LEARNING ALGORITHM FOR SELECTION OF AN AUTOREGRESSIVE MODEL FOR MULTI-STEP AHEAD FORECAST

Mikhail Prokopenko

CSIRO Division of Information Technology,
Locked Bag 17, North Ryde, NSW 2113, Australia
E-mail: mikhail@syd.dit.csiro.au, Phone: [+61 2] 325-3144, Fax: [61 2] 325-3101

ABSTRACT: This paper addresses the problem of learning an order of an autoregressive (AR) model with multi-step ahead prediction properties and describes a computational learning algorithm based on a new selection criterion (*"pattern residuals' interdependence measure estimator"* - PRIME). The PRIME criterion for the selection of an AR model with sufficient predictive power measures interdependence among residuals obtained from a given training set (an observed time series), a time series modelled by an autoregressive-moving average (ARMA) process, and a corresponding deterministic pattern extracted from the original data by a smoothing filter. As a measure of the residuals' interdependence a mean expected log-likelihood function of a correlation coefficient between the residuals is defined. The paper presents the results of Monte Carlo simulation which generate an empirical distribution of the proposed estimator and provide evidence of appropriate identification of data. It illustrates also the results obtained by multi-step ahead forecast for actual data. The comparison between models selected by the proposed criterion and well-known criteria provides favourable evidence for the strength of the PRIME criterion.

1. INTRODUCTION.

Building a system that can learn from its environment has become a major focus of research in Computational Learning Theory and in related areas of Artificial Intelligence such as Machine Learning and Neural Networks. According to the recently proposed definition [7] a system is called intelligent, if it can 1) generate and store information about its environment and its state; 2) create knowledge from this information; 3) use this knowledge for achieving its goals. In particular, the system is allowed to see a set of examples and must develop a hypothesis (a model) that can explain those examples [6]. The process of "detecting and perhaps formally expressing an apparent regularity or pattern in a body of data about many particular instances or events" can be called inductive learning [14] and covers such topics as the acquisition of general theories from particular data, and "explanation" as a way of making data less surprising or more probable under a plausible theory.

At the same time the fundamental principles of computer learning have strong interconnections with traditional methodology developed in the fields of applied mathematics (interpolation and function approximation), engineering (signal filtering and pattern recognition), statistics (regression analysis, ARMA modelling, classification). It is well known in time series analysis that the aim of model selection is not only to check the adequacy of a model, or to identify a system but also to obtain a good predictor [11]. The same idea supports the development of learning systems which attempt to extract useful prediction rules by processing data taken from the past - in other words, based only on cases for which the values of both inputs and outputs of a real system have been determined (a given training set). If the environment in which the system operates changes dynamically or/and the input information is uncertain or otherwise deficient then an accepted hypothesis (a selected model) can quickly become inadequate. These more realistic scenarios suggest to consider models capable of providing an effective multi-step ahead prediction.

The most influential work on time series forecasting was carried out by Box and Jenkins [4], who used ideas of Wold decomposition (any stationary process can be uniquely represented as the sum of linearly deterministic process and purely nondeterministic, or an MA(∞) process, and these components are mutually uncorrelated [15]), formulated the class of ARMA models and developed a model selection strategy. The basis of the Box-Jenkins forecasting method is formed by recognition of statistical patterns - “from what has been observed, infer significant characteristics of the process generating the data such as significant time lags, significant frequencies, extractable signals, and noise” [8]. The model selection strategy consists of three stages: identification, estimation and diagnostic checking. At the first stage the principle of parsimony [4] is of major importance. It is worth noting that the interpretation of learning as a competition [14] uses the similar idea of complexity: other things being equal, simple theories are to be preferred to complex ones (Occam’s Razor). Another analogy between learning and model selection is that a degree of fit to the known data is also a critical criterion. “The primary objective of the concept learner is to infer a classification rule that describes the target concept. While attempting to achieve the primary objective, the secondary objective is to infer a classification rule that is as close as possible to the target concept” [13]. If ARMA model is a classification rule, then a model order selection and parameters estimation correspond to inferring the rule “closest” to the target concept (true model).

