Working Paper 91-08
February 1990

Departamento de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Madrid)

INTERPOLATION, OUTLIERS AND INVERSE AUTOCORRELATIONS

Daniel Peña* and Agustín Maravall**

Abstract

The paper addresses the problem of estimating missing observations in linear, possibly nonstationary, stochastic processes when the model is known. The general case of any possible distribution of missing observations in the time series is considered, and analytical expressions for the optimal estimators and their associated mean squared errors are obtained. These expressions involve solely the elements of the inverse or dual autocorrelation function of the series.

This optimal estimator -the conditional expectation of the missing observations given the available ones- is equal to the estimator that results from filling the missing values in the series with arbitrary numbers, treating these numbers as additive outliers, and removing the outlier effects from the invented numbers using intervention analysis.

Key words
Missing observations; Outliers, Intervention analysis; ARIMA models; Inverse autocorrelation function.

* Departamento de Economía, Universidad Carlos III de Madrid and Laboratorio de Estadística, ETSII, Universidad Politécnica de Madrid.

** European University Institute, Firenze, Italy.

1. INTRODUCTION

In this paper we consider the problem of estimating missing observations in linear, possibly nonstationary, stochastic processes, which we parametrize as parsimonious Autoregressive Integrated Moving Average (ARIMA) models. The model is assumed to be known, and hence we concern ourselves with obtaining analytical expressions for the conditional expectation of the missing observations given the available ones. In terms of the well-known EM algorithm, our aim is to obtain further insights into the E -the expectation- step.

In Section 2 we present the model, the basic assumptions and some background references. Section 3 contains the general result of the paper for the case of any possible distribution of missing observations in the series. Analytical expressions for the optimal estimator of the missing observations and the associated Mean Squared Error (MSE) are obtained, and the relationship with additive outlier models is discussed. Section 4 presents some conclusions.

2. THE MODEL AND SOME BACKGROUND RESULTS

Let the series z_t follow the general ARIMA model

$$\phi(B)\ z_t = \theta(B)\, a_t, \qquad (2.1)$$

where $\phi(B)$ and $\theta(B)$ are finite polynomials in the lag operator B, and a_t is a Gaussian white-noise process with variance σ^2_a. The autoregressive (AR) polynomial $\phi(B)$ contains the stationary as well as the nonstationary roots; the moving average (MA) polynomial $\theta(B)$ is assumed to be invertible and hence, if $\pi(B)$ $(\equiv 1-\pi_1 B -\pi_2 B^2 -...)$ denotes the convergent polynomial $\phi(B)\ \theta(B)^{-1}$, model (2.1) can be expressed as the pure autoregression

$$\pi(B)\ z_t = a_t.$$

Define the inverse or dual model of (2.1) as the one that results from interchanging the AR and MA polynomials; that is

$$\theta(B)\ z_t^D = \phi(B)\, a_t, \tag{2.2}$$

or, equivalently,

$$z_t^D = \pi(B)\ a_t$$

Since (2.1) is invertible, (2.2) will be stationary; its variance (v_D) and autocovariance generating function ($\Gamma^D(B)$) are given by

$$v_D = \sigma_a^2 \sum_{i=0}^{\infty} \pi_i^2 \qquad (\pi_0 = 1) \tag{2.3}$$

$$\Gamma^D(B) = v_D + \sum_{i=1}^{\infty} \Gamma_i^D\ (B^i + F^i) = \sigma_a^2\, \pi(B)\, \pi(F), \tag{2.4}$$

where $F = B^{-1}$ is the forward operator. Then, the dual autocovariance at lag i will be given by

$$\Gamma_i^D = \sigma_a^2 \sum_{i=0}^{\infty} \pi_j \pi_{j+i}. \tag{2.5}$$

Following Cleveland (1972), the function

$$\rho^D(B) = \sigma_a^2\, \pi(B)\ \pi(F)/v_D \tag{2.6}$$

will be denoted the Inverse or Dual Autocorrelation Generating Function (DAGF). (Trivially, from the ARIMA expression of the model, the DAGF is immediately available.)

