

SUBSAMPLING AND MODEL SELECTION IN TIME SERIES ANALYSIS REVISED

By Jun-ichiro Fukuchi

Faculty of Economics, Hiroshima University

Higashi-hiroshima, 739 Japan

Email:jfukuch@ipc.hiroshima-u.ac.jp

Summary

In this article, the subsampling method of Carlstein (1986) is used to estimate the risk of prediction for time series data. First, we extend Carlstein's result by proving strong consistency of the subsampling estimator. Second, we propose a procedure of selecting a time series model empirically from a set of possibly nonnested and misspecified models by using estimated risk of prediction as a selection criterion. Specifically, when this procedure is applied to the selection of the order of an autoregressive model, it is shown to be a consistent order selector if an appropriate subsample size is chosen. We propose a practical model selection procedure with a common subsample size chosen by Hall and Jing (1996)'s procedure.

Some key words: Autoregressive model; Mean squared error of prediction; Model selection; Subsampling; Threshold autoregressive model.

1. Introduction

The subsampling method is first proposed by Carlstein (1986) as a tool for estimating parameters of the sampling distribution of a statistic computed in a sample from a stationary process. Specifically, he used this method to estimate the variance of a statistic. Let $\{X_t\}_{t \geq 1}$ be a sequence of random variables, which is not necessarily stationary and $\mathbf{X}_n = (X_1, X_2, \dots, X_n)$ be a sample from it. Let \mathcal{F}_n^m denote the σ -algebra generated by $\{X_i, n \leq i \leq m\}$ and define the strong mixing coefficient of $\{X_t\}_{t \geq 1}$ by

$$\alpha(k) = \sup_j \sup_{A \in \mathcal{F}_1^j, B \in \mathcal{F}_{j+k}^\infty} |P(A \cap B) - P(A)P(B)|.$$

The sequence $\{X_t\}_{t \geq 1}$ is said to be strong mixing if $\alpha(k) \rightarrow 0$ as $k \rightarrow \infty$. The subsampling method is described as follows. Let $f_n : \mathbf{R}^n \rightarrow \mathbf{R}$ be a measurable function and suppose we would like to estimate $\Psi_n := E\{f_n(\mathbf{X}_n)\}$. Let $\mathbf{X}_k^i = (X_{i+1}, X_{i+2}, \dots, X_{i+k})$ be a subseries of $\{X_t\}_{t \geq 1}$ starting at X_{i+1} . Write $f_k^i = f_k(\mathbf{X}_k^i)$. Let $\{b_n\}$ be a sequence of positive integers such that $b_n \rightarrow \infty$ as $n \rightarrow \infty$. Then the nonoverlapping and overlapping versions of the subsampling estimator of Ψ_n with the subsample size b_n are defined to be

$$\hat{\Psi}_{n,b}^{(no)} = k_n^{-1} \sum_{i=0}^{k_n-1} f_{b_n}^i, \quad \hat{\Psi}_{n,b}^{(o)} = N_n^{-1} \sum_{i=0}^{N_n-1} f_{b_n}^i,$$

respectively, where $k_n = \lfloor n/b_n \rfloor$ and $N_n = n - b_n + 1$. The subscript n of b_n in the left hand sides of the above definitions is suppressed for notational simplicity. A characteristic of the subsampling method is that it does not require any knowledge of the specific structures of time series other than its asymptotic stationarity and strong mixing property. Carlstein (1986) proved L_2 -consistency of the nonoverlapping subsampling estimator when $\{X_t\}_{t \geq 1}$ is a stationary and strong mixing process. Consistency of the overlapping subsampling estimator will be presented in Section 1. In general, the overlapping version is more efficient than the nonoverlapping one; see Politis and Romano (1993). For this reason, the focus in this paper is centred on the overlapping version. In the following, $\hat{\Psi}_{n,b}$ denotes the overlapping subsampling estimator $\hat{\Psi}_{n,b}^{(o)}$.

For some standard examples, Ψ_n has an expansion

$$\Psi_n = \Psi + a_n^{-1} \eta + o(a_n^{-1}),$$

where $\{a_n\}$ is a sequence of positive integers such that $a_n \rightarrow \infty$. When the constant Ψ is known, use of the extrapolation

$$\tilde{\Psi}_{n,b} = \left(1 - \frac{a_b}{a_n}\right) \Psi + \frac{a_b}{a_n} \hat{\Psi}_{n,b}$$

is more sensible than that of $\hat{\Psi}_{n,b}$. This was proposed by Hall and Jing (1996). They call this method ‘sampling window’ method and showed its second order correctness for the case of estimating the sampling distribution of a function of the mean. The same type of

extrapolation was investigated fully in Bertail (1997). When Ψ is an unknown constant, an extrapolation based on two subsampling estimator $\hat{\Psi}_{n,n_1}$ and $\hat{\Psi}_{n,n_2}$ is possible. This kind of extrapolation was investigated by Bickel, Götze and van Zwet (1997) for the m out of n bootstrap.

The purpose of the present paper is two-fold. First, we establish consistency of the subsampling estimator in a general setup. In Section 2 we extend Carlstein's result on L_2 -consistency of the subsampling estimator to a class of nonstationary processes and prove its strong consistency. Second, we apply the subsampling method to the model selection in time series analysis. In Section 3 the subsampling method is used to estimate the risk of prediction in time series data. Section 4 is devoted to explaining a procedure of selecting a subsample size. In Section 5, we use the estimated risk of prediction as a model selection criterion. Specifically, we prove that this criterion is a consistent order selection criterion for an autoregressive process if an appropriate subsample size is chosen. In Section 6, a practical model selection procedure is defined by specifying how to choose a subsample size, and results of simulation studies on this procedure are presented.

A related method is the blockwise bootstrap of Künsch (1989) and Liu and Singh (1992). The use of the blockwise bootstrap for estimating the risk of prediction would be investigated elsewhere.

2. Consistency of subsampling estimator

In this section we establish consistency of the subsampling estimator. We make the following assumptions.

Assumption A1. There exists $\Psi \in \mathbf{R}$ such that

- (i) $\Psi_n \rightarrow \Psi$ as $n \rightarrow \infty$,
- (ii) $N_n^{-1} \sum_{i=0}^{N_n-1} E(f_{b_n}^i) \rightarrow \Psi$ as $n \rightarrow \infty$, for any $\{b_n\}$ with $b_n/n \rightarrow 0$.

Assumption A2. There exist constants $C_1 > 0$ and $\gamma > 3$ such that $E |f_k^i|^\gamma \leq C_1$, for every $i = 0, 1, \dots$ and $k = 1, 2, \dots$.

Assumption A1 resembles Assumption A in Politis, Romano and Wolf (1997). If

$\{X_t\}_{t \geq 1}$ is stationary or asymptotically stationary, Assumption A1(ii) trivially follows from A1(i). The next theorem is about L_2 - and strong consistency of $\hat{\Psi}_{n,b}$.

