1988

Advances In Time Series Analysis With Hydrological Applications

Mosquera Carlos Jimenez

Advances in Time Series Analysis with Hydrological Applications

by

Carlos José Jiménez Mosquera

Department of Statistical and Actuarial Sciences

Submitted in partial fulfilment
of the requirements for the degree of
Doctor in Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario
August 1988

ISBN 0-315-43295-0

ABSTRACT

The objective of this thesis is to develop and refine statistical methods which can be used for solving a variety of challenging problems which arise in the field of stochastic hydrology and elsewhere: Four areas in which more research is required in stochastic hydrology are long term memory processes and the Hurst phenomenon, seasonal geophysical processes, bootstrapping time series models and nonlinear and/or nonGaussian features present in geophysical tim^ series.

For addressing the first problem, we study the family of fractionally differencing autoregressive and moving-average processes (FARMA). Specifically, we characterize the efficiency of the sample mean and the efficiency of ordinary least squares estimates of regression parameters when the statistical error belongs to the FARMA family. Also, we prove that for FARMA models it is true that the logarithm of the determinant of the autocovariance matrix, divided by the number of observations converges to the logarithm of the variance of the white noise. Finally, the finite sample properties of the maximum likelihood method of

estimation are compared with the average likelihood method.

For the second problem in stochastic hydrology stated above, we investigate the family of periodic autoregressive and moving-average models (PARMA). We characterize these processes from the point of view of the state-space formulation. We give an exact maximum likelihood algorithm for fitting a PARMA model. Also, we give algorithms for computing the autocovariance matrix, the inverse autocorrelations, and the information matrix associated with the PARMA model.

An additional problem that initially was formulated for its relevance to stochastic hydrology is an improved bootstrapping procedure to investigate parameter uncertainty. We generalize and study some of the asymptotic properties of this procedure.

With respect to the fourth problem in stochastic hydrology, we investigate nonparametric function fitting when applied to time series processes. We give strong approximation results for these estimates, subject to certain mixing properties of the stochastic process. We extend a data based method for choosing the smoothing constant, and propose another method for doing this. We investigate a convex combination of a parametric estimate plus a nonparametric function estimate. Also, we study within a time series context a semi-parametric approach for function estimation. Finally, we use the kernel method for estimating state-dependent models.

ACKNOWLEDGEMENTS

Without the never ending support of both my supervisors Dr. McLeod and Prof. Hipel, I would have never been able to finish this thesis. Therefore, I wish to express my deepest thanks for their support in every aspect of this work.

I thank to the Department of Statistical and Actuarial Sciences: faculty, staff, students, past and present; because it was always been a wonderful place to work. I thank especially to Dr. MacNeill that helped in many different ways.

I wish to thank my parents and Tere's parents.

And, of course! I thank Tere, Verónica and Micaela, because they made every moment that I spent working in the thesis and not with them, both, sad and valuable.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

Autoregressive-moving average (ARMA) time series models have been popular among time series modelers ever since the publication of the seminal book of Box and Jenkins (1976). This is also true for researches investigating hydrological, and, in general, geophysical stochastic processes. However, it has also been apparent that Box-Jenkins models are not appropriate for every situation encountered by a time series analyst. In stochastic hydrology, this inadequacy is most apparent. In this thesis we investigate extensions of Box-Jenkins models, nonparametric function fitting and bootstrapping procedures that are motivated by hydrological applications. It should be noted that these applications are not unusual for hydrological time series analysis.

