The Annals of Statistics
1989, Vol. 17, No. 2, 941–946

STRONG CONSISTENCY OF THE PLS CRITERION FOR ORDER DETERMINATION OF AUTOREGRESSIVE PROCESSES

BY E. M. HEMERLY[1] AND M. H. A. DAVIS

Imperial College of Science and Technology, London

This note concerns the problem of order determination for autoregressive models. Rissanen's "Predictive least squares principle" prescribes that one should choose as order estimate $\hat{k}(n)$ at time n the order of the model which has given the least mean square prediction error up to that time. We show that this procedure is strongly consistent, that is, that $\hat{k}(n) \to p$ a.s. as $n \to \infty$ when the data are generated by an AR process of order p, given an upper bound p^*.

1. Introduction. Several criteria have been proposed to solve the problem of order determination for autoregressive processes. As representative works we can mention Anderson's (1963) multiple decision procedure, Akaike's (1974) AIC criterion, whose consistency properties were analysed by Shibata (1976), and Rissanen's (1978, 1980) MDL criterion [see also Schwarz (1978)] and the $\phi(k)$ criterion proposed by Hannan and Quinn (1979). An altogether different criterion has been proposed by Rissanen (1986a). The corresponding principle of modelling, the PMDL (predictive minimum description length principle), unlike the maximum likelihood method, permits optimal identification of parameters both in their values and in their number. When specialized to Gaussian models, the PMDL gives rise to the PLS (predictive least squares principle), Rissanen (1986b). Whereas the usual least squares minimizes the mean prediction error, the PLS minimizes the prediction errors on the observations. In so doing, the minimized criterion can be interpreted as representing the least total accumulated "honest" prediction errors (where the "honest" denotes that only past data are used to identify the predictor parameters), and as being the stochastic complexity of the data.

The first study concerning the consistency of the PLS was carried out by Rissanen (1986b), who considered linear regression models with Gaussian noise. Based on first and second moments of some random variables of interest and relying on Chebyshev's inequality, it is shown that $\hat{k}(n) \to p$ in probability, where p is the dimension of the regressor vector, with $p \in M = \{1, 2, \ldots, p^*\}$, $p^* < \infty$, and $\hat{k}(n)$ is the PLS estimate at time n. The next study was done by Wax (1986, 1988), who obtained the same result for autoregressive models without requiring the Gaussian assumption. Besides giving the proof, Wax

Received November 1987; revised June 1988.
[1]Research supported by ITA and CAPES, Brazil.
AMS 1980 *subject classifications*. Primary 62M10, 93E12; secondary 60F15, 62M20.
Key words and phrases. Autoregressive process, martingale difference, order determination, predictive least squares, strong consistency, structure identification.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*.
www.jstor.org

presented a computationally efficient way of evaluating the PLS estimates by using predictive lattice filters.

In this work, by relying on Wei's (1987) result for multiple regression models, we show that as conjectured by Rissanen (1986c) the PLS estimates are also strongly consistent, that is, $\hat{k}(n) \to p$ almost surely for AR models.

2. Problem formulation and regularity assumptions. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t=0,1,\ldots}, P)$ be a filtered probability space and $\{w(t)\}$ a martingale difference process with respect to $\mathcal{F}_t$. The sampled data $\{y(t),\ t \geq 0\}$, with $y(t) = 0,\ \forall\ t < 0$, is generated by the pth-order autoregressive process

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \cdots + a_p y(t-p) + w(t). \tag{1}$$

We assume that an upper bound p^* for the model order p is known. Competing models of order $k = 1, 2, \ldots, p^*$ are fitted by ordinary least squares. Let

$$\hat{\Theta}^{\mathrm{T}}(k,t) = [\hat{a}_{k,1}(t) \quad \hat{a}_{k,2}(t) \quad \ldots \quad \hat{a}_{k,k}(t)] \tag{2}$$

denote the estimated predictor coefficients in a kth-order model at time t; then

$$\Theta(k,t) = \left(\sum_{j=1}^{t} \Phi(k,j)\Phi^{\mathrm{T}}(k,j) \right)^{-1} \sum_{j=1}^{t} \Phi(k,j)y(j), \tag{3}$$

where $\Phi(k,t)$ denotes the regressor vector

$$\Phi^{\mathrm{T}}(k,t) = [y(t-1) \quad y(t-2) \quad \ldots \quad y(t-k)]. \tag{4}$$

