Biometrika (1999), **86**, 4, *pp.* 965–972

Printed in Great Britain

Statistical analysis of incomplete long-range dependent data

BY WILFREDO PALMA AND GUIDO DEL PINO

Departamento de Estadística, Pontificia Universidad Católica de Chile, Casilla 306, Santiago 22, Chile

wilfredo@mat.puc.cl gdelpino@mat.puc.cl

SUMMARY

This paper addresses both theoretical and methodological issues related to the prediction of long-memory models with incomplete data. Estimates and forecasts are calculated by means of state space models and the influence of data gaps on the performance of short and long run predictions is investigated. These techniques are illustrated with a statistical analysis of the minimum water levels of the Nile river, a time series exhibiting strong dependency.

Some key words: ARFIMA model; Incomplete data; Linear predictor; Long-memory; Maximum likelihood; Mean square prediction error; State space system.

1. INTRODUCTION

Long-range dependent data arise in a wide variety of scientific disciplines, from hydrology to economics; see for example Bloomfield (1992), Robinson (1993), Beran (1994) and Ray & Tsay (1997). A well-known class of long-memory processes are the autoregressive fractionally integrated moving average (ARFIMA) models, defined by the discrete-time equation

$$\Phi(B)(1-B)^d y_t = \Theta(B)\varepsilon_t,$$

for $t = 1, \ldots, n$, where $|d| < \frac{1}{2}$, $\{\varepsilon_t\}$ is a white noise sequence with zero mean and variance σ_ε^2, B is the backshift operator $By_t = y_{t-1}$, $\Phi(B)$ and $\Theta(B)$ are polynomials of degrees p and q respectively with no common zeros and all their roots outside the unit circle, and $(1-B)^d$ is the fractional difference operator. The ARFIMA models have long memory because their autocorrelations decay to zero at a hyperbolic rate, that is $\rho_k \sim |k|^{-\alpha}$ $(\alpha > 0)$, for large k. Estimation of these long-range dependent models is discussed by Fox & Taqqu (1986), Dahlhaus (1989) and Sowell (1992), among others. The problem of data gaps in time series has received a great deal of attention; see for example Jones (1980), Ansley & Kohn (1983) and Penzer & Shea (1997). In particular, Palma & Chan (1997) and Chan & Palma (1998) develop state space methods for dealing with missing observations in the long-memory context.

The main objective of this paper is to investigate the effects of missing values or data irregularities on the behaviour of prediction errors. If y_t denotes the value of the series at time t, the complete and the observed series are $(y_t, t \in I_n = \{1, \ldots, n\})$ and $(y_t, t \in K_n = \{k_1, \ldots, k_{r_n}\} \subseteq I_n)$ respectively. Information may be available in the pattern of missing observations, which is equivalent to the K_n, but this set is normally considered as fixed, or inference is performed conditional on it. For likelihood inference this may be justified by appealing to the 'missing at random' condition of Little & Rubin (1987, p. 10), which means here that K_n is not affected by the parameter θ specifying the time series model or by the unobserved values $\{y_t : t \in I_n, t \notin K_n\}$. The likelihood function $L(\theta)$ is just the joint density of the observed values, which may be specified in many ways, leading to different formulae for the likelihood, e.g. integrated likelihood or recursive likelihood. Nevertheless, they all lead to equivalent functions. This paper focuses on the recursive likelihood for a Gaussian time series, which can be calculated by means of the Kalman filter. Let $\hat{y}_t$ be the best linear predictor

of y_t given the observed values up to time $t-1$ and let Δ_t be the variance of the one-step prediction error $y_t - \hat{y}_t$. With this notation the recursive likelihood is

$$L(\theta) \equiv (2\pi)^{-r_n/2} \left(\prod_{t \in K_n} \Delta_t \right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2} \left\{ \sum_{t \in K_n} \frac{(y_t - \hat{y}_t)^2}{\Delta_t} \right\} \right]. \tag{1}$$

Explicit state space formulae for calculating the predictions $\hat{y}_t$ and Δ_t may be found in Palma & Chan (1997).

Section 2 addresses the behaviour of short and long run forecasts of ARFIMA models in the presence of data gaps. Applications of these procedures to the analysis of the annual minimum water levels of the Nile river are discussed in § 3.