The first example of a hydrological application that motivates the use of a new time series model, is the Hurst phenomenon (Hurst, 1951). This phenomenon has motivated the definition of long term memory and of long-term memory processes (see for example McLeod and Hipel, 1978). It should be noted that long-term memory models are significant also from a probabilistic point of view, as they constitute one example where the classical conditions of the Central Limit Theorem for the sample mean of a stationary linear process are not fulfilled. Fractionally differenced autoregressive-moving average (FARMA) processes (Granger an Joyeux, 1980) are time series models that exhibit long-term memory, yet they are simple to describe. Previous studies of FARMA models have assumed a known mean. We relax this condition and show that maximum likelihood estimation and the approximate maximum likelihood estimation algorithm of Li and McLeod (1986) are asymptotically normal random variables. However, when the differencing parameter is negative, the sample mean is not asymptotically efficient. For exact maximum likelihood estimation we need to compute the determinant of the autocovariance matrix of the process. The Box-Jenkins' (1976) backforecasting approach for computing the sum of squares or Newbold's (1974) exact maximum likelihood method are not ideally suited for computing this determinant. Hence, there have appeared in the literature a variety of approximations for the determinant term, that are more easily computable. An extremely accurate approximation for the determinant of the autocovariance matrix of an ARMA process was given by McLeod (1977). In this thesis, we investigate if a similar approximation is possible for FARMA models. Also, both the behaviour of the determinant and the efficiency of the sample mean are linked to the behaviour of the extreme eigenvalues of the autocovariance matrix. Moreover, the behaviour of the extreme eigenvalues also affects the efficiency of estimates of regression parameters when the regression model has FARMA noise We investigate the asymptotic behaviour of these eigenvalues

The second example of a new time series model motivated by stochastic hydrology is the family of periodic ARMA or PARMA models Periodicity enters Box-Jenkins models multiplicatively. This implies that the governing equations of the process are constant over time. However, it seems reasonable to assume that for example, for a seasonal riverflows series that the riverflows in July depend on the riverflows in June differently than the riverflows in April depend on the riverflows in March. PARMA models allow a different ARMA model for each season. However, they are not a new type of model as they are equivalent to a multivariate ARMA process. Nonetheless, it seems worthwhile to study them in their own right because of the physical motivation just stated. Also, in most applications, fewer parameters are required to fit a PARMA model to a time series than to fit an equivalent multivariate model to the time series under consideration. Finally, PARMA models are amendable to generalizations to nonlinear and/or nonGaussian time series models that may be physically motivated.

It seems that the description of these processes from a state-space point of view would be useful for engineering purposes and also for the estimation of the parameters and for forecasting. Because this has not been done in the literature, we present several state-space representations, suited for different occasions. Based on the state-space equations, we give conditions to ensure that the governing equations of an PARMA model define a stationary time series model. We give an exact maximum likelihood algorithm based on the Kalman filter. This algorithm is faster than previous algorithms (Vecchia, 1985), making it possible to fit PARMA models instead of just PAR models. Additionally, we give expressions for the autocovariance matrix plus two algorithms for computing it. Autocorrelations, partial autocorrelations, and inverse autocorrelations are useful for identification of the time series process under study. However, it was stated by Sakai (1982) that partial correlations are not useful for the identification of periodic models. We rederive Sakai's Levison-Durbin equations for the partial autocorrelations to show that they have a cut-off property that may be used for identification. We give a definition of the inverse autocorrelations of PARMA models. They exhibit a truncation property that may be used for identification. The inverse autocorrelations may be estimated using a algorithm similar to the one used for estimating inverse autocorrelations of ARMA processes. For purposes of assessing how good is the fit that PARMA models provide for a particular time series, it is useful to compute the information matrix. We give two algorithms for computing the information matrix. Also, for purposes of forecasting it is useful to have an expression of the mean square error (MSE) of the forecast when the parameters are estimated. We present one. Finally, autoregressive models may be generalized to models with exponential marginal univariate distributions but whose autocorrelation function behaves as an autoregressive model. This is the EAR(p) and the next to exponential autoregressive, NEAR(2), type of models. We generalize these models to the case of a periodic time series having an exponential marginal density.

A third example of a statistical methodology that has been proposed by considerations about stochastic hydrology is the bootstrap methodology put forward by Cover and Unny (1986). This approach to the bootstrap was motivated by concerns about parameter uncertainty in the fitting of time series models. We generalize their procedure by envisioning the bootstrap as the resampling of the loss function that we are using for estimating the parameters. Also, we give another procedure for bootstrapping, which is based on slight perturbations of the observations. Both definitions are interesting because the other bootstrapping procedure for regression like models (Freedman, 1981) imposes a particular model on the observations, even if the observations do not follow the model. Our definitions do not have this drawback.

We give several asymptotic results to investigate if the bootstrap distribution coincides asymptotically with the distribution of the estimated parameters. For the mean of a linear process it is possible to give more accurate results than for the general case. We give these results.