Theorem 1 *Let $\{X_t\}_{t \geq 1}$ be strong mixing.*

(a) *Suppose Assumption A1 holds and that $\{(f_n^i)^2, i = 0, 1, \dots, N_n - 1, n = 1, 2, \dots\}$ is uniformly integrable. If $b_n = o(n)$, then $\hat{\Psi}_{n,b} \rightarrow \Psi$ in L_2 , as $n \rightarrow \infty$.*

(b) *Suppose Assumption A1 and A2 hold and that $\alpha(k) \leq d^{-1} \exp(-dk)$, $k = 1, 2, \dots$, for some $d > 0$. If $b_n = o(n^\delta)$ for some $0 < \delta < (\gamma - 3)/(\gamma - 1)$, then $\hat{\Psi}_{n,b} \rightarrow \Psi$ almost surely, as $n \rightarrow \infty$.*

Theorem 1(a) is an extension of Theorem 2 in Carlstein (1986) for a stationary process to a nonstationary process.

3. Estimation of the risk of prediction

In this section we consider the estimation of the risk of prediction by subsampling. In the following, for the sake of simplicity, we assume that $\{X_t\}_{t \geq 1}$ is not only strong mixing but also stationary. But all the results continue to hold under some extra conditions without assuming stationarity.

Let $\hat{X}_{n+1} = g_n(\mathbf{X}_n)$ be a predictor of X_{n+1} based on \mathbf{X}_n , where $g_n : \mathbf{R}^n \rightarrow \mathbf{R}$ is measurable. Let $\text{PMSE}_n = E(\hat{X}_{n+1} - X_{n+1})^2$ be the mean squared error of predictor \hat{X}_{n+1} and let $\hat{X}_{i+b_n}^{(i)} = g_{b_n-1}(\mathbf{X}_{b_n-1}^i)$ be the predictor of X_{i+b_n} based on the subsample $(X_{i+1}, X_{i+2}, \dots, X_{i+b_n-1})$. The (overlapping version of) subsampling estimator of PMSE_n is given by

$$\widehat{\text{PMSE}}_{n,b_n}^{(o)} = N_n^{-1} \sum_{i=0}^{N_n-1} \left(\hat{X}_{i+b_n}^{(i)} - X_{i+b_n} \right)^2, \quad (1)$$

where $N_n = n - b_n + 1$. The next theorem is an immediate consequence of Theorem 1.

Theorem 2 *Assume that $\text{PMSE}_n \rightarrow \sigma^2$ for some $\sigma^2 > 0$, as $n \rightarrow \infty$.*

(a) *Assume that $\{(\hat{X}_{n+1} - X_{n+1})^4\}_{n \geq 1}$ is uniformly integrable. If $b_n = o(n)$, then*

$\widehat{\text{PMSE}}_{n,b_n}^{(o)} \rightarrow \sigma^2$ *in L_2 , as $n \rightarrow \infty$.*

(b) *Assume that $\alpha(k) \leq d^{-1} \exp(-dk)$, $k = 1, 2, \dots$, for some $d > 0$, and*

$E(|\hat{X}_{n+1} - X_{n+1}|^{2\gamma}) \leq C_1$ *for some $\gamma > 3$ and $C_1 > 0$. If $b_n = o(n^\delta)$, for some $0 < \delta < (\gamma - 3)/(\gamma - 1)$, then $\widehat{\text{PMSE}}_{n,b_n}^{(o)} \rightarrow \sigma^2$ almost surely, as $n \rightarrow \infty$.*

Remark 1. Apparently, Theorem 2 can be extended to a more general form of prediction risk. Such generalizations include the h -step ($h > 1$) ahead prediction and the loss function which is asymmetric in prediction error around zero.

From Theorem 2, advantages of this method are apparent. First, the method can apply to a general form of predictor $\hat{X}_{n+1} = g_n(\mathbf{X}_n)$. Second, for the subsampling estimator to be consistent, requirements on the data generating process (DGP) are only stationarity and a condition on the mixing coefficient of the process. Up to now, it has been shown that the class of stationary and strong mixing processes contains a wide variety of linear and nonlinear processes (cf. Doukhan, 1994, chapter 2). Thus this second characteristic of the method allows us to estimate the risk of predictor even when the model being used is misspecified.

Several methods of estimating the mean squared error of prediction are reviewed in Bhansali (1992). Some methods require that the DGP be an autoregressive or autoregressive moving average model of known order (Fuller and Hasza, 1981, Ansley and Newbold, 1981, and Stine, 1987). Some other methods require that the DGP be a linear process (Hannan and Nicholls, 1977, Shibata, 1980, and Bhansali, 1992). But, to the knowledge of the present author, there seems to be no literatures on the estimation of the risk of a predictor when the DGP belongs to a wider class than the class of linear processes.

A disadvantage of this method is that $\widehat{\text{PMSE}}_{n,b_n}$ has a bias as an estimator of PMSE_n ; see (2). A possible solution to it is a type of extrapolation which was investigated by Bickel, Götze and van Zwet (1997) for the case of independently and identically distributed random variables. Their method extrapolates an estimator from two ‘ m out of n bootstrap’ estimates obtained by putting $m = n_0$ and $m = n_1$, where $n_0, n_1 = o(n)$, and the same idea can be applied to subsampling. However, from our experience of a simulation study on our problem, this extrapolation sometimes produces erroneous estimates depending on values of n_0 and n_1 , so that we do not use it in this paper.

4. Choice of subsample size

An apparent obstacle in implementing the subsampling method is the choice of the subsample size. Let $\Psi_n = E\{f_n(\mathbf{X}_n)\}$ be the quantity of interest and $\hat{\Psi}_{n,b}$ be its subsam-

pling estimate using a subsample size b . When the optimal subsample size is $b \sim Cn^\delta$, where $\delta \in (0, 1)$ is a known real number, the following method is considered.

1. Fix $m < n$. Compute subsampling estimate $\hat{\Psi}_{n,m}$ from the entire data set (X_1, X_2, \dots, X_n) . 2. For each $b < m$, estimate the mean squared error of $\hat{\Psi}_{m,b}$ by

$$\widehat{\text{MSE}}(\hat{\Psi}_{m,b}) := (n - m + 1)^{-1} \sum_{i=1}^{n-m+1} (\hat{\Psi}_{m,b}^{(i)} - \hat{\Psi}_{n,m})^2,$$

where $\hat{\Psi}_{m,b}^{(i)}$ is the subsampling estimate of Ψ_m , computed from $(X_i, X_{i+1}, \dots, X_{i+m-1})$, using the subsample size b . 3. Select the value of b , say \hat{b}_m , which minimizes $\widehat{\text{MSE}}(\hat{\Psi}_{m,b})$. Take $\hat{b}_n = (n/m)^\delta \hat{b}_m$ as an estimate of the optimal subsample size.