2. MODEL SELECTION AND THE PRIME CRITERION.

A typical way of realising balancing behaviour between overfitting and underfitting risks is to select orders p (AR) and q (MA) which minimise

$$\mathbf{T + \alpha\,(p + q),} \qquad \mathbf{(1)}$$

where T is one of test statistics. The second term can be considered as a penalty term for the complexity of the model, and a term for compensating the random fluctuation of T [11]. There exist several so called criterion procedures: Akaike Information Criterion (AIC) [1,2], Schwarz Bayesian Criterion (SBC) [12], or ϕ [5]. All criteria are of the form of (1) with

$$\mathbf{T = -\,2\,\log(maximum\ likelihood)} \qquad \mathbf{(2)}$$

The choice of α depends on the aim of the selection (prediction, identifying, etc.). And “the conclusion that the optimum value of α depends, in a complicated way, on unknown parameters is unhelpful for the analysis of data” [3]. The principle of statistical model building based on the maximum likelihood estimate and the minimum information theoretic criterion estimate have been introduced by Akaike in order to explicitly formulate the problem of statistical identification as a problem of estimation and completely eliminate “the need of the subjective judgement required in the hypothesis testing procedure for the decision on the levels of significance” [2]. The well-known AIC [1,2] approximates discrimination between density functions of a true model and a candidate used for prediction and selects the model with minimal value of

$$\mathbf{(-2)\ \ \log(maximum\ likelihood)\ +\ 2k,} \qquad \mathbf{(3)}$$

where the number k of independently adjusted parameters within the model is added to correct the downward bias [2]. The criterion provides an asymptotically efficient solution to the problem but suffers some serious drawbacks. In particular, it often selects models of very high dimension if these are included as candidates. Moreover, the AIC procedure is not consistent. On the other hand since the criterion involves σ^2, the one-step prediction variance, the identified model may possess certain optimality properties for one-step ahead predictions. However, if the goal is multi-step ahead predictions, then it is plausible that the criterion must be modified to reflect this goal [9].

A filtered series or deterministic pattern (a signal extracted from a signal corrupted by noise) is expected to be very close to a true series and has to have the same order of AR components as a well-fitted AR model generated by the equation

$$\mathbf{z_t = \sum^p_{j=1} a_j z_{t-j} + \varepsilon_t,} \qquad \mathbf{(4)}$$

where ε_t is a zero-mean white. In order to construct a linear filter $L(x_t)$ of order k

$$\mathbf{L(x_t) = \sum^k_{j=0} c_j x_{t-j}} \qquad \mathbf{(5)}$$

representing a deterministic pattern of true series we associate $x_t = z_t + e_t$, where x_t is a series of observed data and z_t is a true series, with some certain model of AR(p) process, namely with a model where p = k. In general, the white noise disturbance term ε_t differs from the noise $e_t = x_t - z_t$, which emerged when the realisation x_t was observed from an unknown true stochastic process z_t, and includes also an unknown stochastic term (shock) δ_t: $\varepsilon_t = e_t + \delta_t$. Hence the expression (4) can be transformed to

$$\mathbf{z_t = \sum^p_{j=1} a_j z_{t-j} + x_t - z_t + \delta_t}$$

or

$$\mathbf{z_t - (1/2)\delta_t = (1/2)x_t + (1/2) \sum^p_{j=1} a_j z_{t-j}} \qquad \mathbf{(6)}$$

Definition of a filtered series y_t as a deterministic pattern of the true series in the way [10]

$$\mathbf{y_{t-s} = \begin{cases} z_{t-s} - (1/2)\,\delta_{t-s}, & s = 0 \\ z_{t-s}, & s > 0 \end{cases}} \qquad \mathbf{(7)}$$

allows to rewrite (6) in the recursive form

$$\mathbf{y_t = (1/2)x_t + (1/2) \sum^p_{j=1} a_j y_{t-j}} \qquad \mathbf{(8)}$$

Any recursive linear filter of order k

$$\mathbf{y_t = L^r(x_t) = \alpha_0 x_t + \sum^k_{l=1} \alpha_l y_{t-l}} \qquad \mathbf{(9)}$$