In addition, we give some applications of these bootstrapping procedures.

Finally, the necessity of using nonlinear and nonGaussian time series models has arisen in many contexts and not just stochastic hydrology. A particular behaviour in hydrology that may be seen as nonGaussian and/or nonlinear is the seesaw appearance of seasonal riverflows. We investigate the extension of the kernel density and function estimation to the context of a time series process. We give results similar to those of Silverman (1978), for both density estimation and function estimation. Also, the bias and the variance of the nonparametric estimates are investigated. We extend Rudemo's (1981) data based approach for choosing the smoothing constant. Moreover, we also propose a quick method of choosing the constant by preserving a set of quantiles. We obtain for the case when we allow the kernel to belong to the class of kernels of order p optimal kernels for density estimation.

On many occasions, one assumes that a particular parametric family may give a good approximation to the true density or function that we are estimating. We attempt to blend this parametric estimate with a nonparametric estimate by using a convex combination of both estimates. Although, the asymptotic rate of convergence is as slow as the rate of convergence of the nonparametric estimate, it may be sensible to use this convex combination from a preasymptotic point of view. We present some results about the consistency of the parametric estimate.

Finally, we generalize the autoregressive equations to allow for models with state-dependent coefficients (Priestly, 1980). We use a least squares type of procedure with state-dependent weights to obtain estimates of these autoregressive coefficients. Additionally, we prove the consistency and asymptotic normality of these estimates.

References

1. Box, G.E.P., and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control* (Second Edition). Holden-Day, San Francisco, California.

2. Cover K. A.,and Unny, T. E. (1986). Application of computer intensive statistics to parameter uncertainty in streamflow synthesis. *Water Resources Bulletin*, **22**, 495–499.

3. Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Mathematical Statistics*, **9**, 1218–1228.

4. Granger, C. W. J. and R. Joyeux (1980). An introduction to long memory time series models and fractional differencing. *J. Time Series Analysis*, **1**, 15–29.

5. Hurst, H. E. (1951). Long-term storage capacity of reservoirs *Trans. Amer. Soc. Civil Eng.*, **116**, 770–808.

6. Li, W. K. and A. I. McLeod (1986). Fractional time series modelling. *Biometrika*, **73**, 165–176.

7. McLeod, A. I. (1977). Improved Box–Jenkins estimators. *Biometrika*, 64, 531–534.

8. McLeod, A. I. and K. W. Hipel (1978a). Preservation of the rescaled adjusted range, part one. A reassesment of the Hurst phenomenon. *Water Resources Research*, 14, 491–508.

9. Newbold, P. (1974). The exact likelihood for a mixed autoregressive moving-average process. *Biometrika*, , 61, 248–256.

10. Priestly, M. B. (1980). State-dependent models: A general approach to non-linear time series analysis. *J. Time Series*, , 1, 47–71.

11. Rudemo, M. (1982). Empirical Choice of histogram and kernel density estimators. *Scan. J. Statist.*, 9, 65–128.

12. Sakai, H. (1982). Circular Lattice Filtering Using Pagano's Method. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-30, 279–286

13. Silverman, B. W. (1978). Weak and strong uniform consistency of a density estimate and its derivatives. *Annals of Mathematical Statistics*, 6, 177–184.

14. Vecchia, A. V. (1985). Maximum Likelihood Estimation for Periodic Autoregressive-Moving Average Models. *Technometrics* 27,

15. Vecchia, A. V. (1986): Periodic Autoregressive-Moving Average (PARMA) Modelling with Applications to Water Resources. In Special publication on Time Series Analysis in Water Resources. K. Hipel, editor.

Chapter 2

FRACTIONALLY DIFFERENCED MODELS

2.1 Introduction

Some geophysical processes exhibit what can be described as persistent behaviour. Persistence or long term memory is the term used to describe a time series whose autocorrelation structure decays slowly to zero or, equivalently, whose spectral density is highly concentrated at frequencies close to zero. Such autocorrelation structure suggests that the process must depend strongly upon values of the time series far away in the past; hence, in order to model the process the whole past should be incorporated into the description of the process. Because the autocorrelation structure of an autoregressive-moving average (ARMA) process damps out exponentially, some researches think that ARMA models are inappropriate to describe persistent time series which have autocorrelation functions (ACF) which decay to zero slowly as the lag increases.