This method of selecting a block length was proposed by Hall and Jing (1996) for the sampling window method, and a similar method was used in Hall, Horowitz and Jing (1995) for the blockwise bootstrap. Politis et al.(1997) suggested that this method can be used to select a subsample size in the subsampling. In the rest of the paper, we call the procedure defined by 1 to 3 above Hall and Jing's procedure.

In order to use this method, it is necessary to evaluate the mean squared error of the subsampling estimator. In the next theorem, we obtained the form of the bias and the variance of the nonoverlapping subsampling estimator of PMSE_n :

$$\widehat{\text{PMSE}}_{n,b_n}^{(no)} = k_n^{-1} \sum_{i=0}^{k_n-1} \left(\hat{X}_{(i+1)b_n}^{(ib_n)} - X_{(i+1)b_n} \right)^2.$$

Theorem 3 *Let $\{X_t\}_{t \geq 1}$ be a stationary strong mixing sequence. Assume*

- (i) $E \left(\left| \hat{X}_{n+1} - X_{n+1} \right|^\gamma \right) < c_1 < \infty, n = 1, 2, \dots$, for some $c_1 > 0$ and $\gamma > 4$,
- (ii) $\text{var} \left\{ (\hat{X}_{n+1} - X_{n+1})^2 \right\} \rightarrow c_2$, as $n \rightarrow \infty$, for some $c_2 > 0$,
- (iii) $\text{cov} \left\{ (\hat{X}_{n+1}^{(0)} - X_{n+1})^2, (\hat{X}_{2n+2}^{(n+1)} - X_{2n+2})^2 \right\} \rightarrow c_3$, as $n \rightarrow \infty$, for some $c_3 \in \mathbf{R}$,
- (iv) $\sum_{k=1}^{\infty} \alpha(k)^{(\gamma-4)/\gamma} < \infty$,
- (v) $\text{PMSE}_n = \sigma^2 + n^{-1}\eta + o(n^{-1})$ for some $\sigma^2 > 0$ and $\eta \in \mathbf{R}$,
- (vi) $b_n = o(n)$.

Then

$$\text{bias} \left(\widehat{\text{PMSE}}_{n,b_n}^{(no)} \right) := E \left(\widehat{\text{PMSE}}_{n,b}^{(no)} - \text{PMSE}_n \right)$$

$$= C_1 b_n^{-1} + o(b_n^{-1}), \quad (2)$$

$$\begin{aligned} \text{var} \left(\widehat{\text{PMSE}}_{n,b_n}^{(no)} \right) &:= E \left\{ \widehat{\text{PMSE}}_{n,b}^{(no)} - E(\widehat{\text{PMSE}}_{n,b}^{(no)}) \right\}^2 \\ &= C_2 b_n n^{-1} + o(b_n n^{-1}), \end{aligned} \quad (3)$$

where $C_1 \in \mathbf{R}$ is a constant and C_2 is a positive constant, which do not depend on n and b_n . Assume, in addition

$$(iv)' \quad \alpha(k) \leq C k^{-2\gamma/(\gamma-4)} \text{ for every } k = 1, 2, \dots,$$

$$(v)' \quad \text{PMSE}_n = \sigma^2 + n^{-1}\eta + O(n^{-3/2}),$$

$$(vi)' \quad b_n \gg n^{2/7},$$

where C is a positive constant. Then

$$\begin{aligned} \text{mse} \left(\widehat{\text{PMSE}}_{n,b_n}^{(no)} \right) &:= E \left(\widehat{\text{PMSE}}_{n,b}^{(no)} - \text{PMSE}_n \right)^2 \\ &= C_2 b_n n^{-1} + C_3 b_n^{-2} + o(b_n n^{-1} \vee b_n^{-2}). \end{aligned} \quad (4)$$

for some positive constant C_3 , where $a \vee b := \max(a, b)$.

In fact, the form of the bias (2) is obtained from condition (v) only, without assuming (i)-(iv). Condition (ii) states that the covariance of squared prediction errors computed on adjacent nonoverlapping subsamples converges to a constant. For the case of the prediction based on fitting an AR model of a finite order, conditions (i)-(v), (iv)' and (v)' are fulfilled; see Appendix 2.

It is readily shown that if (4) holds the optimal subsample size, in the sense of minimum mean squared error, for the nonoverlapping subsampling estimator is $b_n \sim C n^{1/3}$, where the positive constant C does not depend on n and b_n . Thus we can use Hall and Jing's procedure with $\delta = 1/3$.

For the overlapping subsampling estimator, it seems difficult to derive the form of asymptotically optimal subsample size. Let $\widehat{\text{PMSE}}_{n,b_n}^{(o)}$ be the overlapping subsampling estimator of PMSE_n defined by (1). It is easy to see the bias of $\widehat{\text{PMSE}}_{n,b_n}^{(o)}$ is the same as (2). If the data generating process (DGP) is an $\text{AR}(p_0)$ process and a predictor based on the $\text{AR}(p)$ model, where $p \geq p_0$, is used, then

$$\text{var} \left\{ \widehat{\text{PMSE}}_{n,b_n}^{(o)} \right\} = C_4 n^{-1} + o(n^{-1}), \quad (5)$$

where a positive constant C_4 depends on p and p_0 ; see Appendix 2. We are not able to obtain an expansion of the mean squared error of $\widehat{\text{PMSE}}_{n,b_n}^{(o)}$ because of the difficulty in deriving a further expansion of $\text{var} \left\{ \widehat{\text{PMSE}}_{n,b_n}^{(o)} \right\}$. However, it is apparent that the optimal subsample size for the overlapping subsampling estimator is different from that for the nonoverlapping subsampling estimator. This is a clear contrast to the result of Politis and Romano (1993). It can be readily seen from their results that for estimation of the variance of a ‘general linear statistic’, in their sense, the order of optimal subsample size for the overlapping and the nonoverlapping subsampling estimators are the same. The reason for the difference in these orders of optimal subsample size in our problem is that the squared error of prediction is asymptotically a function of only a fixed number of observations in the end of a sample when a Markov model such as an AR model is used.

5. Model selection

Theorem 2 leads to the idea of using the estimated risk of prediction by the subsampling as a model selection criterion. In this paper, a set of models is said to be ‘correctly specified’ if it contains the true model and ‘misspecified’ if it does not. Suppose we would like to select a model from a finite set of models $\mathcal{M} = \{M_1, M_2, \dots, M_K\}$ based on their predictive power. The candidate models M_1, M_2, \dots, M_K may be nested or nonnested, and correctly specified or misspecified. Let $\hat{X}_n(M)$ be a predictor constructed from the model $M \in \mathcal{M}$, $\text{PMSE}_n(M) := E\{\hat{X}_{n+1}(M) - X_{n+1}\}^2$, and $\widehat{\text{PMSE}}_{n,b_n}(M)$ be the overlapping subsampling estimator of $\text{PMSE}_n(M)$. Since the subsampling estimator $\widehat{\text{PMSE}}_{n,b_n}(M)$ consistently estimates $\text{PMSE}_n(M)$ for each model $M \in \mathcal{M}$ and no model assumptions are required for this result to be valid, it provides a natural means of comparing different models.