Now define

$$e(k,t+1) = y(t+1) - \hat{y}(k,t+1) = y(t+1) - \Phi^{\mathrm{T}}(k,t+1)\hat{\Theta}(k,t) \tag{5}$$

[this is the "honest" prediction error, in the sense that calculation of $\hat{y}(k,t+1)$ involves only the data $y(1), \ldots, y(t)$] and let

$$\mathrm{PLS}(k,n) = (1/n) \sum_{t=1}^{n} e^2(k,t). \tag{6}$$

The order estimate $\hat{k}(n)$ at time n is then

$$\hat{k}(n) = \min_{k \in M} \mathrm{PLS}(k,n), \tag{7}$$

where $M = \{1, 2, \ldots, p^*\}$. Thus $\hat{k}(n)$ is the order of the model which has given the least mean square prediction error up to time n. To state our results, the following mild regularity conditions on the process (1) are required.

(A.1) The roots of the characteristic polynomial $z^p - a_1 z^{p-1} - \cdots - a_p = 0$ are all inside the unit circle.

(A.2) The linear innovations $\{w(t)\}$ satisfy

$$\begin{gathered} E[w(t)|\mathcal{F}_{t-1}] = 0 \quad \text{a.s.}, \qquad E[w^2(t)|\mathcal{F}_{t-1}] = \sigma^2 \quad \text{a.s.}, \\ E[|w|^{\alpha}|\mathcal{F}_{t-1}] < \infty \quad \text{a.s. for some } \alpha > 2. \end{gathered} \tag{8}$$

THEOREM 1. *Suppose conditions* (A.1) *and* (A.2) *hold. Then the* PLS *criterion is strongly consistent, that is,*

$$\hat{k}(n) \to p \quad a.s.\ as\ n \to \infty. \tag{9}$$

The proof is given below in Section 3. Having established (9), it is immediate that under conditions (A.1) and (A.2) we have strong consistency of the parameter estimates, that is,

$$\hat{\Theta}\big(\hat{k}(n), n\big) \to \Theta(p) \quad \text{a.s. as } n \to \infty, \tag{10}$$

where $\Theta(p) = [a_1\ a_2\ \dots\ a_p]$.

3. Proof of Theorem 1. Since the arguments for the undermodelled ($k < p$) and the overmodelled ($k > p$) cases are different, we will consider them separately.

3.1. *Overmodelled case.* We rewrite (6) as

$$\begin{aligned} n\,\mathrm{PLS}(k,n) &= \sum_{t=1}^{n} w^2(t) + 2\sum_{t=1}^{n} (e(k,t) - w(t))w(t) \\ &\quad + \sum_{t=1}^{n} (e(k,t) - w(t))^2 \end{aligned} \tag{11}$$

and since $(e(k,t) - w(t))$ is $\mathcal{F}_{t-1}$-measurable and $\{w(t)\}$ is a martingale difference, from Chow (1965) we have

$$\begin{aligned} &n\,\mathrm{PLS}(k,n) \\ &\quad = \sum_{t=1}^{n} w^2(t) + (1 + o(1)) \sum_{t=1}^{n} (e(k,t) - w(t))^2 + O(1) \quad \text{a.s.} \end{aligned} \tag{12}$$

Now, from Wei (1987), Theorem 3, we will have

$$\sum_{t=1}^{n} (e(k,t) - w(t))^2 = (1 + o(1))\sigma^2 \log\det \sum_{t=1}^{n} \Phi(k,t)\Phi^{\mathrm{T}}(k,t) \quad \text{a.s.} \tag{13}$$

if

$$\Phi^{\mathrm{T}}(k,n)\left(\sum_{t=1}^{n} \Phi(k,t)\Phi^{\mathrm{T}}(k,t)\right)^{-1} \Phi(k,n) \to 0 \quad \text{a.s. as } n \to \infty \tag{14}$$

and

$$\lambda_{\min}\left(D^{-1}(k,n)\left(\sum_{t=1}^{n} \Phi(k,t)\Phi^{\mathrm{T}}(k,t)\right)D^{-1}(k,n)\right) \to \infty \quad \text{a.s. as } n \to \infty,$$