Assume the series z_t has a single missing value for t = T. The minimum MSE estimator of z_T (the conditional expectation of z_T given the available observations) is a linear combination of the observed values, where the weights depend on the covariance structure of the process. Several authors have shown how to compute the estimator recursively using the Kalman filter (Jones, 1980; Harvey and Pierse, 1984; Kohn and Ansley, 1983). Wincek and Reinsel (1986), and Ljung (1982, 1989) have concentrated on the explicit form of the likelihood function and the maximum likelihood estimates of the missing values. The analytical expression for the conditional expectation of a missing value in a stationary stochastic process has been available for some time (Grenander and Rosenblatt, 1957); it can be expressed as

$$\hat{z}_T = -\sum_{i=1}^{\infty} \rho_i^D (z_{T-i} + z_{T+i}) \qquad (2.7)$$

where ρ_i^D is the lag-i dual autocorrelation of z_t. Moreover, the MSE of the estimator is found to be

$$MSE(\hat{z}_T) = \sigma_a^4 / v_D^{-1} = \sigma_a^2 / \sum_{j=0}^{\infty} \pi_i^2 ,$$

where v_D is given by (2.3).

Brubacher and Wilson (1976) obtained (2.7) by least squares for a seasonal nonstationary ARMA process and showed the relation between interpolation and the inverse autocorrelation function. Kato (1984) showed that the inverse autocorrelation function at lag K is equal to the negative of the partial correlation between z_t and z_{t+k} after elimination of the influence of z_h, (h≠t, t+k). Battaglia and Bhansali (1987) used the interpolation error variance to build an index of linear determinism.

Pourahmadi (1989) has studied the estimation and interpolation of several missing values of a stationary time series. Finally, Bruce and Martin (1989) and Peña (1987, 1990) have shown the relationship between interpolation and additive outlier estimation.

Focussing on this last extension, if the series has an additive outlier at period T, so that the observation for that period is (z_T+w) where w is the (unknown) outlier effect, then the optimal estimator of w is given by (see Chang, Tiao and Chen, 1988)

$$\hat{w} = \rho^D(B)\ Z_T, \qquad (2.8)$$

where Z denotes the observed series such that $Z_T = z_T+w$, and $Z_t = z_t$ for $t \neq T$. Thus, back to the missing observation case, if the "hole" in the series is filled with an arbitrary number Z_T and this "invented" observation is treated as an outlier, the missing observation estimator can be obtained through

$$\hat{z}_T = Z_T - \hat{w}, \qquad (2.9)$$

and it is easily seen that (2.8) and (2.9) yield expression (2.7). Furthermore, since $z_T-\hat{z}_T = \hat{w}-w$, it follows that MSE($\hat{w}$) is also equal to σ_a^4/v_D.

Therefore, when the model is known, optimal estimation of an additive outlier is equivalent to the following procedure: First, assume the outlier is a missing observation and obtain its optimal estimate given the rest of the observations. Then, compute the outlier effect as the difference between the outlier and the interpolated value. Alternatively, estimation of a missing observation can be seen as the result of the following procedure: First, fill the "hole" in the series with an arbitrary number. Then treat this number as an additive outlier, and remove the outlier effect with intervention analysis.

3. THE GENERAL CASE

Assume that, in all generality, the finite series z_t has k missing values at periods T_1, T_2,..., T_k, were $T_i < T_j$ if i < j. We can always fill the k holes in the series with arbitrary numbers, Z_{T_i}, and construct an "observed" series Z_t by

$$\begin{aligned} Z_t &= z_t + w_t \qquad t = T_1, \ldots, T_K \\ Z_t &= z_t \qquad \textit{otherwise.} \end{aligned}$$

where w_t is an unknown parameter. In matrix notation, we can write

$$Z = z + Hw \tag{3.1}$$

where Z and z are the series Z_t and z_t expressed as vectors, H is a matrix with k columns such that $H_{T_j,j} = 1$, for j = 1,..., k, and $H_{ij} = 0$, otherwise, and w is a k-dimensional vector with elements $w_j (\equiv Z_{T_j} - z_{T_j})$. Let Σ be the covariance matrix of the series z_t; then the generalized least squares estimator of w is given by

$$\hat{w} = (H' \Sigma^{-1} H)^{-1} H' \Sigma^{-1} Z. \tag{3.2}$$

Treating the starting values as fixed constants, (Ansley, 1979), the Cholesky factorization of Σ leads to

$$\Sigma^{-1} = \pi' \pi \, \sigma_a^{-2},$$

where π is a lower triangular matrix with the $-\pi_j$'s on the various lower diagonals. Then we can write (3.2) as

$$\hat{w} = (X'X)^{-1} X'Y, \tag{3.3}$$

where $X = \pi H$, $Y = \pi Z$.