can be transformed [10] to the non-recursive form (5)[1] by the transformation

$$\mathbf{c_0 = \alpha_0}$$
$$\mathbf{c_l = \sum^l_{i=1} \alpha_i c_{l-i}} \qquad \mathbf{(10)}$$

for l = 1, ... , k, which set the dependency between coefficients c_l of $L(x_t)$ and α_l of $L^r(x_t)$. So in order to obtain coefficients c_l of the smoothing linear filter $L(x_t)$ (5) one need to apply the transformation (10) to the recursive filter $L^r(x_t) = y_t$, where k = p, $\alpha_0 = ½$ and $\alpha_j = (1/2)a_j$. Since by construction $L(x_t) = L^r(x_t) = y_t$ and $\delta_t = 2 (z_t - y_t)$, the estimate of the specification error ε_c can be decomposed as

$$\mathbf{\varepsilon_t^{hat} = (x_t - L(x_t)) + (X_t^{hat} - L(x_t)),} \qquad \mathbf{(11)}$$

[1] In general, a linear filter with non-recursive structure is more numerically stable.

where X_t^{hat} denotes estimate of z_t by the well-fitted AR model (4) [10].

The PRIME (*pattern residuals' interdependence measure estimator* [10]) criterion can not be viewed entirely as an information criterion. However, it incorporates a likelihood and a penalty term to take away some of the "natural growth" that absolute value of log(likelihood) has for large orders. In other words, the criterion has the structure similar to (1)-(2). The criterion is based on the assumption that a stable multi-step ahead forecast requires the well-identified plane created by points X_t, X_t^{hat}, and $L(x_t)$ in the n-dimensional space (n - sample size), where X_t^{hat} is an estimate by an AR model and $L(x_t)$ is filter constructed according to (4)-(10). In other words, vectors $(L(x_t) - X_t)$ and $(L(x_t) - X_t^{hat})$ should be orthogonal, if X_t^{hat} is the estimate by a well-fitted model [10]. The "orthogonality" assumption leads to the condition

$$\boldsymbol{\rho = corr(L(x_t) - X_t, L(x_t) - X_t^{hat}) = 0,} \quad \textbf{(12)}$$

for correlation coefficient ρ between residuals $L(x_t) - X_t^{hat}$ and $L(x_t) - X_t$ constructed after ARMA modelling and data filtering. It is worth noting that the estimates of both residuals terms in right-hand side of the decomposition (11) have to be uncorrelated under the orthogonality assumption (12).

The PRIME procedure estimates expected log likelihood of the correlation coefficient ρ for each model and selects the model which attains the minimum expected log likelihood:

$$\boldsymbol{\min_{\rho_k} El(\rho_k) = \min \{ - (n - k)\log (2\pi) - (n - k)\log(1 - \rho_k^2)/2 - (n - k) + k \},} \quad \textbf{(13)}$$

where ρ_k is the maximum likelihood estimator of correlation coefficient in sample of n-k and the additional k corrects the bias analogously to (3). The final expression for the PRIME criterion [10] is obtained by inversion of (13):

$$\boldsymbol{\max_{\rho_k} El(\rho) = \max \{(n - k) \log[4\pi^2(1 - \rho_k^2)] + 2n - 4k\}.} \quad \textbf{(14)}$$

It is clear that the maximum attains when $\rho_k = 0$.

3. COMPUTATIONAL LEARNING ALGORITHM AND EXPERIMENTAL RESULTS.

The section describes a computational learning algorithm of the PRIME selection procedure, covers the results of Monte Carlo simulation which generate an empirical distribution of the proposed estimator and provide evidence of appropriate identification of data. It illustrates also the results obtained by multi-step ahead forecast for actual data. In order to provide evidence for the strength of the PRIME criterion two methods have been utilised. Firstly, the PRIME criterion was compared with AIC and SBC by a Monte Carlo simulation with 100 replications of the Gaussian AR(3) model with 200 observations. Secondly, multi-step ahead predictions of actual data were obtained by models selected by PRIME criterion and AIC and the results were compared.

