Statistica Sinica 3(1993), 255-259

SELECTION OF A LINEAR INTERPOLATOR FOR TIME SERIES

Francesco Battaglia

University "La Sapienza"

Abstract: A criterion is proposed for selecting the order of the linear interpolator of a stationary time series, which may be useful in the problems of missing values and outlier detection. The criterion is based on a mean-square error similar to that leading to the final prediction error criterion for autoregressive model identification, and is called the *final interpolation error* criterion. The behavior of the proposed criterion is illustrated by means of a simulation study.

Key words and phrases: Inverse correlations, linear interpolator, missing values, order selection criteria, outliers, time series.

1. Introduction

Let $\{x_t\}$ denote a second-order stationary zero-mean stochastic process with autocovariance function $R(h)$, inverse covariance function $Ri(h)$ and inverse correlation function $ri(h) = Ri(h)/Ri(0)$ (Cleveland (1972)). The linear interpolator, I_t, of x_t is a linear combination of the observations x_{t-j}, $j \neq 0$ such that the mean square error $E\{x_t - I_t\}^2$ is a minimum; it is well known (see e.g. Grenander and Rosenblatt (1957)) that the weights of such linear combination are equal to $-ri(j)$:

$$I_t = -\sum_{j \neq 0} ri(j) x_{t-j}. \tag{1}$$

Linear interpolators are important for their ability to reconstruct one observation from the remaining data, and arise naturally in outlier detection (e.g. Fox (1972), Peña and Maravall (1991), Chang, Tiao and Chen (1988), Tsay (1988)) and missing data (e.g. Brubacher and Wilson (1976)). In fact, if x_t is missing, its least squares estimator is I_t, and, if x_q, say, is an additive outlier, i.e. it is additively perturbed by a shock ω, the least squares estimator of ω is $x_q - I_q$.

There are two different ways of estimating a linear interpolator: an ARMA model may be fitted to the data and then the interpolator is expressed as a function of the estimated model parameters, or alternatively the coefficients of I_t may be estimated directly by least squares. The latter case seems preferable because the observed mean square interpolation error is minimized; however, for

practical application it is necessary to limit the infinite sum in (1) to a finite number p of coefficients.

The paper concentrates on the choice of the order p, and a procedure is proposed for choosing p according to a mean square error optimality criterion, similar to that leading to the final prediction error for autoregressive model fitting (Akaike (1969)). We apply Akaike's idea for computing the final interpolation error, i.e., the variance of the interpolation error when the interpolator coefficients are estimated using an independent realization. The analogy between autoregressive models and finite linear interpolators is apparent since, in the first case, the observations are regressed on p past values, while, in the second case, they are regressed on both past and future values. The reason for using a different order selection criterion is the different behavior of the coefficients of autoregressive models and linear interpolators. When the series follows exactly a purely autoregressive process of order k, it may be easily shown that $ri(h) = 0$ for $h > k$, therefore $p = k$ is also the correct order for the linear interpolator, but for other processes the inverse correlations do not cut off, so that the choice of p involves an approximation, and there is no evidence that the best approximating orders should coincide for the two purposes. Specifically, the improvement achieved by increasing the order, when fitting an autoregressive model, if measured in terms of residual variance reduction, is determined by the partial correlation, whereas the same measure for linear interpolator fitting is determined by the inverse correlation. It has been pointed out (Cleveland (1972), Abraham and Ledolter (1984)) that these two functions behave differently, though both have the same cut-off lag for purely autoregressive processes, and, in particular, their decay rates may be considerably different.

The present framework is suitable only for stationary time series, and in non-stationary cases the interpolator can be derived only for the differenced series. However, inverse correlations may be defined for non-stationary series also, using the concept of pseudo-spectral density (Hillmer and Tiao (1982)); sample properties are not known in this case and are currently being investigated.

We finally note that some problems arise in handling the data at the beginning and at the end of the series, since computing interpolators at time t requires sufficient observations before and after t. Forecasts or backforecasts based on unilateral models may be useful in such cases.

2. Derivation of the FIE Criterion and Applications

Suppose that two independent realizations $\{x_t\}$ and $\{y_t\}$ of the same gaussian process with $ri(h) = 0$ for $h > p$ are available. We compute the least squares estimates $\{\hat{ri}(u), u \leq p\}$ on the series $\{y_t\}$ and use them for estimating a linear

finite interpolator for x_t:

$$\hat{I}_t = -\sum_{u=1}^{p} \hat{r}i(u)(x_{t-u} + x_{t+u}).$$

The mean square interpolation error $E(x_t - \hat{I}_t)^2$, unconditional on $\{y_t\}$, may be derived following the method developed by Akaike (1969) for prediction errors. The following asymptotic expression is obtained:

$$\text{FIE}(p) = \frac{1}{Ri(0)}\left\{1 + \frac{2}{N}\text{tr}[\{Ri(0)S_p\}^{-1}]\right\} \tag{2}$$

where N is the series length and S_p the matrix with elements $s_{i,j} = R(i-j) + R(i+j)$ $(i,j = 1,2,\ldots,p)$. We call (2) the *final interpolation error*; its value is jointly influenced by the amount of variability unexplained after interpolation and the sample variability related to coefficients estimation. The structure is similar to that of the familiar FPE: two additive terms are present, one accounting for the interpolation error variance, $1/Ri(0)$, and the other one, $\frac{2}{N}\text{tr}[\{Ri(0)S_p\}^{-1}]$, decreasing with N and increasing with p.

In order to compute FIE on an actual series, ordinary autocovariance estimates may be used for the matrix S_p, while estimates of $Ri(0)$ have been proposed by Bhansali (1980) and Battaglia (1988).