In the rest of this section, we provide a theoretical justification of this model selection procedure for a simple case that the DGP is an autoregressive process of a finite order and the candidate models are correctly specified. Suppose $\{X_t\}_{t=-\infty}^{\infty}$ is a doubly infinite sequence of random variables generated from AR(p_0) process

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_{p_0} X_{t-p_0} + \varepsilon_t, \quad -\infty < t < \infty. \quad (6)$$

We make the following assumptions:

Assumption B1. $1 - \sum_{i=1}^{p_0} \alpha_i z^i \neq 0$ for every complex number z with $|z| \leq 1$.

Assumption B2. ε_t is independently, identically, and symmetrically distributed around zero with finite variance.

Assumption B3(s). $E(|\varepsilon_t|^s) < \infty$ for some $s \geq 2$.

Assumption B4. The distribution of ε_t has an absolutely continuous component.

Assumption B5(p). $E\{\|\hat{\Gamma}_n^{-1}(p)\|^{2k}\}$ ($k = 1, 2, \dots, k_0$) is bounded for $n \geq n_0$ for some k_0 and n_0 , where $\hat{\Gamma}_n(p) = (n-p)^{-1} \sum_{t=p}^{n-1} X_{t,p} X'_{t,p}$ and $X_{t,p} = (X_t, X_{t-1}, \dots, X_{t-p+1})'$. Here the matrix norm for A is defined by $\|A\| = \sup(\beta' A \beta)$, where \sup is taken over all vectors satisfying $\|\beta\| \leq 1$.

Remark 2. Fuller and Hasza (1981) showed that Assumption B5 is satisfied if $\{X_t\}_{t \geq 1}$ is a Gaussian process.

Remark 3. If Assumptions B1, B2 and B4 are satisfied, the AR process (6) is geometrically absolutely regular (Doukhan, 1994, pp. 99). Hence it is strong mixing and its strong mixing coefficient satisfies $\alpha(k) \leq d^{-1} \exp(-dk)$, $k = 1, 2, \dots$, for some $d > 0$.

Consider fitting the AR(p) model to (X_1, X_2, \dots, X_n) and predict X_{n+1} by

$$\hat{X}_{n+1}(p) = \hat{\alpha}_{n,1}(p)X_n + \hat{\alpha}_{n,2}(p)X_{n-1} + \dots + \hat{\alpha}_{n,p}(p)X_{n-p+1}, \quad (7)$$

where $\hat{\alpha}_n(p) := \{\hat{\alpha}_{n,1}(p), \hat{\alpha}_{n,2}(p), \dots, \hat{\alpha}_{n,p}(p)\}'$ is given by

$$\hat{\alpha}_n(p) = \left(\sum_{t=p}^{n-1} X_{t,p} X'_{t,p} \right)^{-1} \sum_{t=p}^{n-1} X_{t,p} X_{t+1}.$$

Let $\text{PMSE}_n(p) = E\{\hat{X}_{n+1}(p) - X_{n+1}\}^2$ be the mean squared error of the predictor $\hat{X}_{n+1}(p)$. It can be deduced from Theorem 3 and Corollary 6 of Kunitomo and Yamamoto (1986) (hereafter referred as KY) that if Assumptions B1, B2, B3(s) and B5(p) are satisfied with $s = 32$, then

$$\text{PMSE}_n(p) = \sigma^2(p, p_0) + n^{-1} \eta(p, p_0) + O(n^{-\frac{3}{2}}), \quad (8)$$

as $n \rightarrow \infty$, for some $\sigma(p, p_0) > 0$ and $\eta(p, p_0) \in \mathbf{R}$.

Let $\hat{X}_{i+b_n}^{(i)}(p)$ be the predictor of X_{i+b_n} , obtained by fitting the AR(p) model to the subsample $(X_{i+1}, X_{i+2}, \dots, X_{i+b_n-1})$. The subsampling estimator of $\text{PMSE}_n(p)$ is given by

$$\widehat{\text{PMSE}}_{n,b_n}(p) = N_n^{-1} \sum_{i=0}^{N_n-1} \left\{ \hat{X}_{i+b_n}^{(i)}(p) - X_{i+b_n} \right\}^2, \quad (9)$$

where $N_n = n - b_n + 1$. Consistency of this estimator is given by the following.

Theorem 4 (a) *Suppose Assumptions B1, B2, B3(s), B4 and B5(p) hold with $s = 20$.*

If $b_n = o(n)$, then $\widehat{\text{PMSE}}_{n,b_n}(p) \rightarrow \sigma^2(p, p_0)$ in L_2 , as $n \rightarrow \infty$.

(b) *Suppose Assumptions B1, B2, B3(s), B4 and B5(p) hold with $s = 12\gamma$ for some integer*

$\gamma \geq 4$. If $b_n = o(n^\delta)$ where $0 < \delta < (\gamma - 3)/(\gamma - 1)$, then $\widehat{\text{PMSE}}_{n,b_n}(p) \rightarrow \sigma^2(p, p_0)$ almost surely, as $n \rightarrow \infty$.

Now we define an order selector for an AR model. Assume that the true order p_0 is unknown but it is less than or equal to K . Theorem 4 suggests that $\widehat{\text{PMSE}}_{n,b_n}(p)$ can be used as an order selection criterion, i.e. select \hat{p}_{n,b_n} , where

$$\hat{p}_{n,b_n} = \underset{1 \leq p \leq K}{\operatorname{argmin}} \widehat{\text{PMSE}}_{n,b_n}(p).$$

In a finite sample, the $\text{AR}(p_0)$ model may not be optimal, in terms of the mean squared prediction error, among AR models of finite order; see KY. However, it is asymptotically optimal, namely there exists N such that $\text{PMSE}_n(p_0) < \text{PMSE}_n(p)$ for any $n \geq N$ and $p \neq p_0$. Thus it is desirable for an order selector to converge to the true order p_0 as $n \rightarrow \infty$. The next theorem is about the weak and strong consistency of \hat{p}_{n,b_n} .