The concept of persistence was introduced into hydrology and then into statistics as a result of the investigation of the Hurst phenomenon (Hurst, 1951). A good exposition of the Hurst phenomenon with respect to many geophysical processes is given in Hurst et al. (1965) and in Hipel and McLeod (1989). In addition to other authors, Mandelbrot and Wallis (1968), Klèmes (1974), Hipel and Mcleod (1978), McLeod and Hipel (1978a), Boes and Salas (1973, 1978), and Salas et al. (1979) describe the Hurst phenomenon and suggest various approaches for modelling persistence in hydrological time series.

Various definitions of persistence are available in the literature. Perhaps the most valuable definition, because of its simplicity and because it captures the essence of 'persistence', was given by McLeod and Hipel (1978a) in connection with the Hurst phenomenon. They classified a time series process according to the behaviour of the memory of the process where memory is defined as

$$\mathcal{M} = \sum_{0}^{\infty} |\rho_n|, \tag{2.1}$$

where ρ_n is the theoretical ACF at lag n for the process. A long term memory process is defined as a process with $\mathcal{M} = \infty$, whereas a short term memory process has $\mathcal{M} < \infty$. The $\mathcal{M}$ value had already been used as a mixing coefficient for stationary time series (Brillinger, 1976) giving an index of the rate at with the present values of the time series are dependent on the values in the far away past. If $M < \infty$, one could assume that points in the time series that are well separated are asymptotically independent. This asymptotic independence is useful to extend to time series processes probabilistic results that are valid when the observations are independent, such as: normality of averages, the asymptotic behaviour of a quantity like the sample ACF, parameter estimates obtained either by maximum likelihood estimation or by the method of moments, hypothesis testing, Portmanteau tests, and so on. Thus, $M < \infty$ has been traditionally assumed in time series analysis. But then most of the results usually used in time series analysis, as given in books such as those by Brillinger (1976) and Hannan (1970), are not necessarily true for long term memory processes because these processes have an infinite memory.

An alternative definition of long term memory, essentially equivalent to the definition given above, is to consider time series processes whose ACF decays as

$$\rho_n = O(n^{-\alpha}), \tag{2.2}$$

where α lies in the interval $(0, 1)$.

Besides geophysical time series, the classification of time series according to short and long memory is useful in other applied areas (Cox, 1984; Parzen, 1982) such as economies (Granger, 1980; Granger and Joyeux, 1980)and statistical mechanics (Burton and Waymire, 1983; Newman, 1980; Taqqu 1979). The use of a memory index has also been found useful with other types of stochastic processes (Cox, 1984), although the memory has had different definitions.

Several parametric models that are able to exhibit persistence have appeared in the literature. These long memory models include fractional Gaussian noise (Mandelbrot and Wallis, 1969a,b,c), broken-line models (Mejia et al., 1972), shifting level processes (Boes and Salas, 1978; Ballerini and Boes, 1985), and fractionally differenced autoregressive-moving average (FARMA) models (Granger, 1980; Granger and Joyeux, 1980; Hosking, 1981, 1984, 1985; Li and McLeod, 1986). Many of the foregoing long memory models are also described in textbooks on stochastic hydrology by authors such as Kottegoda (1980), Salas et al. (1980), Brass and Rodriguez-Iturbe (1984) and Hipel and McLeod (1989).

2.2 Fractionally differenced autoregressive moving average time series models

A device frequently used in time series modelling is to difference the series if it is thought that its mean function is time dependent. A time-dependent mean could produce sample autocorrelations that do not decay to zero exponentially, but instead decay to zero much more slowly. If the rate of decay of the ACF seems to depend linearly upon the lag, the usual approach is to use the first difference of the time series. For the type of processes studied in this chapter, the ACF decays to zero at a rate slower than exponential but faster than linear. In fact, the rate of decay may be described as hyperbolic. This suggests the use of an operator similar to the usual differencing operator to model time series having a slowly decaying ACF with long memory. FARMA models generalize in a natural form the concept of differenced ARMA time series models.

