Order Estimation by Accumulated Prediction Errors

JORMA RISSANEN

Abstract

This paper presents a new criterion based on prediction error which allows the estimation of the number of parameters as well as structures in statistical models. The criterion is valid for short and long samples alike. Unlike Akaike's earlier criterion, also based on prediction error, the criterion proposed here appears to produce consistent error estimates in ARMA processes.

ARMA PROCESSES; MINIMUM DESCRIPTION LENGTH; MODEL COST; PARAMETERS; CONSISTENCY

1. Introduction

Based upon the reasoning that since the ultimate use of most models of time series is to provide predictions, there has long been a desire to base the entire estimation procedure on minimization of prediction errors. Indeed, such estimators have been shown to possess desirable properties, and, in fact, in the Gaussian ARMA processes they are comparable with the ML-estimators (Ljung and Caines (1979)). Moreover, as shown by Davisson (1965) and Akaike (1974), if one includes in the prediction error the effect of the estimation errors in the parameters, a criterion results which automatically penalizes the number of parameters in the model. This is an important innovation, for in the past a separate hypothesis testing was required to estimate the number of the parameters.

A meticulous analysis, above all by Shibata (1976), (1980), has revealed that the order estimates minimizing Akaike's AIC criterion, which in the Gaussian ARMA processes is quite equivalent to Davisson's prediction error criterion, have interesting and useful properties, except that they are not consistent even in the case of AR processes. Although one may argue that consistency in itself is not all that important, in particular if the 'true' data-generating system is infinite-dimensional, this author nevertheless feels that the AIC criterion fails an analyzable test of performance. And this, of course, does nothing to increase one's confidence in it when it is to be applied to non-analyzable cases. We do not, incidentally, accept the premise that any physical system is infinite-dimensional. In fact, we regard the finite amount of observed data to *be* the 'true' system, and together with possible prior information the data are all we have about the process; any other 'system' explanation we invent is simply a model, which may well be infinite-dimensional, and as such it will have to compete with other models we care to consider.

Prompted by this shortcoming of the AIC criterion, we proposed in Rissanen (1978), (1983a), (1985) an altogether different principle, the so-called MDL principle (MDL for minimum description length), based on a purely information-theoretic idea: pick the parameters so that the model they define permits a redescription of the observed sequence with the smallest number of binary digits. This principle indeed has been shown to produce consistent order estimates in ARMA processes (Hannan (1980)). But despite the soundness and success of the MDL principle, the failure of such an intuitively attractive principle as one based on prediction errors remained in this author's mind as a puzzling issue, so much so that he suspected that the prediction errors are not fairly and properly represented in the above-mentioned attempts. This, indeed, appears to be the case, and when corrected we do arrive at a criterion based on prediction error, which is not only valid asymptotically like the previous ones but which is also perfectly justified even for short samples. This criterion appears to be asymptotically equivalent to the MDL criterion for Gaussian ARMA processes, and hence their estimates should be consistent.

2. Accumulated prediction error criterion

We consider an observed sample $x = x_1, \cdots, x_n$, where the numbers x_i are delivered to us one after another so that at every time instant $t = 0, 1, \cdots, n-1$ we are given the past sequence $x^t = x_0 x_1 \cdots x_t$. Here, x_0 is a constant, say 0, representing the string of no observations. Suppose now that at each t we are to make a prediction of the next value x_{t+1}, based upon the sequence x^t so far seen. How should we form a measure of the prediction errors? It seems quite natural to define the following accumulated measure:

$$V(k, x) = n^{-1} \sum_{0}^{n-1} (x_{t+1} - \hat{x}_{t+1})^2, \tag{2.1}$$

where $\hat{x}_{t+1} = f(x^t, \theta(x^t))$ denotes the prediction made at time t based upon the past sequence x^t with use of parameters estimated in some way, collected in the k-component vector $\theta(x^t)$, which also must depend only on the past observations. Applying the sensible reasoning that we should act on the principle that has worked best in the past (indeed, we cannot think of a better principle for statistical inference!), these estimates should clearly be determined by minimization of the summed past prediction errors. The number of parameters k is then determined so that $V(k, x)$ is minimized.