Theorem 5 *Suppose Assumptions B1, B2, B3(s), B4 and B5(p) hold for some $s \geq 2$ and for every $1 \leq p \leq K$.*

(a) *If $s = 32$ and $b_n = o(n^{1/3})$, then $\hat{p}_{n,b_n} \rightarrow p_0$ in probability as $n \rightarrow \infty$.*

(b) *If $s = 12\gamma$ for some integer $\gamma \geq 4$ and $b_n = o(n^\delta)$ where $0 < \delta < (\gamma - 3)/(3\gamma - 1)$, then $\hat{p}_{n,b_n} \rightarrow p_0$ almost surely as $n \rightarrow \infty$.*

Theorem 5 states that \hat{p}_{n,b_n} converges to the true order of the process if b_n satisfies a certain condition. However, the conditions on b_n required for the consistency of \hat{p}_{n,b_n} do not uniquely determine the value of b_n for given n .

Note that the same b_n is used in estimating every $\text{PMSE}_n(p)$, $p = 1, 2, \dots, K$. Our experience in a simulation study showed that the distribution of \hat{p}_{n,b_n} is much affected by the value of b_n and the characteristics of the DGP. In the next section, we propose a method of choosing a subsample size for the purpose of model selection.

6. Simulation study

A simulation study was conducted to assess small sample properties of the proposed model selection procedure. In the first simulation, samples are generated from an AR(2) process:

$$\text{DGP 1: } X_t = 1.4X_{t-1} - 0.8X_{t-2} + \varepsilon_t.$$

This model is from Bhansali (1981) and KY. Error terms $\{\varepsilon_t\}$ are independently and identically distributed standard normal random variables all through this section. The set of candidate models consists of autoregressive models of order 1 to 5 and thus it is correctly specified. The mean squared prediction error (PMSE) of the predictors were computed by a Monte Carlo simulation. Those based on AR models of order 1 to 5 for sample size 50 are 2.79, 1.04, 1.07, 1.09 and 1.12 respectively. Thus the order 2 is optimal.

———— Table 1 and 2 should be inserted around here ————

Table 1 presents the distribution of the order selector $\hat{p}_{n,b}$ with various values of subsample size b . When b is between 9 and 15, the optimal order 2 is selected in more than 90% of samples generated. The frequency of underestimating the optimal order increases as b gets smaller and that of overestimating increases as b gets larger. Underestimation should be avoided since it results in a severe increase in the PMSE. Another important point that is seen from Table 1 is that if an appropriate subsample size ($9 \leq b \leq 15$, in this example) is used to estimate each PMSE of different autoregressive order, the resulting order selector performs very well. This observation leads to the following method of order selection.

Step 1. Choose $p_1 \in \{1, 2, \dots, K\}$ and $m \in \{1, 2, \dots, n\}$.

Step 2. Estimate the optimal subsample size for $\widehat{\text{PMSE}}_{n,b_n}(p_1)$ by Hall and Jing's procedure with m . The estimated subsample size is denoted by $\hat{b}_n(p_1)$.

Step 3. Compute $\widehat{\text{PMSE}}_{n,\hat{b}_n(p_1)}(p)$, $p = 1, 2, \dots, K$.

Step 4. Select p which minimizes $\widehat{\text{PMSE}}_{n,\hat{b}_n(p_1)}(p)$, $p = 1, 2, \dots, K$. The selected p is denoted by \hat{p}_n .

This procedure is generalized to the model selection from a general set of candidate models $\mathcal{M} = \{M_1, M_2, \dots, M_K\}$ as follows.

Step 1'. Choose $M \in \{M_1, M_2, \dots, M_K\}$ and $m \in \{1, 2, \dots, n\}$.

Step 2'. Estimate the optimal subsample size for $\widehat{\text{PMSE}}_{n,b_n}(M)$ by Hall and Jing's procedure with m , where $\widehat{\text{PMSE}}_{n,b_n}(M)$ is the subsampling estimate of the PMSE of the predictor based on the model M . The estimated subsample size is denoted by $\hat{b}_n(M)$.

Step 3'. Compute $\widehat{\text{PMSE}}_{n,\hat{b}_n(M)}(M_i)$, $i = 1, 2, \dots, K$.

Step 4'. Select the model M_i which minimizes $\widehat{\text{PMSE}}_{n,\hat{b}_n(M)}(M_i)$, $i = 1, 2, \dots, K$. The selected model is denoted by \hat{M}_n .

As was stated in the end of Section 4, it is difficult to choose an appropriate value for δ in Hall and Jing's procedure. Several values of δ between $1/3$ and $2/3$ were used in preliminary simulations and we found that the value $\delta = 0.40$ works relatively well for both the first simulation and the second simulation which is described below. Results with $\delta = 0.40$ are presented in the following.

It should be noted that in the above procedure only one value is chosen as a common subsample size b and it is used to compute $\widehat{\text{PMSE}}_{n,b}$ for every candidate model. This means that candidate models are compared in each subsample in terms of the squared prediction error, and the model with the minimum squared prediction error averaged over all subsamples will be selected. This interpretation makes the proposed model selection procedure intuitively appealing.

Table 2 presents the distribution of \hat{p}_n with various values of (p_1, m) . The proposed method avoids severe underestimation and its overall performance seems to be satisfactory.

In the second simulation, samples are generated from a threshold AR process:

$$\text{DGP 2: } X_t = \begin{cases} -0.14 + 0.10X_{t-1} + \varepsilon_t & \text{if } X_{t-1} < -0.2, \\ 0.80X_{t-1} + \varepsilon_t & \text{if } X_{t-1} \geq -0.2. \end{cases}$$

Candidate models are autoregressive models of order 1 to 3 and threshold autoregressive models of the form:

$$X_t = \begin{cases} \alpha_1^{(1)} X_{t-1} + \alpha_2^{(1)} X_{t-2} \cdots + \alpha_p^{(1)} X_{t-p} + \varepsilon_t & \text{if } X_{t-1} < 0.0, \\ \alpha_1^{(2)} X_{t-1} + \alpha_2^{(2)} X_{t-2} \cdots + \alpha_p^{(2)} X_{t-p} + \varepsilon_t & \text{if } X_{t-1} \geq 0.0, \end{cases}$$

where $p = 1, 2, 3$. These threshold autoregressive models are denoted by $\text{TAR}(p)$, $p = 1, 2, 3$ in the following. Since the threshold value of candidate threshold AR models is different from that of the DGP, the set of candidate models does not contain the DGP,

i.e. it is misspecified and nonnested. Parameters of a TAR(p) model are estimated by the method of conditional least squares. A general account of threshold AR models is given in Tong (1990). The PMSE of predictors based on AR models of order 1 to 3 for sample size 100 are 1.09, 1.10 and 1.11 respectively, and those based on TAR models of order 1 to 3 are 1.01, 1.04 and 1.06 respectively. The TAR(1) model is optimal. For the sample size 200, those PMSEs are 1.10, 1.10, 1.10, 1.01, 1.02 and 1.04 in the respective order and the TAR(1) model is optimal.

The model selection procedure defined by Step 1' to Step 4' is applied. In Step 1' and Step 2' of the procedure, the TAR(p_1) model is used to choose a subsample size. Table 3 and 4 present the distribution of model selectors \hat{M}_n respectively for the sample size $n = 100$ and 200 respectively.