DEFINITION 2.1 To define FARMA models, first the concept of differencing is generalized by means of the filter

$$\nabla^d(B) = (1 - B)^d = \sum_0^\infty c_n B^n, \tag{2.3}$$

where

$$c_n = (-1)^n \binom{d}{n}$$

d is a real number and B denotes the backshift time operator. Then, a FARMA time series process $\{X_t\}$ is a process that satisfies the following equation

$$\Phi(B)\nabla^d(X_t) = \Theta(B)a_t. \tag{2.4}$$

That is, $\{X_t\}$ is the output that results after the application of the filter ∇^d and a ARMA filter $\Theta(B)/\Phi(B)$ to a white noise, $\{a_t\}$, time series process; where $\Theta(B)$ and $\Phi(B)$ are the moving average and autoregressive operators, respectively. In addition, we will assume that $\{a_t\}$ is Gaussian. ∇

The parameter d controls the memory of the process. If d is nonpositive, the memory is finite, otherwise the process has long term memory. Also, when considering the Hurst phenomenon, d is equal to the Hurst coefficient minus 1/2.

The qualitative difference between ARMA processes and FARMA processes as seen through (2.4) is that the partial correlation coefficients of ARMA processes approach zero exponentially fast; whereas, the partial regression coefficients of FARMA processes approach zero as $1/n^a$ (for FARMA$(0,d,0)$, $\alpha = 1$).

For a nonseasonal FARMA model, we employ the notation FARMA(p,d,q) where p and q are the orders of the autoregressive and moving average operators, respectively, and d is the parameter in the filter in (2.4) and takes on real values. When d is a positive integer, the FARMA(p,d,q) model is equivalent to an ARIMA(p,d,q) model where the acronym ARIMA stands for autoregressive integrated moving average. If d is equal to zero, the FARMA(p,d,q) model is identical to a short memory ARMA(p,q) model. Seasonal FARMA models can be defined in a manner similar to that used for seasonal ARMA and ARIMA models.

REMARK 2.2 The definition of the filter ∇^d involves, formally, the whole past. However, for Gaussian FARMA time series processes we could give another equivalent definition as follows. Starting from some point in time, say $t = 1$, we can think of a FARMA process as a stationary time series process with an autocovariance structure identical to the autocovariance structure of the process defined in (2.4). Then, using the partial linear regression coefficients $\{\phi_{t,k}\}$, we can define the FARMA process by

$$X_t = \sum_{k=1}^{t} \phi_{t,k} X_{t-k} + e_t, \tag{2.5}$$

where $\{e_t\}$, is a white noise process with variance

$$\mathrm{var}(e_t) = \mathrm{var}(e_{t-1})(1 - \phi_{t,t}^2). \tag{2.6}$$

The interest of this definition is that it does not involve the infinite past. The advantage of the definition (2.4) is its simplicity in the sense that just a few parameters need to be specified. However, using (2.5) we could define new families of Gaussian long-term dependent processes if we choose the partial correlation coefficients, $\{\phi_{t,t}\}$, such that they decay to zero hyperbolically.

2.3 Properties

Some basic properties of FARMA processes have been described by Hoskings (1981) and by Granger and Joyeux (1980). They found among other things that:

(a) In order for the process to be stationary $d < 0.5$.

(b) In order for the process to be invertible $d > -0.5$.

(c) The ACF behaves as

$$\rho_n = O(n^{-1+2d}). \tag{2.7}$$

Several researchers (Rosenblatt 1961, 1979, 1981; Taqqu 1975, 1979) have studied the behaviour of statistics derived from time series processes where the ACF behaves as in (2.7) and d is positive. They found that:

(d) The sample mean times $N^{1/2+d}$, where N is the number of observations, converges in law to a normal random variable.

(e) The sample autocovariances do not converge asymptotically to a normal random variable.

The result in (d) about the mean is of some interest in applications because it has been found that processes thought to possess long term memory have a sample mean that seems to indicate slow changes in trend (Boes and Salas, 1978). This wandering of the sample mean may be explained in terms of (d) above. The existence of several approaches for producing persistent behaviour indicates that in order to show that the presence of persistent behaviour in a time series processes is due to a slowly changing trend we need actual insight into the generating mechanism and not only an analysis of the statistical behaviour. Note also that the above results are not restricted to the FARMA process case but that they are valid for any time series whose ACF behaves as in (2.6).