Expression (3.3) can alternatively be obtained as follows: Let d_t^j, j = 1,..., k, be a set of dummy variables such that $d_t^j = 1$ for $t = T_j$ and zero otherwise. Then, from (2.2), the following intervention model (Box and Tiao, 1975) is obtained

$$\pi(B)\ (Z_t - \sum_j w_j d_t^j) = a_t, \qquad (3.4)$$

which can be rewritten as

$$y_t = \sum_j w_j\, x_{jt} + a_t\,, \qquad (3.5)$$

where $y_t = \pi(B)Z_t$, $x_{jt} = \pi(B)d_t{}^j$. Expression (3.5) is a regression equation with x_{jt} deterministic and a_t white-noise; it is immediately seen that the minimum MSE estimator of w is then given by (3.3).

Expressions (3.2) and (3.3) apply to a finite series. Now, let us consider the case of a series of infinite length. The results obtained in this case will be approximately exact when the sample size is large and the missing values are not near the end of the series. Using (2.3), (2.5) and (2.6), straightforward computation yields

$$\Sigma\ y_t x_{jt} = \pi(B)\ \pi(F)\, Z_{T_j} = (v_D/\sigma_a^2)\,\rho(B)\, Z_{T_j} \qquad (3.6)$$

$$\Sigma\ x_{tj}^2 = (v_D/\sigma_a^2) = \sum_{i=0}^{\infty} \pi_i^2 \qquad (3.7)$$

$$\Sigma\ x_{jt}\, x_{j+h,\,t} = -\pi_h + \sum_{i=1}^{\infty} \pi_i \pi_{i+h} = \Gamma_h^D/\sigma_a^2, \qquad (3.8)$$

where $\Gamma_h{}^D$ denotes the lag-h autocovariance of the dual process defined in (2.5). Using (3.7) and (3.8), X'X is a (kxk) symmetric matrix with the (i,j)-th element given by $\Gamma^D{}_{T_i - T_j}$. The elements are thus autocovariances of the dual process for lags reflecting the relative distances between the missing observations. Therefore,

$$X'X = (v_D/\sigma_a^2)\, R_D \qquad (3.9)$$

where R_D is the (kxk) symmetric matrix

$$R_D = \begin{vmatrix} 1 & \rho^D_{T_2-T_1} & \rho^D_{T_3-T_1} & \cdots & \rho^D_{T_k-T_1} \\ & 1 & \rho^D_{T_3-T_2} & \cdots & \rho^D_{T_K-T_2} \\ & & 1 & & \cdot \\ & & & \cdot & \cdot \\ & & & \cdot & \cdot \\ & & & & 1 \quad \rho^D_{T_k-T_{k-1}} \\ & & & & 1 \end{vmatrix} \qquad (3.10)$$

with elements the dual autocorrelations of the process.

If Z^k denotes the vector of invented observations $Z_{T_1},\ldots, Z_{T_k})'$, from (3.3), (3.6) and (3.9), the estimator of w becomes

$$\hat{w} = R_D^{-1}\rho_D(B)\ Z^k, \qquad (3.11)$$

and the vector of missing observations estimators $\hat{z}^k = (\hat{z}_{T_1}, \ldots, \hat{z}_{T_k})'$ can then be obtained through

$$\hat{z}^k = Z^k - \hat{w}. \qquad (3.12)$$

Equations (3.11) and (3.12) are the generalization of equations (2.8) and (2.9), derived for the single missing observation, to the case of any possible sequence of missing observations. The estimator $\hat{z}^k$ can be seen as the result of the following procedure: First, fill the holes in the series with arbitrary numbers, which then are treated as additive outliers. Removing from the arbitrary numbers the outlier effects, the missing observation estimator is obtained. It is straightforward to show that the estimator $\hat{z}^k$ does not depend on the vector Z^k of arbitrary numbers.

As for the MSE of the estimator, from (3.11) and noticing that $z_{T_j} - \hat{z}_{T_j} = \hat{w}_j - w_j$, it follows that

$$MSE(\hat{z}^k) = MSE(\hat{w}) = (\sigma_a^4/v_D)\ R_D^{-1}.$$

Therefore, both the estimator and its MSE can be expressed in terms of the dual autocorelations of the series.

It is immediately seen that, for the scalar case, the matrix R_D of (3.10) becomes simply 1 and equation (2.8) is obtained. More relevantly, for the case of k consecutive missing observations, R_D becomes the symmetric Toeplitz matrix

$$R_D = \begin{vmatrix} 1 & \rho_1^D & \rho_2^D & \cdots & \rho_{k-1}^D \\ & 1 & \rho_1^D & \cdots & \rho_{k-2}^D \\ & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot \\ & & & & \rho_1^D \\ & & & & 1 \end{vmatrix}$$

that is, the (kxk) dual autocorrelation matrix. The MSE of the estimator is then the inverse of the (kxk) autocovariance matrix of the dual process, also a symmetric Toeplitz matrix with all the elements of the i-th diagonal equal to Γ_i^D.