We illustrate the use of this estimation principle in the case of ARMA models. Suppose that we are willing to model the data as being a sample from some stationary zero-mean one-sided moving-average ARMA (p, q) process:

$$x_t + a_1 x_{t-1} + \cdots + a_p x_{t-p} = e_t + b_1 e_{t-1} + \cdots + b_{t-q} e_{t-q} \tag{2.2}$$

where $\{e_t\}$ is an uncorrelated, zero-mean process such that the sequence $\{e_t, e_{t-1}, \cdots\}$ spans the same linear space as $\{x_t, x_{t-1}, \cdots\}$. The process x is then defined up to the second moments by the $p+q$ parameters $\theta = (a_1, \cdots, a_p, b_1, \cdots, b_q)$ and the variance σ^2 of the process e_t.

In view of the class of models selected we consider a linear predictor with the prediction error $\varepsilon_{t+1} = x_{t+1} - \hat{x}_{t+1}$ given by $\varepsilon_{t+1} = e_{t+1,t}$, which is determined by $p+q$ parameters $\theta(t) = (a_{1,t}, \cdots, a_{p,t}, b_{1,t}, \cdots, b_{q,t})$ and the data as follows: for $i = 0, \cdots, t+1$, put

$$e_{i,t} + b_{1,t} e_{i-1,t} + \cdots + b_{q,t} e_{i-q,t} = x_i + a_{1,t} x_{i-1} + \cdots + a_{p,t} x_{i-p}, \tag{2.3}$$

where $x_i = e_{i,t} = 0$ for $i \leqq 0$. The parameter vector $\theta(t)$, in turn, should be determined so that the criterion

$$s^2(t) = t^{-1} \sum_{0}^{t-1} e^2_{i+1,t}, \tag{2.4}$$

is minimized. Finally, the numbers p and q are determined so that the accumulated prediction errors (2.1) are minimized.

An outstanding feature of the criterion (2.1) is that it uses the same given set of observations both as the basis for estimation and as a test of the validity of the estimates. We may compare this with Akaike's criterion AIC, which can also be interpreted in terms of prediction errors. It follows from a result in Davisson (1965), when specialized to a stationary AR (p^0)-process (see also Fuller and Hasza (1981)) that

$$E_{\theta^0}(x_{n+1} - \hat{x}_{n+1})^2 = \sigma^2(1 + p/n) + o(n^{-1}) \tag{2.5}$$

where $\hat{x}_{n+1}$ denotes the predictor obtained with the least squares estimator $\theta(x^n)$ having $p, p \geqq p^0$, components. In other words,

$$\hat{x}_{n+1} = -\hat{a}_1 x_n - \cdots - \hat{a}_p x_{n-p+1},$$

where $\hat{a}_i$ are determined from the observed sequence x^n by minimization of the error squares (2.4) for $t = n$. Further, σ^2 denotes the variance of the stationary process $\{e_t\}$ which with the p^0-component θ^0 determines the process x_t. We thus see that the more parameters (above the 'true' number p^0) we pick in the model, the greater the mean prediction error (2.5). The next step is to replace the variance σ^2 by the estimate, the minimized error squares $s^2(n)$ in (2.4). This estimator has a bias, which is asymptotically given by $-\sigma^2 p/n$. By correcting this and substituting the result in (2.5), we get Akaike's AIC criterion after taking the natural logarithm:

$$2 \ln s(n) + 2p/n. \tag{2.6}$$

It seems to us that both applications of the prediction error principle are meaningful, although in (2.6) no specific samples are used as a means of validating the estimates. Instead, the asymptotic mean is taken to provide a sort of validation. However, something appears to be lost in such a substitute validation procedure, which becomes evident in the lack of consistency of the resulting order estimates. Our criterion (2.1), instead, forces a validation after each observation is received, which leads to a greater penalty on the number of parameters used. That this, in turn, should produce consistent order estimates can be seen from the asymptotic analysis carried out in the next section, but its plausibility is easy to see intuitively. It certainly seems reasonable to expect that a model's predictive capability cannot be improved by estimating excessive unnecessary parameters, while an improvement does result if a new relevant parameter is added to the model. Hence, the best predictions are obtained when the model has as many parameters as the—this time imaginary—data-generating system.