$$\text{where } D(k,n) = \left\{\mathrm{diag}\left(\sum_{t=1}^{n} \Phi(k,t)\Phi^{\mathrm{T}}(k,t)\right)\right\}^{1/2}, \tag{15}$$

where $\lambda_{\min}$ stands for minimum eigenvalue. These conditions are obviously satisfied because (A.1) and (A.2) imply

$$(16)\qquad (1/n)\sum_{t=1}^{n}\Phi^{\mathrm{T}}(k,t)\Phi(k,t)\to R(k)\quad \text{a.s. as } n\to\infty,$$

$$\text{where } R(k)=E\left[\Phi(k,n)\Phi^{\mathrm{T}}(k,n)\right].$$

Therefore, from (13) and (16),

$$(17)\qquad \sum_{t=1}^{n}\left(e(k,t)-w(t)\right)^2=(1+o(1))\sigma^2 k\log n\quad \text{a.s.}$$

and then, from (12) and (17),

$$(18)\qquad n(\mathrm{PLS}(k,n)-\mathrm{PLS}(p,n))=(1+o(1))\sigma^2(k-p)\log n\quad \text{a.s.},$$

which obviously implies

$$(19)\qquad \mathrm{PLS}(k,n)-\mathrm{PLS}(p,n)>0\quad \text{a.s. for } n \text{ large enough, } \forall\, k>p.$$

3.2. *Undermodelled case.* The model set now is considered to be $\{1,2,\ldots,p-1,p\}$, where p is the order of the system generating the data. The analysis here is straightforward because all models of smaller order than p give asymptotically a variance which is larger than σ^2.

For any model with order $k<p$, we define

$$(20)\qquad \tilde{\Theta}(p,t)=\left[a_1-\hat{a}_{k,1}(t-1)\quad\ldots\quad a_k-\hat{a}_{k,k}(t-1)\quad a_{k+1}\quad\cdots\quad a_p\right]$$

and from (1), (2) and (4)–(6) we obtain

$$\begin{aligned} n\,\mathrm{PLS}(k,n) &= \sum_{t=1}^{n}\left(\Theta^{\mathrm{T}}(p)\Phi(p,t)-\hat{\Theta}^{\mathrm{T}}(k,t-1)\Phi(k,t)+w(t)\right)^2\\ &= \sum_{t=1}^{n}\left(\tilde{\Theta}^{\mathrm{T}}(p,t)\Phi(p,t)+w(t)\right)^2\\ &= (1+o(1))\sum_{t=1}^{n}\left(\tilde{\Theta}^{\mathrm{T}}(p,t)\Phi(p,t)\right)^2+\sum_{t=1}^{n}w^2(t)+O(1)\quad \text{a.s.}, \end{aligned}\qquad (21)$$

where the last equality follows from Chow (1965) and the fact that $\tilde{\Theta}^{\mathrm{T}}(p,t)\Phi(p,t)$ is $\mathcal{F}_{t-1}$-measurable. Considering now that $\tilde{\Theta}(p,t)$ is converging a.s. as $t\to\infty$, say to $\tilde{\Theta}(p)$, we can rewrite (21) as

$$\begin{aligned} n\,\mathrm{PLS}(k,n) &= (1+o(1))\tilde{\Theta}^{\mathrm{T}}(p)\left(\sum_{t=1}^{n}\Phi(p,t)\Phi^{\mathrm{T}}(p,t)\right)\tilde{\Theta}(p)\\ &\quad+\sum_{t=1}^{n}w^2(t)+O(1)\quad \text{a.s.}, \end{aligned}\qquad (22)$$

where, from (20), we can estimate

$$\tilde{\Theta}^{\mathrm{T}}(p)\left(\sum_{t=1}^{n}\Phi^{\mathrm{T}}(p,t)\Phi(p,t)\right)\tilde{\Theta}(p)$$

$$\geq |\tilde{\Theta}(p)|^2\lambda_{\min}\left(\sum_{t=1}^{n}\Phi(p,t)\Phi^{\mathrm{T}}(p,t)\right) \tag{23}$$

$$\geq a_p^2\lambda_{\min}\left(\sum_{t=1}^{n}\Phi(p,t)\Phi^{\mathrm{T}}(p,t)\right).$$