2. INFLUENCE OF MISSING VALUES ON PREDICTION

2·1. *Preliminaries*

This section studies the evolution of the one-step mean square prediction error, $E(y_t - \hat{y}_t)^2$, for ARFIMA models, during and after a block of missing data. The results are then specialised to the case of a single missing observation.

For mathematical convenience, the analysis is carried out by taking into account the full past, $y_{t-1}, y_{t-2}, \ldots$, instead of the finite past, $y_{t-1}, y_{t-2}, \ldots, y_1$, of the time series. Throughout this section we use the concepts of exponential and hyperbolic rates of convergence of a sequence $\{x_k\}$ to its limit x as $k \to \infty$. Exponential convergence corresponds to $|x_k - x| \leqslant C_1 a^{-k}$, for large k, for some $a > 1$ and a positive constant C_1, and hyperbolic convergence to $|x_k - x| \leqslant C_2 k^{-\alpha}$, for some $C_2 > 0$ and $\alpha > 0$.

2·2. *Influence of a block of missing values*

When m consecutive observations, $y_{t_0}, \ldots, y_{t_0+m-1}$, are missing, the standard deviation of the prediction error increases during the data gap and then decreases, as new information is added. By stationarity, we may take $t_0 = 0$.

The following two propositions characterise this behaviour during and after the data gap and specify the convergence rates. Proofs are given in the Appendix.

THEOREM 1. *Let y_t be a stationary invertible process with* AR(∞) *decomposition* $y_t = \varepsilon_t - \sum_{j=1}^{\infty} \pi_j y_{t-j}$ *and* MA(∞) *decomposition* $y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$. *Suppose that the observations* $y_0, \ldots, y_{m-1}$ *are missing and let* $\mathcal{M}_{km} = (y_t, t < k, t \notin \{0, 1, \ldots, m-1\})$. *Denote by* $e(t, m, k)$ *the error of the best linear predictor of* y_t *given* $\mathcal{M}_{km}$, *that is given all the available information before time* k, *and by* $\sigma^2(t, m, k)$ *its variance. Then*

(a) *for* $k = 0, \ldots, m$, $\sigma^2(k, m, k) = \sigma_\varepsilon^2 \sum_{j=0}^{k} \psi_j^2$;

(b) *for* $k > m$, $\sigma^2(k, m, k) - \sigma_\varepsilon^2 \leqslant \sigma_y^2 m^2 \max_{j \geqslant k-m} \pi_j^2$.

Theorem 1(a) shows that, as expected, during the data gap the mean square prediction error increases monotonically up to time m, since no new information is being added. In contrast, after time m the variance of the prediction error decreases, as new observations are incorporated. The next result specifies the rates at which these error variances increase or decrease during and after the gap.

THEOREM 2. *With the notation of Theorem* 1, *for a long-memory* ARFIMA *model,*

(a) $\sigma^2(k, m, k) - \sigma_y^2 \sim C_1 k^{2d-1}$, *for some constant* $C_1 > 0$ *and large* k ($k \leqslant m$);

(b) $\sigma^2(k, m, k) - \sigma_\varepsilon^2 \sim C_2 k^{-2d-2}$, *for some constant* $C_2 > 0$ *and large* k ($k > m$).

Also, for a short-memory ARMA *model,*

(c) $\sigma^2(k, m, k) - \sigma_y^2 \sim C_3 a_1^{-k}$, *for some constants* $C_3 > 0$, $a_1 > 1$ *and large* k ($k \leqslant m$);

(d) $\sigma^2(k, m, k) - \sigma_\varepsilon^2 \sim C_4 a_2^{-k}$, *for some constants* $C_4 > 0$, $a_2 > 1$ *and large* k ($k > m$).

According to Theorem 2, there are two different hyperbolic rates for the mean square prediction error during and after the data gap. For example, if $d = 0{\cdot}40$, the variance of the prediction error

during the gap increases at rate $O(k^{-0\cdot2})$, whereas it decreases to σ_ε^2 at rate $O(k^{-2\cdot8})$. Thus information is lost during the block of missing observations at a much slower rate than that at which it is gained after the data gap.