— Table 3 to 4 should be inserted here —

Small sample properties of the order selectors $\hat{p}_{n,b}$, \hat{p}_n and the model selector \hat{M}_n , found in the above simulations are summarized as follows.

(1) The distribution of $\hat{p}_{n,b}$ depends on the value of b . The use of too small b results in underfitting and too large b results in overfitting. The variance of $\hat{p}_{n,b}$ increases as b gets larger. (2) By using Hall and Jing's procedure of choosing the value of b , \hat{p}_n and \hat{M}_n avoid severe underfitting and overfitting. (3) Distributions of \hat{p}_n and \hat{M}_n depend on the DGP and the sample size. The more complex the DGP is, the larger sample size is needed to obtain a satisfactory large frequency of selecting the optimal model.

7. Conclusion

Most of known model selection criteria such as Akaike information criterion and Schwarz information criterion are devised to select a model from a set of nested candidate models. In practice, however, one often needs to select a model from a set of nonnested and possibly misspecified candidate model. Recently some model selection criteria are developed for the situation that candidate models are nonnested and possibly misspecified; see Konishi and Kitagawa (1996) for independently and identically distributed observations and Sin and White (1996) for dependent observations. Sin and White showed that

the penalized likelihood criterion is a consistent model selection criterion if the penalty term in the criterion satisfies certain conditions on the rate of growth to infinity. However, how to select an exact value of the penalty term in practice is still an open problem. Also, their results are restricted to the case where parameters of a candidate model are estimated by the quasi-maximum-likelihood method. Yao and Tong (1994) proposed a method of subset selection of stochastic regressors based on a cross-validation method. Their approach seems to be extendable to the model selection problem considered in this paper and it will be investigated elsewhere. In Section 6 of this paper, we proposed a model selection procedure based on estimated prediction risk, estimated by subsampling, of a candidate model. Although we did not provide theoretical results except the case that candidate models are AR models of finite order and the DGP is an AR model, the simulation results showed in a limited way usefulness of the proposed procedure for the model selection in nonnested and possibly misspecified candidate models.

Acknowledgements

I wish to thank two referees and the editor Professor D. M. Titterton for their suggestions that led to an improved version of the paper. Thanks go to Kaoru Futagami for her suggestions in FORTRAN programming in Section 6 and to Michio Hatanaka and Nouredine Rhomari for constructive comments. I also thank Patrice Bertail for the provision of his technical paper, from which I learned the Rhomari's result. Also, thanks go to Peter Bickel, Peter Hall, Dimitris Politis, and Taku Yamamoto for the provision of their technical papers. A revision of this work was done while I was visiting Centre for Mathematics and its Applications, Australian National University. I would like to thank Peter Hall for his kind hospitality.

Appendix 1

Proofs

Proofs of theorems are sketched below. The detailed proofs are available from the author on request.

Proof of Theorem 1(a). We have

$$\begin{aligned}
N_n^2 \text{var} \left(\hat{\Psi}_{n,b} \right) &\leq \sum_{i=0}^{N_n-1} \sum_{j=0}^{N_n-1} | \text{cov}(f_{b_n}^i, f_{b_n}^j) | \\
&= \sum_{i=0}^{b_n-1} \sum_{j=0}^{N_n-1} + \sum_{i=b_n}^{N_n-1-b_n} \sum_{j=0}^{i-b_n} + \sum_{i=b_n}^{N_n-1-b_n} \sum_{j=i-b_n+1}^{i+b_n-1} + \sum_{i=b_n}^{N_n-1-b_n} \sum_{j=i+b_n}^{N_n-1} + \sum_{i=N_n-b_n}^{N_n-1} \sum_{j=0}^{N_n-1} \\
&= A_{1,n} + A_{2,n} + A_{3,n} + A_{4,n} + A_{5,n} \quad (\text{say}),
\end{aligned}$$

where summands are suppressed on the right hand side of the first equality. Then a modification of the proof of Carlstein (1986), Theorem 2 yields the result. \square

A main tool for proving Theorem 1(b) is an exponential inequality of the sum of strong mixing random variables. For a recent development of this type of inequalities, we refer Bertail (1997) and Bosq (1993). For a triangular array of strong mixing sequences of bounded random variables, Rhomari (1993) proved an exponential inequality and derived a ready-to-use corollary for our purpose. Let $\{X_t\}_{t \geq 1}$ be strong mixing and let $h_n : \mathbf{R}^{b_n} \rightarrow \mathbf{R}$ be a measurable function such that

$$| h_n(X_{i+1}, X_{i+2}, \dots, X_{i+b_n}) | \leq M_n \text{ and } E \{ h_n(X_{i+1}, X_{i+2}, \dots, X_{i+b_n}) \} = 0$$

for every $i \in \mathbf{N}$. Then

Lemma 1 (Rhomari, 1993) *If $b_n < p_n \leq N_n/2$, then*

$$\begin{aligned}
\text{pr} \left\{ \left| \sum_{i=0}^{N_n-1} h_n(X_{i+1}, X_{i+2}, \dots, X_{i+b_n}) \right| > \varepsilon N_n \right\} &\leq 4 \exp \left\{ - \frac{N_n \varepsilon^2}{8 M_n (4 \alpha_{n,p_n}^* M_n + p_n \varepsilon)} \right\} \\
&+ 33 \left(\frac{M_n}{\varepsilon} \right)^{\frac{1}{2}} \frac{N_n}{p_n} \alpha(p_n - b_n),
\end{aligned}$$

where $\alpha_{n,p_n}^* = 1 + 2b_n + 8 \sum_{i=1}^{p_n-b_n} \alpha(i)$.

As is noted in Rhomari's paper, if $\sum_{i=1}^{\infty} \alpha(i) < \infty$, then $\alpha_{n,p_n}^* \leq cb_n$ for some $c > 0$.