Expression (2) is derived under the hypothesis that $ri(h) = 0$ for $h > p$; in other words, that the correct order is not larger than p. If p is less than the true order, the observed interpolation error $1/\hat{R}i(0)$ will generally decrease on increasing p. On the other hand, FIE(p) is increasing as p increases over the correct order. Thus, a reasonable procedure for selecting the order of the interpolator is to compute FIE(p) for a range of possible values of p, and select the one corresponding to the minimum estimated FIE.

In order to gain some insight into the behavior of the proposed criterion, a simulation study is presented. Sets of hundred series of 50 observations were simulated according to some ARMA gaussian models; on each series, the FPE criterion for autoregressive order selection, and the FIE for interpolation order choice were computed for orders ranging from 0 to 9. Results are shown in Table 1.

For a first order autoregressive process with parameter $\phi = 0.5$ the behavior of the two criteria is broadly consistent. Similar results are found for a second order autoregressive process with parameters $\phi_1 = 1.2$ and $\phi_2 = -0.6$, and are omitted here to save space. The results for a first order moving average process with parameter $\theta = 0.35$ show larger differences between the two criteria: FIE generally tends to select smaller orders than FPE; in particular, very large orders

are less frequently selected. We note that for MA(1) models the ratio of inverse to partial correlation equals $-(1-\theta^2)(1-\theta^{2h+2})$, therefore the inverse correlations are about 15% smaller than the partial correlations for the present series. Finally, a second order moving average process with parameters $\theta_1 = 0.1, \theta_2 = -0.8$ was simulated. In this case the orders selected by means of FIE were generally larger than those chosen by FPE. Here the inverse correlations exhibit a pseudo-periodical behavior with period about four, and their absolute values are large (-0.79 at lag 2) and very slowly decaying: $ri(8)$ is about 0.37 and $ri(10)$ is -0.27. The behavior of the partial correlations is similar but with rather smaller values (0.48 at lag 2, -0.14 at lag 8, 0.10 at lag 10).

In order to check to what extent the different behaviour of the order selection criteria is relevant for interpolation accuracy, a figure of merit is evaluated as follows. For each simulated series, and according to the order selected by each criterion, the finite linear interpolator is estimated by least squares, and the observed interpolation error variance (average of the squared interpolation error for t ranging from $p+1$ to $N-p$) is computed. Table 2 reports the averages of such quantities based on the hundred replications. Figures for FIE are always less than for FPE, differences are smaller for purely autoregressive processes than for moving average, and range from about 5% to 10%.

Summarizing, the simulation results suggest that for purely autoregressive processes both criteria generally behave in a similar way, while for other processes the orders selected by the two procedures may differ considerably, depending on the different behavior of the partial and inverse correlations. In both cases, use of FIE provides a reduction in the overall observed mean square error of interpolation.

Table 1. Frequencies of selected orders in 100 series of length 50 simulated from: (i) an AR(1) model with $\phi = 0.5$; (ii) a MA(1) model with $\theta = 0.35$, and (iii) a MA(2) model with $\theta_1 = 0.1$ and $\theta_2 = -0.8$.

	AR(1)		MA(1)		MA(2)	
order	FPE	FIE	FPE	FIE	FPE	FIE
0	0	5	0	14	0	0
1	73	76	54	52	0	0
2	10	9	23	22	23	7
3	10	9	8	8	3	2
4	7	1	8	4	35	29
5			3	0	16	10
6			2	0	11	23
7					3	7
8					8	15
9					1	7

Table 2. Observed interpolation error variance according to the chosen order selection criterion. Figures are averages on 100 series of length 50 from: (i) an AR(1) model with $\phi = 0.5$; (ii) a MA(1) model with $\theta = 0.35$; (iii) an AR(2) model with $\phi_1 = 1.2$ and $\phi_2 = -0.6$; (iv) a MA(2) model with $\theta_1 = 0.1$ and $\theta_2 = -0.8$.

model	FPE	FIE
AR(1)	0.79	0.75
MA(1)	0.82	0.79
AR(2)	0.36	0.34
MA(2)	0.60	0.54

References

Abraham, B. and Ledolter, J. (1984). A note on inverse autocorrelations. *Biometrika* **71**, 609–614.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21**, 243–247.

Battaglia, F. (1988). On the estimation of the inverse correlation function. *J. Time Ser. Anal.* **9**, 1–10.

Bhansali, R. J. (1980). Autoregressive and window estimates of the inverse correlation function. *Biometrika* **67**, 551–566.

Brubacher, S. R. and Wilson, G. T. (1976). Interpolating time series with application to the estimation of holiday effects on electricity demand. *Appl. Statist.* **25**, 107–116.

Chang, I., Tiao, G. C. and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics* **30**, 193–204.

Cleveland, W. S. (1972). The inverse autocorrelations of a time series and their applications. *Technometrics* **14**, 277–293.

Fox, A. J. (1972). Outliers in time series. *J. Roy. Statist. Soc. Ser.B* **34**, 350–363.

Grenander, U. and Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series*. John Wiley, New York.

Hillmer, S. C. and Tiao, G. C. (1982). An ARIMA-model-based approach to seasonal adjustment. *J. Amer. Statist. Assoc.* **77**, 63–70.

Peña, D. and Maravall, A. (1991). Interpolation, outliers and inverse autocorrelations. *Comm. Statist. Theory Methods* **20**, 3175–3186.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *J. Forecast.* **7**, 1–20.

Department of Sociology, University "La Sapienza", via Salaria, 113, 00198 Rome, Italy.

(Received July 1989; accepted May 1992)