Theorems 1 and 2 assume the data gap length to be fixed. However, if the length of the gap increases to infinity, then the prediction process is statistically equivalent to that during the transition period at the beginning of the time series, with no previous observation. The following result characterises the convergence of the prediction error in this case.

THEOREM 3. *Let $\sigma^2(k, \infty, k)$ be the variance of the prediction error $e(k, \infty, k)$. Then, as $k \to \infty$,*

$$\sigma^2(k, \infty, k) - \sigma_\varepsilon^2 \sim Ck^{-1}.$$

An interesting relationship between predicting with finite and infinite past may be drawn from Theorem 3. Let $\tilde{e}(t, m, k)$ be the error of the best linear predictor of y_t based on the finite past $(y_{t-1}, y_{t-2}, \ldots, y_1)$, and let $\tilde{\sigma}^2(k, \infty, k)$ be its variance. Then

$$\tilde{\sigma}^2(k, \infty, k) = \sigma_y^2 \prod_{i=1}^{k-1} (1 - \phi_{ii}^2) = \sigma^2(k, \infty, k).$$

According to Theorem 3, this term converges hyperbolically to σ_ε^2, the error variance of the best linear predictor of y_t based on the infinite past $(y_{t-1}, y_{t-2}, \ldots)$.

2·3. *Influence of a single missing value*

THEOREM 4. *Under the conditions of Theorem* 1, *the mean square prediction error of an isolated missing observation is as follows*:

(a) *for $k > 0$, $\sigma^2(k, 1, k) - \sigma_\varepsilon^2 = \pi_k^2 \sigma^2(0, 1, k)$;*

(b) *for $k > 0$, if π_k is a monotonically decreasing sequence then $\sigma^2(k, 1, k) - \sigma_\varepsilon^2$ is a monotonically decreasing sequence converging to zero.*

According to Theorem 4(a) there is a jump in the mean square prediction error after a missing value. Its magnitude is $\pi_1^2 \sigma_\varepsilon^2$, unless $\pi_1 = 0$, as is true for some ARFIMA models. For instance, in an ARFIMA$(1, d, 1)$ process $\pi_1 = \theta - \phi - d$, and so there is no jump in the one-step mean square prediction error if $d = \theta - \phi$. The monotonicity condition in Theorem 4(b) is shared by all fractional noise processes with long memory parameter d. In fact, $\pi_j = \prod_{k=1}^{j} (k - 1 - d)/k > 0$ and so $\pi_k/\pi_{k+1} = (k + 1)/(k - d)$. However, for $d \in (-1, 1)$ and $k \geqslant 1$, $(k + 1)/(k - d) > 1$, proving that $\pi_k/\pi_{k+1} > 1$.

Figure 1 depicts the evolution of the mean square prediction error, $\sigma^2(k, 1, k)$, after a single missing value for ARFIMA$(0, d, 0)$ models with parameters $d = 0\cdot10$, $d = 0\cdot40$ and $d = 0\cdot49$, respect-

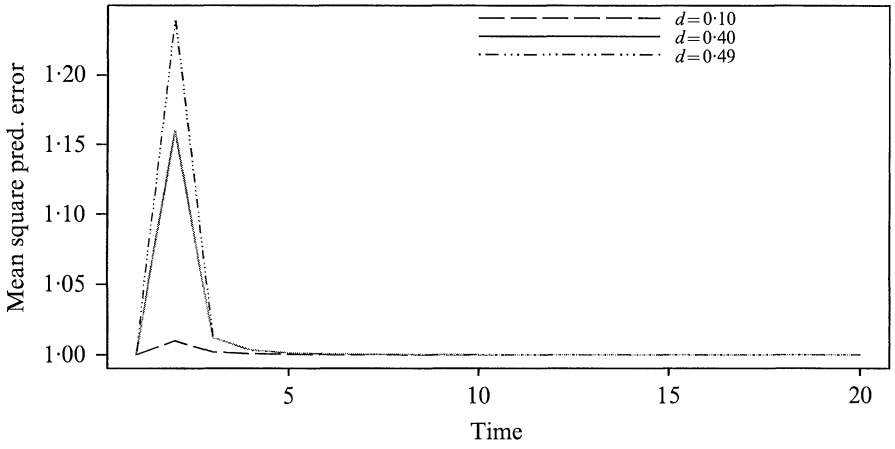

Fig. 1. Mean square prediction error for fractional noise processes.

ively. If we take $\sigma_\varepsilon^2 = 1$ it follows from $\pi_1 = \theta - \phi - d$ that the magnitude of the jump is d^2. Observe that, after the first peak, the mean square prediction error decays to one monotonically. Furthermore, the difference between the mean square prediction error and one is not noticeable after about six steps from the missing value.