Proof of Theorem 1(b). Let $M_n = n^\eta$, where η is a constant such that $(\gamma - 1)^{-1} < \eta < (1 - \delta)/2$. This is possible because of the assumption on δ . Let K be a constant such that $dK > 2 + \eta/2$. Let $p_n = b_n + K \log n$. Since $p_n \leq N_n/2$ for sufficiently large n , from Lemma 1, for any $\varepsilon > 0$,

$$\text{pr} \left(\left| \sum_{i=0}^{N_n-1} f_{b_n}^i \right| > \varepsilon N_n \right) \leq \text{pr} \left\{ \left| \sum_{i=0}^{N_n-1} f_{b_n}^i I(|f_{b_n}^i| \leq M_n) \right| > \frac{\varepsilon}{2} N_n \right\}$$

$$\begin{aligned}
& + \text{pr} \left\{ \left| \sum_{i=0}^{N_n-1} f_{b_n}^i I(|f_{b_n}^i| > M_n) \right| > \frac{\varepsilon}{2} N_n \right\} \\
& \leq 4 \exp \left\{ -\frac{N_n \varepsilon^2}{C M_n^2 b_n + 8 \varepsilon M_n p_n} \right\} + 33 \left(\frac{2 M_n}{\varepsilon} \right)^{\frac{1}{2}} \frac{N_n}{p_n} \alpha(p_n - b_n) \\
& + \frac{2}{\varepsilon} N_n^{-1} E \left\{ \left| \sum_{i=0}^{N_n-1} f_{b_n}^i I(|f_{b_n}^i| > M_n) \right| \right\}. \tag{10}
\end{aligned}$$

It is straightforward to show that the first two terms in (10) are summable. It follows from Hölder's and Markov's inequalities that the third term is bounded by $C n^{-\eta(\gamma-1)}$ for some $C > 0$ and thus it is summable since $\eta(\gamma-1) > 1$. Borel-Cantelli lemma completes the proof. \square

The proof of Theorem 3 is straightforward and is omitted.

Proof of Theorem 4. (a). Let $\alpha(p) := \{\alpha_1(p), \alpha_2(p), \dots, \alpha_p(p)\}'$ be the probability limit of $\hat{\alpha}_n(p)$. Define $X_{n,p} = (X_n, X_{n-1}, \dots, X_{n-p+1})'$, and $\tilde{X}_{n+1} = \alpha(p)' X_{n,p}$. Since $\{X_t\}_{t \geq 1}$ is causal (Brockwell and Davis, 1990, pp. 83), it follows from Theorem 2 of von Bahr and Esseen (1965) and Theorem 4 of von Bahr (1965) that $E(|X_t|^s) < \infty$ for some $s > 0$ if $E(|\varepsilon_t|^s) < \infty$. Once we observe that $\{\hat{X}_{n+1}(p) - X_{n+1}\}^4$ is bounded by $2^3 [\|\hat{\alpha}_n(p) - \alpha(p)\|^4 \|X_{n,p}\|^4 + (\tilde{X}_{n+1} - X_{n+1})^4]$, it can be readily shown from Hölder's inequality and Lemma A.2 of KY that $\{(\hat{X}_{n+1}(p) - X_{n+1})^4\}_{n \geq 1}$ is uniformly integrable. Theorem 2(a) completes the proof.

(b). Similar to the proof of (a). Use Theorem 2(b). \square

The next lemma will be used in the proof of Theorem 5.

Lemma 2 *Suppose Assumptions B1, B2, B3(s), B4 and B5(p) hold for some $s \geq 2$.*

(a) *If $s = 28$ and $b_n = o(n^{1/3})$, then*

$$\widehat{\text{PMSE}}_{n,b_n}(p) = \sigma^2(p, p_0) + b_n^{-1} \eta(p, p_0) + o_p(b_n^{-1}). \tag{11}$$

(b) *If $s = 12\gamma$ for some integer $\gamma \geq 4$ and $b_n = o(n^\delta)$ where $0 < \delta < (\gamma - 3)/(3\gamma - 1)$, then the remainder term $o_p(b_n^{-1})$ in (11) is replaced by $o(b_n^{-1})$ almost surely.*

Proof. (a). It is enough to show that $\widehat{\text{PMSE}}_{n,b_n}(p) - \text{PMSE}_{b_n}(p) = o_p(b_n^{-1})$. First we decompose the variance of $\widehat{\text{PMSE}}_{n,b_n}(p)$ into the five terms which are similar to those in the proof of Theorem 1(a). Schwarz's and Davydov's inequalities yield the result.

(b) Similar to the proof of Theorem 1(b).□

Proof of Theorem 5. (a). Let $p < p_0$. Then it follows from Theorem 3 of KY that $\lim_{n \rightarrow \infty} \{\text{PMSE}_n(p) - \text{PMSE}_n(p_0)\} = \sigma^2(p, p_0) - \sigma^2(p_0, p_0) > 0$. Thus it follows from Lemma 2(a) that

$$\text{pr} \left\{ \widehat{\text{PMSE}}_{n, b_n}(p_0) - \widehat{\text{PMSE}}_{n, b_n}(p) > 0 \right\} \rightarrow 0. \quad (12)$$

Next, let $p \geq p_0$. From the proof of Theorem 4 in KY, it is easily seen that if Assumptions B1, B2, B3(s) and B5(p) hold with $s = 32$, then for $p \geq p_0$

$$\lim_{n \rightarrow \infty} n \{\text{PMSE}_n(p+1) - \text{PMSE}_n(p)\} = a' \Gamma a,$$

where a is a nonzero constant and its form is given in KY, and Γ is the variance-covariance matrix of (X_1, X_2, \dots, X_p) . Since Γ is positive definite and $a \neq 0$, $a' \Gamma a > 0$. This implies that $\sigma^2(p, p_0) = \sigma^2(p_0, p_0)$ and $\eta(p, p_0) > \eta(p_0, p_0)$ for $p > p_0$. Again, by Lemma 2(a), it is readily shown that (12) holds. Therefore

$$\text{pr}(|\hat{p}_{n, b_n} - p_0| > \varepsilon) \leq \sum_{\substack{1 \leq p \leq K \\ p \neq p_0}} \text{pr} \left\{ \widehat{\text{PMSE}}_{n, b_n}(p) < \widehat{\text{PMSE}}_{n, b_n}(p_0) \right\} \rightarrow 0,$$

as $n \rightarrow \infty$.

(b) Similar to (a).□

Appendix 2

Additional results

The proofs of Theorem 6 and 7 are omitted and are available from the author.

Theorem 6 *Let $\{X_t\}_{t=-\infty}^{\infty}$ be an AR process (6) satisfying Assumptions B1, B2, B3(s), B4, B5(p) with $s = 20 + \varepsilon$ for some $\varepsilon > 0$. Let $\hat{X}_{n+1} = \hat{X}_{n+1}(p)$ be defined by (7). Then conditions (i)-(v), (iv)' and (v)' of Theorem 3 are satisfied.*

Theorem 7 *Let $\{X_t\}_{t=-\infty}^{\infty}$ be an $AR(p_0)$ process, where p_0 is finite, satisfying Assumption B1, B2, B3(s), B4 and B5(p) with $s = 20$. Let $\widehat{\text{PMSE}}_{n, b_n}(p)$ be defined as in (9). If $p \geq p_0$, then*

$$\text{var} \left\{ \widehat{\text{PMSE}}_{n, b_n}(p) \right\} = C_4 n^{-1} + o(n^{-1}), \quad (13)$$

where C_4 is a positive constant.

