\begin{aligned}|\zeta_T'| &= \frac{T\zeta_T}{\xi_T(1+\lambda\zeta_T)}\\ &\le \frac{T\zeta_T}{\xi_T}\end{aligned}$$

and $T(\bar{\sigma}^2+o(1))\zeta_T$ is bounded by (5.7). The desired conclusion now follows, with the right hand side of (5.8) equal to $\pm c_T'(\omega)$ as $|C(\mathcal{M}_1,T)-C(\mathcal{M}_2,T)| = \pm(C(\mathcal{M}_1,T)-C(\mathcal{M}_2,T))$, since $\log\lambda_{\max}(\boldsymbol{X}_J'\boldsymbol{X}_J)/\xi_T = O(1)$ a.s. This completes the proof. □

Proof of Theorem 3.2. To show Part (i) assume that $\widehat{\mathcal{M}}$ does not enter $\{\mathfrak{T}(\delta^*)\cup\mathfrak{N}(\delta^*)\}$. Then eventually $\widehat{\mathcal{M}}=\mathcal{M}_2$ where $\mathcal{M}_2$ is a false model with proximity bounded by $\delta^*$. By Lemma 3.1 (i) however, this implies that $SC_T(\mathcal{M}_2)>SC_T(\mathcal{M}_1)$ infinitely often for any $\mathcal{M}_1\in\{\mathfrak{T}(\delta^*)\cup\mathfrak{N}(\delta^*)\}$, contradicting the definition of $\widehat{\mathcal{M}}$. Hence $\widehat{\mathcal{M}}$ must enter $\{\mathfrak{T}(\delta^*)\cup\mathfrak{N}(\delta^*)\}$ eventually. Part (ii) follows similarly. Assume $\mathcal{M}_1,\mathcal{M}_2\in\{\mathfrak{T}(\delta^*)\cup\mathfrak{N}(\delta^*)\}$ and $C(\mathcal{M}_1,T)<C(\mathcal{M}_2,T)$ a.s. Then by Lemma 3.1 (ii)

$$\frac{T\{SC_T(\mathcal{M}_1)-SC_T(\mathcal{M}_2)\}}{\xi_T}\to -c_T'(\omega)\le -\gamma<0.$$

Thus the possibility that $\widehat{\mathcal{M}}$ will eventually equal $\mathcal{M}_2$ is excluded. Hence we can conclude that eventually $\widehat{\mathcal{M}}$ must either equal $\mathcal{M}_1$, or $\widehat{\mathcal{M}}$ does not enter $\{\mathfrak{T}(\delta^*)\cup\mathfrak{N}(\delta^*)\}$. If $\delta^*=0$ then $\mathcal{M}_1$ is a true model of minimal complexity, if $\delta^*>0$ then $\mathcal{M}_1$ is a pseudo true $\delta^*$-neighbourhood model of minimal complexity. □

## Acknowledgements

The author wishes to thank the referee for helpful comments. I am grateful to Les Godfrey and Peter Burridge for helpful discussions, they are absolved from any responsibility for the views expressed here, however, and I alone am responsible for any remaining errors.

## References

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

Apostol, T. M. (1960). *Mathematical Analysis*, Addison-Wesley, Reading.

Burridge, P. and Hristova, D. (2007). Consistent estimation and order selection for nonstationary autoregressive processes with stable innovations, Tech. rep., Department of Economics, City University, London.

Chow, Y. S. (1965). Local convergence of martingales and the law of large numbers, *Annals of Mathematical Statistics*, **36**, 552–558.

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

Knight, K. (1989). Consistency of Akaike's information criterion for infinite variance autoregressive processes, *The Annals of Statistics*, **17**, 824–840.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems, *The Annals of Statistics*, **10**, 154–166.

McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*, World Scientific Publishing Company, Singapore.

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

Paulsen, J. and Tjøstheim, D. (1985). Least squares estimates and order determination procedures for autoregressive processes with time dependent variance, *Journal of Time Series Analysis*, **6**, 117–133.

Phillips, P. (1990). Time series regression with a unit root and infinite-variance errors, *Econometric Theory*, **6**, 44–62.

Poskitt, D. S. (1987). Precision, complexity and Bayesian model determination, *Journal of the Royal Statistical Society*: *Series B*, **49**, 199–208.

Potscher, B. M. (1989). Model selection under nonstationarity: Autoregressive models and stochastic linear regression models, *Annals of Statistics*, **17**, 1257–1274.

Potscher, B. M. (1991). Effects of model selection on inference, *Econometric Theory*, **7**, 163–185.

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

Shen, X., Huang, H.-C. and Ye, J. (2004). Inference after model selection, *Journal of the American Statistical Association*, **99**, 751–762.

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

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, **57**, 307–333.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation, *Biometrika*, **92**, 937–950.