The results discussed in this section indicate a sharp contrast between ARMA and ARFIMA processes. For example, from Theorem 2, the influence of a data gap on the mean square prediction error vanishes at an exponential rate for ARMA models and at a hyperbolic rate for ARFIMA processes. Thus, the influence on the prediction errors persists longer in the latter processes. On the other hand, as a consequence of Theorem 1(a) the forecasting error variance after the end of the series grows more slowly for ARFIMA than ARMA models, giving long-memory processes a clear advantage over short-memory processes.

3. APPLICATION: THE NILE RIVER DATA REVISITED

The annual minimum water levels of the Nile river measured at the Roda gorge is a well-known time series exhibiting long-range dependency; see for example Hosking (1984), Beran (1994, Ch. 1) and Hipel & McLeod (1994, Ch. 10). These measurements, available from Statlib at www.stat.cmu.edu, are displayed in Fig. 2(a) spanning a time period from AD 622 to AD 1921. Several blocks of repeated observations, i.e. consecutive years having exactly the same minimum water level, have been removed. Since the observations are specified by four digits, the repetitions are probably the result of a lack of new information.

From AD 622 to AD 1281, the period analysed by Beran (1994), there are 48 repeated values. This figure rises to 344 when the full period is considered, corresponding to roughly 27% of the sample size of 1297 observations. The problem is especially critical after the fifteenth century, when blocks of up to 55 consecutive repetitions can be found.

Following Beran (1994, p. 11), we fitted an ARFIMA$(0, d, 0)$ model to the Nile river data and the maximum likelihood estimates are presented in Table 1. The full period from AD 622 to AD 1921 was divided into two sub-periods, from AD 622 to AD 1281, which coincides with Beran's analysis, and from AD 1282 to AD 1921, a period in which 46% of the observations are missing. The means of the time series in Fig. 2(a), displayed in the third column of Table 1, are similar in the three

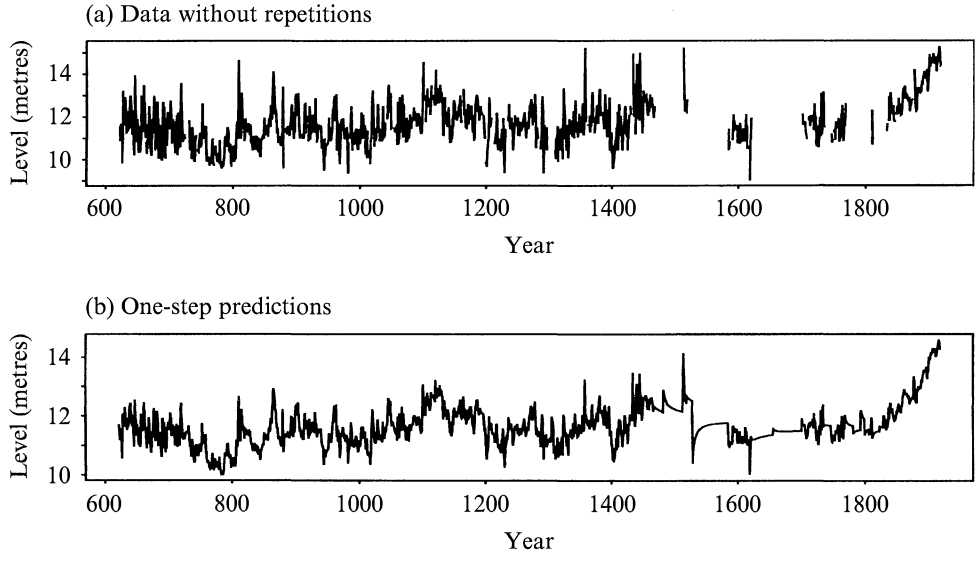

Fig. 2. Annual minimum water levels of the Nile river (AD 622–AD 1921).

periods, whereas the standard deviations of the time series, shown in the fourth column, present small changes from period to period.