Recalling that from (12) and (17)

$$n\,\mathrm{PLS}(p,n) = (1+o(1))\sigma^2 p\log n + \sum_{t=1}^{n} w^2(t) \quad \text{a.s.}, \tag{24}$$

from (22)–(24) results

$$\mathrm{PLS}(k,n) - \mathrm{PLS}(p,n)$$

$$\geq (1+o(1))a_p^2\frac{1}{n}\lambda_{\min}\left(\sum_{t=1}^{n}\Phi(p,t)\Phi^{\mathrm{T}}(p,t)\right) + O\left(\frac{\log n}{n}\right) \quad \text{a.s.}, \tag{25}$$

and since from (16)

$$\liminf_{n\to\infty}(1/n)\lambda_{\min}\left(\sum_{t=1}^{n}\Phi(p,t)\Phi^{\mathrm{T}}(p,t)\right) > 0 \quad \text{a.s.}, \tag{26}$$

from (25) we conclude that

$$\mathrm{PLS}(k,n) - \mathrm{PLS}(p,n) > 0 \quad \text{a.s. for } n \text{ large enough}, \forall\, k < p, \tag{27}$$

and then, from (19) and (27), (9) is proved. □

4. Final remarks. The extension to ARMA models is not straightforward. This is so because in the ARMA overmodelled case the parameter estimates provided by the recursive prediction error method may not be well defined. Recently, however, Veres (1988) has established results concerning this case.

We should mention that the strong consistency of the PLS criterion and the possibility of evaluating it in real time suggest its application in areas as signal processing and adaptive control, among others, where recursive computation is essential. Moreover, the recursive computation can be carried out in an efficient way by using the lattice form for parameter estimation, since in this case all the prediction errors $e(1,t),\ldots,e(p^*,t)$ are calculated at once [see Wax (1988) for details].

Finally, an independent, and much more involved, proof of the strong consistency of the PLS criterion for AR processes has been provided by Hannan, McDougall and Poskitt (1987).

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723.

ANDERSON, T. W. (1963). Determination of the order of dependence in normally distributed time series. In *Proc. Symposium Time Series Analysis Brown Univ.* (M. Rosenblatt, ed.) 425–446. Wiley, New York.

CHOW, Y. S. (1965). Local convergence of martingales and the law of large numbers. *Ann. Math. Statist.* **36** 552–558.

HANNAN, E. J. and QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195.

HANNAN, E. J., MCDOUGALL, A. J. and POSKITT, D. S. (1989). Recursive estimation of autoregressions. *J. Roy. Statist. Soc. Ser. B* **51**. To appear.

RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

RISSANEN, J. (1980). Consistent order estimates of autoregressive processes by shortest description of data. In *Analysis and Optimization of Stochastic Systems* (O. Jacobs, M. Davis, M. Dempster, C. Harris and P. Parks, eds.). Academic, New York.

RISSANEN J. (1986a). Stochastic complexity and modeling. *Ann. Statist.* **14** 1080–1100.

RISSANEN J. (1986b). A predictive least-squares principle. *IMA J. Math. Control Inform.* **3** 211–222.

RISSANEN, J. (1986c). Order estimation by accumulated prediction errors. In *Essays in Time Series and Allied Processes* (J. Gani and M. B. Priestley, eds.). *J. Appl. Probab.* **23A** 55–61.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126.

VERES, S. (1988). Strong consistency for the accumulated prediction error criterion for multivariate processes. Preprint, Imperial College London.

WAX, M. (1986). Order selection for AR models by predictive least-squares. *Proc. of Twenty-Fifth CDC* 1481–1486.

WAX, M. (1988). Order selection for AR models by predictive least squares. *IEEE Trans. Acoust. Speech Signal Process.* **36** 581–588.

WEI, C. Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Ann. Statist.* **15** 1667–1687.

INSTITUTO TECHNOLÓGICO DE AERONÁUTICA
DIVISÃO DE ELETRÔNICA—CTA
12225 SÃO JOSÉ DOS CAMPOS-SP
BRAZIL

DEPARTMENT OF ELECTRICAL ENGINEERING
IMPERIAL COLLEGE OF SCIENCE AND TECHNOLOGY
EXHIBITION ROAD
LONDON SW7 2BT
ENGLAND