A computational algorithm of the PRIME procedure [10] involves construction of a smoothing filter by the transformation (10) applied to AR coefficients for every candidate model. When both AR model and filtered series are obtained, the algorithm calculates ρ_k and PRIME value for a given order k. Finally, it selects the model orders corresponding to local maxima of (14). One potential benefit of the PRIME procedure is that obtained "smoothed" filter estimates of the observed data do not have to be adapted for each step of prediction. The computational algorithm of the Monte Carlo simulation has the following steps:

Step 1. Generating ARMA(3,0) model: X(T) = 0.9 X(T-1) - 0.7 X(T-2) + 0.4 X(T-3) + E(T), where E(T) is white noise drawn from a standardised normal distribution.
Step 2. Centring the generated data X(T).
Step 3. Specifying the order k.
Step 4. Running the 200 observations X(T) through ARIMA procedure with the parameters:
AR order P = the given order k; MA order Q = 0.
Storing k AR coefficients, the modelled series X^{hat}(T), and the test statistics (AIC, SBC).
Step 5. Construction of a linear filter C(J), J = 1, ..., k + 1 by the transformation (10) applied to AR coefficients for the given order k.
Step 6. Filtering the observed data X(T) and its lagged values X(T-1) ... X(T-k) by the C(J).
Pattern L(T) extraction.
Step 7. Construction of the residuals D1 = L(T) - X(T) and D2 = L(T) - X^{hat}(T).
Estimation of the variances of D1 and D2 and the covariance between them.
Estimation of the correlation coefficient and calculation of PRIME value for the given order k.
Step 8. Storing the results into output file.
Step 9. If the order k does not exceed 24 (the margin for a data set of 200 observations) then return to the step 3 with k = k + 1, else select local optima for the AIC, SBC, and PRIME.

The steps 1 - 9 were repeated 100 times with the different seeds of random normal distribution for Monte Carlo simulation. The SAS Institute's statistical computer package was used as a simulation tool. The PRIME procedure selects the 3rd order by maximising (14). With growth of order k the PRIME value is steadily decreasing. The AIC demonstrates similar behaviour, although after correct selection of the 3rd order as a global minimum it indicates local minima for large k (Fig. 1). It is well known that AIC has a tendency to select longer lags than the true number of lags. SBC obviously does not have such fluctuations. That on the other hand might be the reason for slightly less predictive power for multi-step ahead forecast of SBC than AIC - in order to predict many steps ahead a model has to include low-frequency components (longer lags). Also the PRIME provides sharper results in indication of the global optimum than AIC and SBC.

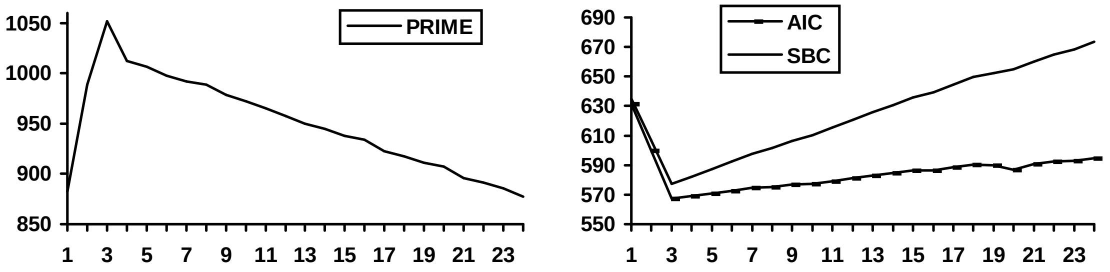

Figure 1. Typical plots of the compared criteria dependent on model order (axis X).

The multi-step ahead forecasting for actual data[1] has been obtained by two models. The basic model was selected by AIC procedure and the alternative one - by PRIME criterion. The results of the unadjusted time series (1969.01 - 1992.12) identification are summarised in the following

Table 1. Growth Rate of the U.S. Civilian Labor Force (1969.01 - 1992.12). Unadjusted Data[2].