Table 1: *Nile river data. Maximum likelihood estimation of d and* σ_ε^2

Period	Missing	$\bar{y}$	$\hat{\sigma}_y$	$\hat{d}$	$t_{\hat{d}}$	$\hat{\sigma}_\varepsilon$
AD 622–AD 1281	7%	11·50	0·89	0·3712	11·96	0·921
AD 1282–AD 1921	46%	12·08	1·17	0·4385	14·92	1·095
AD 622–AD 1921	27%	11·71	1·04	0·4141	18·21	0·733

From Table 1, it can be observed that the maximum likelihood estimates of the long-memory parameter d in the three periods are similar. The estimate found by Beran (1994, p. 125) is $\hat{d} = 0{\cdot}40$, for the period AD 622–AD 1281 without removing the repeated values. The maximum likelihood estimate of d for the second and third periods with the repeated values is 0·49 in both cases, indicating almost nonstationary models.

The t-statistics for $\hat{d}$ in the studied periods are highly significant. The estimated standard deviations of the noise $\hat{\sigma}_\varepsilon$ are close to one in the first and second periods. However, when the full period is considered, this estimate drops to around 0·7.

The influence of the data gaps on the forecasts can be analysed from the Kalman filter output. Figure 2(b) depicts one-step predictions. The residuals, $e_t = y_t - \hat{y}_t$, and the predictions' standard deviations are displayed in Fig. 3. The evolution of these standard deviations is explained by the theoretical results from § 2. Two typical situations are shown in Fig. 4. After a single missing value at time $t = 791$, see Fig. 4(a), the residual standard deviation jumps to roughly $0{\cdot}8 \simeq \sigma_\varepsilon\sqrt{(1+\pi_1^2)}$, and then decays to $\sigma_\varepsilon = 0{\cdot}733$ monotonically, in agreement with Theorem 4. It can be also observed in Fig. 4(a) that it takes approximately six steps after the missing value to reach the 0·733 level, analogously to the theoretical result displayed in Fig. 1 for $d = 0{\cdot}40$.

For a string of missing values, see Fig. 4(b), the mean square prediction error approaches the upper limit σ_y^2 monotonically at a hyperbolic rate $O(k^{-0{\cdot}26})$, see Theorem 2(a). Then, as new information is added, the error variance decays to σ_ε^2 at a hyperbolic rate $O(k^{-2{\cdot}74})$ as indicated

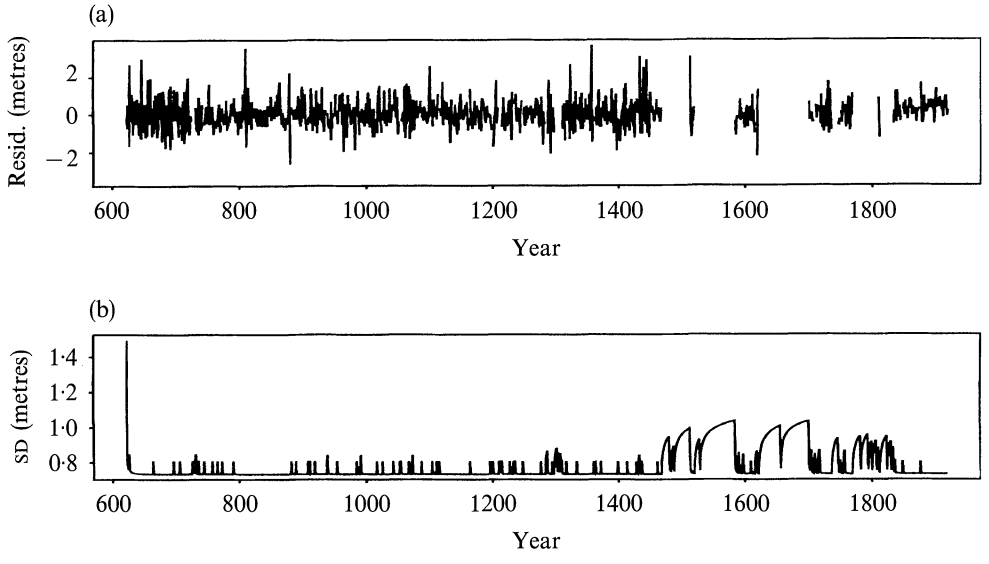

Fig. 3. (a) Residuals: $y_t - \hat{y}_t$, and (b) standard deviations of one-step predictions, for the Nile river data.