In Stone (1977) and Geisser and Eddy (1979) another predictive approach to model selection was described, which also uses the common batch of data both for estimation and validation. However, that approach is not 'honestly' predictive in the same sense as ours, and, in fact, Stone in the case with independence shows that the resulting criterion is asymptotically equivalent to Akaike's criterion. Hence, cross-validation in itself does not seem to guarantee consistency, and indeed why should it?

3. Asymptotic properties

In order to analyze further the proposed criterion and its estimators, suppose that the strings x are generated by a process in the class of Gaussian ARMA processes. We wish to find out how small the mean of the accumulated prediction error criterion (2.1) then can be made. In Rissanen (1984) we proved an asymptotic result, which states the following.

No matter how many parameters we estimate, and no matter how we estimate them, the inequality

$$E_{\theta^0} V(k, x) \geqq \sigma^2[1 + ((k^0 - \varepsilon)/n) \ln n], \tag{3.1}$$

holds for all positive numbers ε and all 'true' parameters θ^0, defining stationary the standardized quantile function $QI(u)$ defined by

$$QI(u) = Q(u) - Q(0.5)/2(Q(0.75) - Q(0.25)).$$

We call $QI(u)$ an identification-quantile function; it equals 0 at $u = 0.5$ and its slope at $u = 0.5$ is approximately equal to 1. It is an easily estimated approximation to the unit-quantile function $Q1(u)$ defined by

$$Q1(u) = \{Q(u) - Q(0.5)\}/q(0.5).$$

The corresponding density function $f1(x)$ is normalized so that the median $= 0$ and $f1(0) = 0$. The unit-density function $f1(x)$ is: $\exp(-\pi x^2)$ for normal distributions; $\exp(-2x)$, $x \geqq -0.5 \log 2$, for exponential distributions.

natural logarithmic unit in the AR case, is as follows:

$$\ln s(n) + \tfrac{1}{2}((p+1)/n) \ln n. \tag{3.3}$$

If we replace in (3.2) σ^2 by its bias-corrected estimate, as above following (2.5), we get the criterion

$$2 \ln s(n) + (p/n) \ln n, \tag{3.4}$$

which differs from (3.3) in having one fewer parameter. The extra parameter in (3.3) is σ^2, whose best estimate is $s(n)$ no matter how the other p parameters are estimated. Hence, for estimating these the two criteria are equivalent. We conjecture that the difference between $V(p, x)$ and (3.4), multiplied by $n(\ln n)^{-1}$, tends to 0. If this is true, then because the order estimates obtained by minimization of (3.3) have been shown to be consistent (Hannan and Quinn (1979), Rissanen (1980)), so should be the estimates obtained by minimization of the accumulated prediction errors (2.1).

4. Simulations

We illustrate the use of the criterion (2.1) by applying it to a sequence of observations generated by a Gaussian ARMA system. We fitted models of type ARMA (p, q) with $(p, q) = (1, 0)$, $(2, 0)$, $(1, 1)$, and $(0, 2)$. Table 1 gives the minimized criterion $V(k, x^n)$ for five different values of n along the single sample of size 600. If we add that models $(2, 2)$ and $(0, 1)$ gave uniformly worse values than the two best models $(1, 0)$ and $(1, 1)$ in the table (we did not calculate the last entry for the two worst models), the reader can conclude that a system $(1, 1)$ was the one that generated the data. Notice, however, that up to the sample size 200 the simpler first-order AR model $(1, 0)$ performed better than the eventual winner $(1, 1)$.