References

- Ansley, C.F. and Newbold, P. (1981). On the bias in estimates of forecast mean square error. *J. Amer. Statist. Ass.* **76**, 569-578.
- Bhansali, R. J. (1981). Effects of not knowing the order of an autoregressive process on the mean squared error of prediction-I. *J. Amer. Statist. Ass.* **76**, 588-597.
- Bhansali, R. J. (1992). Autoregressive estimation of the prediction mean squared error and an R^2 measure : an application. In *New Directions in Time Series Analysis, Part I* pp. 9-24. Ed. D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt and M. Taqqu. Springer-Verlag, New York.
- Bertail, P. (1997). Second order properties of an extrapolated bootstrap without replacement: the i.i.d and strong mixing cases. *Bernoulli* **3**(2), 149-179.
- Bickel, P. J., Götze, F. and van Zwet, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Stat. Sinica* **7**, 1-31.
- Bosq, D. (1993) Bernstein-type large deviations inequalities for partial sums of strong mixing processes. *Statistics* **24**, 59-70.
- Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods, Second Edition*. Springer-Verlag, New York.
- Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14**, 1171-1179.
- Davydov, Y. A. (1970). The invariance principle for stationary processes. *Theory Prob. Appl.* **14**, 487-498.
- Doukhan, P. (1994). *Mixing: Properties and Examples*. Springer-Verlag, New York.
- Fuller, W. A. and Hasza, D. P. (1981). Properties of predictors for autoregressive time series. *J. Amer. Statist. Ass.* **76**, 155-161.
- Hall, P., Horowitz, J. L. and Jing, B. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika* **82**, 561-574.
- Hall, P and Jing, B. (1996). On sample reuse methods for dependent data. *J. R. Statist. Soc. B* **58**, 727-737.
- Hannan, E. J. and Nicholls D. F. (1977). The estimation of the prediction error variance. *J. Amer. Statist. Ass.* **72**, 834-840.

- Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika* **83**, 875-890.
- Kunitomo, N. and Yamamoto, T. (1985). Properties of predictors in misspecified autoregressive time series models. *J. Amer. Statist. Ass.* **80**, 941-950
- Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217-1241.
- Liu, R. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap*. Ed. R. Lepage and L. Billard, pp. 225-248. Wiley, New York.
- Politis, D. N. and Romano, J. P. (1994). On the sample variance of linear statistics derived from mixing sequences. *Stochastic Process Appl.* **45**, 155-167.
- Politis, D. N., Romano, J. P. and Wolf, M. (1997). Subsampling for heteroskedastic time series. *J. Econometrics.* **81**, 281-317
- Rhomari, N. (1993). Remarque sur l'inégalité de type exponentiel pour des sommes partielles d'un processus fortement mélangeant. Preprint, L.S.T.A. Paris VI and CREST-ENSAE.
- Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147-164.
- Sin, C. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *J. Econometrics* **71**, 207-225.
- Stine, R. A. (1987). Estimating properties of autoregressive forecasts. *J. Amer. Statist. Ass.* **82**, 1072-1078.
- Tong, H. (1990). *Non-linear Time Series: A Dynamical Systems Approach*, Oxford University Press, Oxford.
- Yao, Q. and Tong, H. (1994). On subset selection in non-parametric stochastic regression. *Stat. Sinica* **4**, 51-70.
- von Bahr, B. (1965). On the convergence of moments in the central limit theorem. *Ann. Math. Statist.* **36**, 808-818.
- von Bahr, B. and Esseen, C. (1965). Inequalities for the r th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.* **36**, 299 -303.

Table 1: Distribution of the order selector $\hat{p}_{n,b}$ for AR model with various subsample sizes b . Samples of size $n = 50$ are generated from the DGP 1. Number of simulation is 500.

b	$n = 50$ Selected AR model order				
	1	2	3	4	5
7	26.4	73.6	0.0	0.0	0.0
8	10.4	89.6	0.0	0.0	0.0
9	3.6	96.4	0.0	0.0	0.0
10	2.4	97.4	0.0	0.0	0.0
11	1.4	98.2	1.6	0.0	0.0
12	0.2	98.2	1.6	0.0	0.0
13	0.8	95.4	3.8	0.0	0.0
14	0.6	94.6	4.4	0.4	0.0
15	0.2	93.4	6.2	0.2	0.0
20	1.6	83.8	10.4	3.0	1.2
25	1.2	78.4	13.4	5.4	1.6
30	2.8	67.8	16.6	8.4	4.4
35	4.2	57.6	17.0	12.4	8.8
40	8.4	48.2	19.6	12.4	11.4

Table 2: Distribution of the order selector \hat{p}_n for AR model with various values of (p_1, m) . Samples of size $n = 50$ are generated from the DGP 1. Number of simulation is 500.

(p_1, m)	$n = 50$ Selected AR model order				
	1	2	3	4	5
(1,10)	0.4	95.2	4.0	0.4	0.0
(1,20)	1.4	93.8	3.6	0.8	0.4
(2,10)	0.6	89.4	8.2	1.6	0.2
(2,20)	0.8	87.8	8.8	2.0	0.6
(3,10)	1.2	87.2	8.6	2.6	0.4
(3,20)	0.8	82.4	11.0	4.4	1.4
(4,10)	0.4	79.8	14.6	4.0	1.2
(4,20)	1.2	79.2	13.6	4.2	1.8
(5,10)	2.0	75.8	12.4	6.8	3.0
(5,20)	1.2	77.6	14.0	5.2	2.0

Table 3: Distribution of the model selector \hat{M}_n with various value s of (p_1, m) . Samples of size $n = 100$ are generated from the DGP 2. Number of simulation is 500.

(p_1, m)	$n = 100$					
	Selected	model		TAR	model	
	AR	model	order	1	2	order
	1	2	3	1	2	3
(1,40)	39.6	2.4	1.6	53.6	2.6	0.2
(1,60)	31.8	3.0	2.4	60.4	1.6	0.8
(2,40)	21.4	4.6	3.0	64.0	4.8	2.2
(2,60)	20.6	4.2	2.0	64.0	7.4	1.8
(3,40)	16.2	5.2	2.2	65.0	8.6	2.8
(3,60)	16.8	4.8	2.0	65.4	8.0	3.0

Table 4: Distribution of the model selector \hat{M}_n with various value s of (p_1, m) . Samples of size $n = 200$ are generated from the DGP 2. Number of simulation is 250.

(p_1, m)	$n = 200$					
	Selected	model		TAR	model	
	AR	model	order	1	2	order
	1	2	3	1	2	3
(1, 50)	20.8	0.8	0.4	76.0	2.0	0.0
(1,100)	14.0	0.0	0.0	82.0	3.6	0.4
(2, 50)	7.2	0.0	0.8	87.2	4.4	0.4
(2,100)	4.0	0.4	0.8	85.6	7.6	1.6
(3, 50)	4.8	0.8	1.2	85.6	6.4	1.2
(3,100)	4.4	0.4	1.2	82.4	10.0	1.6