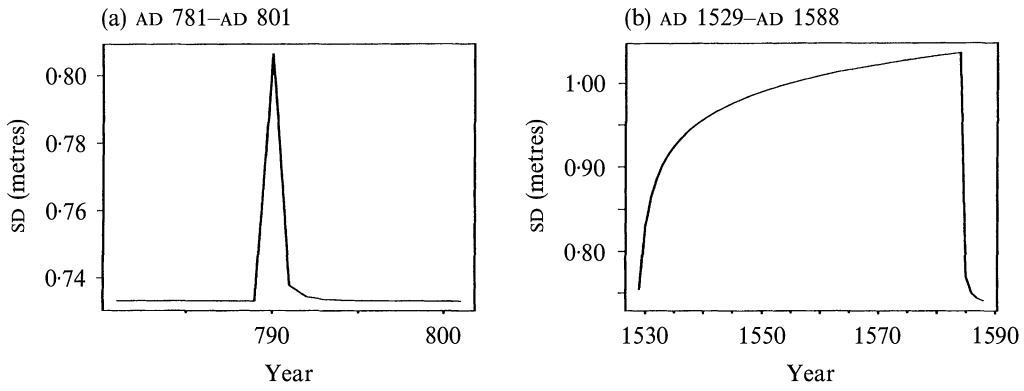

Fig. 4: Nile river data. Evolution of one-step standard deviations (a) AD 781–AD 801 period, (b) AD 1529–AD 1588 period.

by Theorem 2(b). It takes 55 missing observations for the mean square prediction error to increase to about one, but it takes fewer than six observations to regain the original level, σ_ε^2.

Model fitting for the Nile river application has been performed using the state space formulation, and a Fortran program that implements this approach is available from the authors. As shown by this application, data irregularities may severely affect the estimation of long-memory processes, resulting in almost nonstationary models.

ACKNOWLEDGEMENT

Part of this work corresponds to the fourth chapter of W. Palma's Ph.D. dissertation at Carnegie-Mellon University; he would like to thank N. H. Chan, J. B. Kadane and N. Terrin for their guidance and helpful comments and discussions. This research was supported in part by a grant from Fondecyt. The authors wish to thank the editor and two referees for their careful review of the paper.

APPENDIX

Proofs

Proof of Theorem 1. Part (a) is a standard result for the $t-s$ steps ahead error variance; see e.g. Beran (1994, p. 167). To prove (b) take $t=k$ in the AR(∞) decomposition and subtract from both sides the best linear predictor. Then all terms vanish, except for y_k and those associated with the missing observations, which yields the useful identity

$$e(k, m, k) = \varepsilon_k - \sum_{j=k-m+1}^{k} \pi_j e(k-j, m, k). \tag{A1}$$

By the orthogonality of ε_k to all previous observations,

$$\sigma^2(k, m, k) = \sigma_\varepsilon^2 + \operatorname{var}\left\{\sum_{j=k-m+1}^{k} \pi_j e(k-j, m, k)\right\}, \tag{A2}$$

for $m \geqslant 1$. Bounding the sum in (b) and $\sigma^2(j, m, k)$ by σ_y^2 ends the proof. □