Finally, from equation (3.11) another interesting expression for $\hat{z}^k$ is obtained. Let $w_j^{(1)}$ denote the estimator of w_j obtained by assuming that, in the vector Z^k, only the element Z_{T_j} is invented, and using the method of Section 2 for the scalar case. Define the vector $w^{(1)} = (w_1^{(1)}, \ldots, w_k^{(1)})'$; from (2.8), expression (3.11) can be rewritten

$$\hat{w} = R_D^{-1}\ w^{(1)}.$$

Therefore, for the case of a vector of missing observations, the optimal estimator can be seen as a weighted average of the estimators obtained by treating each missing observation as if it were the only missing one (using (2.8) on the arbitrarily filled series). The

weights are the elements of the inverse dual autocorrelation matrix.

4. CONCLUSION

We have considered the general case of any distribution of missing observations in an infinite realization a possibly nonstationary linear stochastic process. Compact analytical expressions for the optimal estimator of the missing observations are derived, which explicitly show the dependence of the conditional expectation on the stochastic structure of the series, and involve only the elements of its inverse or dual autocorrelation function. Furthermore, analytical expressions for the Mean Squared Error matrix are immediately obtained, which are simply functions of the dual autocovariance matrix.

The expression for the conditional expectation of the missing observations given the available ones is the same that results from filling the missing values with arbitrary numbers, treating these numbers as additive outliers, and removing the outlier effects through intervention analysis from the invented observations.

BIBLIOGRAPHY

Ansley, C.F. (1979). 'An algorithm for the exact likelihood of a mixed autoregressive-moving average process,' *Biometrika*, 66, 1, 59-65.

Battaglia, F. and Bhansali, R.J. (1987). 'Estimation of the interpolation error variance and an index of linear determinism,' *Biometrika*, 74, 771-779.

Box, G.E.P. and Tiao, G.C. (1975). 'Intervention Analysis with applications to economic and environmental problems,' *Journal of the American Statistical Association*, 70, 70-79.

Brubacher, S.R. and Tunnicliffe-Wilson, G. (1976). 'Interpolating time series with application to the estimation of holiday effects on electricity demand,' *Applied Statistics*, 25, 2, 107-116.

Bruce, A.G. and Martin, R.D. (1989). 'Leave-k-out diagnostics for time series,' *Journal of the Royal Statistical Society B*, 51, 3, 363-424.

Chang, I., Tiao, G.C. and Chen Ch. (1988). 'Estimation of time series parameters in the presence of outliers,' *Technometrics*, 30, 2, 193-204.

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

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

Harvey, A.C. and Pierse, R.G. (1984). 'Estimating missing observations in economic time series,' *Journal of the American Statistical Association*, 79, 125-132.

Jones, R.H. (1980). 'Maximum likelihood fitting of ARMA models to time series with missing observations,' *Technometrics*, 22, 3, 389-395.

Kato, A.J. (1984). 'A characterization of the inverse autocorrelation function,' *Communications in Statistics, Theory and Methods*, 13, 20, 2503-2510.

Khon, R. and Ansley, C.F. (1983). 'Fixed interval estimation in state space models when some of the data are missing or aggregated,' *Biometrika*, 70,3, 683-8.

Ljung, G.M. (1982). 'The likelihood function for a stationary Gaussian autoregressive-moving average process with missing observations,' *Biometrika*, 69, 1, 265-8.

Ljung, G.M. (1989). 'A note on the estimation of missing values in time series,' *Communication in Statistics, Simulation and Computation*, 18, 2, 459-465.

Peña, D. (1987). 'Measuring the importance of outliers in ARIMA models,' in *New Perspectives in Theoretical and Applied Statistics*, eds. M.L. Puri et al, Wiley, 109 - 118.

Peña, D. (1990). 'Influential observations in time series,' *Journal of Business and Economic Statistics, 8*, 2, 235-241.

Pourahmadi, M. (1989). 'Estimation and interpolation of missing values of a stationary time series,' *Journal of Time Series Analysis*, 10, 2, 149-169.

Wincek, M.A. and Reinsel, G.C. (1986). 'An exact maximum likelihood estimation procedure for regression-ARMA time series models with possibly nonconsecutive data,' *Journal of the Royal Statistical Society*, 48, 3, 303-313.

ACKNOWLEDGEMENTS

Thanks are due to Professor Gregory Reinsel and to two anonymous referees for helpful comments. The first author acknowledges support from DGICYT, project PB86-0538.