TABLE 1
Minimized criterion values for four models

		Length n				
		50	100	200	300	600
	(1, 0)	1.336	1.276	1.101	1.107	1.015
Model	(2, 0)	1.629	1.385	1.156	1.120	—
(p, q)	(1, 1)	1.505	1.307	1.117	1.096	0.996
	(0, 2)	1.925	1.520	1.221	1.159	—

The data-generating system had the parameters $a_1 = 0.5$ and $b_1 = -0.3$. The 300-sample estimate of the (1, 0)-model was $\hat{a}_1 = 0.59$ while the same estimates for the two parameters in the (1, 1)-model were $\hat{a}_1 = 0.335$ and $\hat{b}_1 = -0.405$. The associated minimized sample variance for the best (1, 1)-model was $s^2(300) = 1.024$, which gives the value $1.024(1 + \log 300)/300) = 1.073$ for the MDL criterion (3.3), where we now used the binary logarithm. This is a little smaller than the table entry 1.096. For the final sample size of 600 we got with the same model $s^2(600) = 0.953$, which gives the value 0.988 for the MDL criterion. This, again, is less than the corresponding table entry 0.996. We conclude that the criterion (2.1) imposes a greater penalty on the system complexity than the MDL criterion. This is particularly noticeable for short samples where the relative model cost is greater. Hence, the criterion (2.1) tends to underestimate the number of parameters, which is, perhaps, just as it ought to be. After all, the 'information' in the data, the word taken in the technical sense as the infinum of the code lengths, cannot be defined without including the estimation of the parameters, and, hence, to achieve the total information only those parameters which 'buy' enough performance should be retained. This means that initially when the sample size is small the optimum model necessarily has only a few parameters and others will be included gradually as more data is received. This is in keeping with our general philosophy that there never is any 'true' system nor a 'true' number of parameters—only an optimum number—and an excessively complex model is bad not only because of practicability reasons in being more difficult and expensive to implement, but because it performs worse.

In conclusion, we point out that the criterion (2.1) ought to give reasonable results even when used to estimate the structure of vector ARMA processes. After all, when a model in a 'bad' structure is selected, the parameters are expressed in a coordinate system with some axes tending to be near parallel, and one may expect large estimation errors and hence large prediction errors. For the estimation of structure with a three-term MDL criterion we refer to Rissanen (1983b).

References

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

DAVISSON, L. D. (1965) The prediction error of stationary Gaussian time series of unknown covariance. *IEEE Trans. Inf. Theory* **IT-11**, 527–532.

FULLER, W. A. AND HASZA, D. P. (1981) Properties of predictors for autoregressive time series. *J. Amer. Statist. Assoc.* **76**, 155–161.

GEISSER, S. AND EDDY, W. (1979) A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74**, 153–160.

HANNAN, E. J. (1980) The estimation of the order of an ARMA process. *Ann. Statist.* **8**, 1071–1081.

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

LJUNG, L. AND CAINES, P. (1979) Asymptotic normality of prediction error estimators for approximate system models. *Stochastics* **3**, 29–46.

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 Optimisation of Stochastic Systems*, ed. O. Jacobs, M. Davis, M. Dempster, C. Harris, P. Parks, Academic Press, New York.

RISSANEN, J. (1983a) A universal prior for integers and estimation by minimum description lengths. *Ann. Statist.* **11**, 416–431.

RISSANEN, J. (1983b) Estimation of structure by minimum description length. *Circuits, Systems, and Signal Processing* (*Special Issue on Rational Approximations*) **1**, 395–406.

RISSANEN, J. (1984) Universal coding information, prediction, and estimation. *IEEE Trans. Inf. Theory* **IT-30**, 629–636.

RISSANEN, J. (1985) Minimum description length principle. In *Encyclopaedia of Statistical Sciences*, Vol. 5, ed. S. Kotz and N. L. Johnson. Wiley, New York.

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

SHIBATA, R. (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147–164.

STONE, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion *J. R. Statist. Soc.* B **39**, 44–47.