Proof of Theorem 2. (a) For $k \leqslant m$, $\sigma^2(k, m, k) - \sigma_y^2 = \sigma_\varepsilon^2 \sum_{j=k}^{\infty} \psi_j^2$. Since $\psi_j \sim C_1 j^{d-1}$ for large j, $\sigma_\varepsilon^2 \sum_{j=k}^{\infty} \psi_j^2 \sim C_1 k^{2d-1}$, for large k. (b) For $k \gg m$, from (A2),

$$\pi_{k-m+1}^{-2}\{\sigma^2(k, m, k) - \sigma_\varepsilon^2\} = \operatorname{var}\left\{\sum_{j=k-m+1}^{k} \pi_{k-m+1}^{-1} \pi_j e(k-j, m, k)\right\};$$

for large k,

$$\pi_{k-m+1}^{-2}\{\sigma^2(k, m, k) - \sigma_\varepsilon^2\} \doteq \operatorname{var}\left\{\sum_{j=k-m+1}^{k} e(k-j, m, k)\right\}$$
$$= \operatorname{var}(Z_k),$$

where

$$Z_k = \sum_{j=0}^{m-1} y_j - E\left(\sum_{j=0}^{m-1} y_j | y_k, \ldots, y_m, y_{-1}, y_{-2}, \ldots\right).$$

However, $0 < \operatorname{var}(Z_\infty) \leqslant \operatorname{var}(Z_k) \leqslant \operatorname{var}(Z_m) < \infty$. Therefore

$$\sigma^2(k, m, k) - \sigma_\varepsilon^2 \sim C_2 \pi_{k-m+1}^2 \sim C_2 k^{-2d-2} \quad (k \gg m).$$

Parts (c) and (d) are proved analogously, by observing that for ARMA processes $\psi_j \sim C_3 a_1^{-j}$ and $\pi_j \sim C_4 a_2^{-j}$ for large j. □

Proof of Theorem 3. Let C be a constant which does not depend on k. Then, for large k,

$$\sigma^2(k, \infty, k) - \sigma_\varepsilon^2 = \sigma^2(k, \infty, k)\left\{1 - \frac{\sigma_\varepsilon^2}{\sigma^2(k, \infty, k)}\right\}$$
$$= \sigma^2(k, \infty, k)\left[1 - \exp\left\{\sum_{k+1}^{\infty} \log(1 - \phi_{ii}^2)\right\}\right]$$
$$\sim \sigma_\varepsilon^2\left\{1 - \exp\left(-C \sum_{k+1}^{\infty} i^{-2}\right)\right\} \sim Ck^{-1},$$

as required. □

Proof of Theorem 4. Taking $m = 1$ in (A1) yields (a). Part (b) follows from the monotonic behaviour of $\sigma^2(0, 1, r)$, which decreases from σ_ε^2 to $\sigma^2(0, 1, \infty)$, the variance of the interpolation error given all observations but the missing one. □

References

Ansley, C. F. & Kohn, R. (1983). Exact likelihood of vector autoregressive-moving average process with missing or aggregated data. *Biometrika* **70**, 275–8.

Beran, J. (1994). *Statistics for Long-Memory Processes.* New York: Chapman and Hall.

Bloomfield, P. (1992). Trends in global temperature. *Climatic Change* **21**, 1–16.

Chan, N. H. & Palma, W. (1998). State-space modelling of long-memory processes. *Ann. Statist.* **26**, 719–40.

Dahlhaus, R. (1989). Efficient parameter estimation of self similar processes. *Ann. Statist.* **17**, 1749–66.

Fox, R. & Taqqu, M. S. (1986). Large sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Statist.* **14**, 517–32.

Hipel, K. W. & McLeod, A. I. (1994). *Time Series Modelling of Water Resources and Environmental Systems.* Amsterdam: Elsevier.

Hosking, J. R. M. (1984). Modeling persistence in hydrological time series using fractional differencing. *Water Resour. Res.* **20**, 1898–908.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22**, 389–95.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Palma, W. & Chan, N. H. (1997). Estimation and forecasting of long-memory processes. *J. Forecasting* **16**, 395–410.

Penzer, J. & Shea, B. (1997). The exact likelihood of an autoregressive-moving average model with incomplete data. *Biometrika* **84**, 919–28.

RAY, B. K. & TSAY, R. S. (1997). Bandwidth selection for kernel regression with long-range dependent errors. *Biometrika* **84**, 791–802.

ROBINSON, P. M. (1993). Time series with strong dependency. In *Advances in Econometrics, 6th World Congress*, Ed. C. A. Sims, pp. 47–95. Cambridge: Cambridge University Press.

SOWELL, F. (1992). Maximum likelihood of stationary univariate fractionally integrated time series. *J. Econometrics* **53**, 165–88.

[*Received July* 1998. *Revised May* 1999]