[1] the U.S. civilian labor force for the period: January, 1969 - October, 1993 (288 observations). The source of data is the U.S. Bureau of Labor Statistics.

[2] the numbers represent the selected AR and MA orders of the identified ARMA model.

	AR components	MA components
AIC	13, 17, 19, 24	6, 18, 24
PRIME	2, 5, 10, 18, 24	6, 24

The analysis shows that the basic and the alternative models differ very much: PRIME procedure selected 5 AR lags and only one of them (24th) was indicated amongst 4 lags selected by AIC, although the sum (p + q) of AR and MA orders is the same - 7 for both the basic and the alternative models. In order to compare a predictive power of basic and alternative models multi-step ahead post-sample (1993.01 - 1993.10) forecasts have been generated for each case and an h-step mean squared error of prediction has been estimated ($1 \leq h \leq 10$). The comparison shows that the alternative model has a larger predictive power than the basic one: the h-step mean squared error of prediction was smaller uniformly for the alternative model except the 3rd step ahead.

The comparison between the PRIME criterion and well-known criteria (AIC, SBC) provides favourable evidence for the proposed criterion. The two utilised methods demonstrated that the PRIME procedure takes more into account longer lags if these are included as candidates than SBC does, although it does not have a tendency to overestimate model as AIC. Additional research is required in order to work out more rigorous approach to the orthogonality assumption. A derivation of a special procedure for a proper identification of MA components seems to be necessary as well. Nevertheless, the presented results show that the PRIME criterion can be used as an appropriate tool for learning (selection of) an AR model order.

REFERENCES:

1. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle, in B.N. Petrov and F. Csaki, eds., *2nd International Symposium on Information Theory*, Akademia Kiado, Budapest, 1973, pp. 267-281.
2. Akaike, H. A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, vol. AC-19, No. 6, 1974, pp. 716-723.
3. Atkinson, A.C. Likelihood Ratios, Posterior Odds and Information Criteria, *Journal of Econometrics* 16, 1981, pp. 15-20.
4. Box, G.E.P. and Jenkins, G.M. *Time Series Analysis: Forecasting and Control*, revised edn. San Francisco: Holden Day, 1976.
5. Hannan, E.J. and Quinn, B.G. The determination of the order of an autoregression. *J. Roy. Statist. Soc.* Ser. B 41, 1979, pp. 190-195.
6. Hanson, S.J., Petsche, T., Kearns, M., and Rivest, R.L. *Computational Learning Theory And Natural Learning Systems.* Volume II. The MIT Press, 1994.
7. Michalski, R.S. Learning and Cognition. Proceedings of the Second World Conference on the Fundamentals of Artificial Intelligence, Paris, 3-7 July, 1995, pp. 507 - 510.
8. Parzen, E. Some Recent Advances in Time Series Modeling. *IEEE Transactions on Automatic Control*, vol. AC-19, No. 6, 1974, pp. 723-730.
9. Pourahmadi, M. Fundamental Roles of the Idea of Regression and Wold Decomposition in Time Series, in *New Direction in Time Series Analysis*, vol. 46, 1990, pp. 287-314.
10. Prokopenko, M.I. *An Autoregressive Model Order Selection Based on Estimation of Pattern Residuals Interdependence*, M.A. thesis, University of Missouri - Columbia, U.M.I., Ann Arbor, 1994.
11. Shibata, R. Various Model Selection Techniques in Time Series Analysis, in E.J. Hannan, P.R. Krishnaiah and M.M. Rao (eds.*), Time Series in the Time Domain. Handbook of Statistics*, vol. 5, 1985, pp. 179-187.
12. Schwarz, C. Estimating the dimension of a model. *Ann. Statist.* 6, 1978, pp. 461-464.
13. Utgoff, P.E. *Machine Learning of Inductive Bias.* Kluwer Academic Publishers, 1986.
14. Wallace, C.S. *Machine Learning. Lecture presented at the Second Australian Theory Day*, UNSW, Sydney, 17.02.1995.
15. Wold, H. *A Study in the Analysis of Stationary Time Series*, 2nd ed. Uppsala: Almquist and Wicksell, 1954.