An important consequence of the slow rate of decay to zero of the ACF as given by (2.7) is that Bartlett's formula (Bartlett, 1946) for the variances and the covariances of the estimated autocovariance function, ACVF, $\{\hat{\gamma}_k\}$, has to be modified accordingly. In fact, the exact formula for the variance is given by

$$\mathrm{var}(\hat{\gamma}_k) = N^{-1} \sum_{m=-(N-k)+k}^{(N-k-1)} \{1 - \frac{|m - k}{N}\}\{\gamma_m^2 - \gamma_{m+k}\gamma_{m-k}\}, \qquad (2.8)$$

Using (2.7), we get

$$\mathrm{var}(\hat{\gamma}_k) = \begin{cases} O(N^{-1}), & \text{if } d \leq 0.25; \\ O(N^{4d} - 2) & \text{if } d<0.25. \end{cases} \qquad (2.9)$$

Hence, if $d < 0.25$ then $\mathrm{var}(\hat{\gamma}_k) = O(N^{-1})$, which is the same order as in the case of a short memory process. However, if $0.25 < d < 0.5$ the order of $\mathrm{var}(\hat{\gamma}_k)$ is larger than N^{-1}. In fact, as d approaches 0.5 the variance approaches a quantity of order one. This implies that the stochastic variability of the estimate $\hat{\gamma}_k$ of the ACVF is higher for long term memory processes with $0.25 < d < 0.5$ than for short term memory processes. Moreover, the order of the variance depends on the unknown quantity d. Finally, similar results are valid for the covariances of the estimated ACF.

An important, although seemingly trivial, extension of the original definition by Hosking (1981) and Granger and Joyeux (1980) of FARMA processes is to eliminate the assumption that the mean of the time series is zero. The extension of the model given in (2.4) to the case of a nonzero mean is straightforward. However, it is important to note that if a constant, in particular the mean, is passed through the filter ∇^d, the output, for the case of a positive d, is zero. Hence, the mean of the process does not have to appear in the equations that define the model. Nevertheless, it should be noted that the mean is a well defined quantity for this process when $d < 1/2$.

The aforementioned property is very important for determining the stochastic properties of the estimates for the parameters. This is because the sample mean can be used as an estimate for the mean of the time series and the slow rate of convergence of the sample mean as given in (d) above does not affect the asymptotic rate of convergence of the estimates for the other parameters to a Gaussian random variable, where this rate is the usual N^{-1}.

PROPOSITION **2.3**

(a) If d is positive (2.4) can be written as

$$\Phi(B)\nabla^d(X_t - c) = \Phi(B)\nabla^d(X_t) = \Theta(B)a_t, \tag{2.10}$$

for any constant c.

(b) The operator ∇^d smooths the trend function $f(t) = \binom{-d}{t}$, $t \geq 0$, and 0 otherwise.

Proof This is due to the fact that 1 is a zero of the complex function $(1 - z)^d$ and standard properties of the binomial coefficient.

Part (a) applies in particular to the mean or the maximum likelihood estimate. Part (b) is interesting because it has been suggested that the Hurst phenomenon is due to a slowly changing trend (Leopold et al., 1964; Bhattacharya et al., 1983) and perhaps the usefulness of fractionally differenced time series when applied to rivers flow modelling or some other geophysical modelling could be due to this smoothing of the trend and not to the long memory structure. An interesting consequence of Part (b) is that the filter ∇^d can smooth some particular trends. If $d = 1$, it will smooth a trend that corresponds to a straight line. When $0 \leq d < 1/2$ the filter ∇^d smooths slowly changing trends. Hence, even if the process mean is slowly changing, FARMA models could be used to model the time series in much the same way that ARIMA models are used with a deterministic drift component.

A consequence of part (a) in Proposition 2.3 is that the behaviour (long term) of the process is not dependent on the mean as in a stationary ARMA process. On the other hand, the local behaviour of the process does depend on the mean. This can be seen by considering the value of the process conditioned on the past as given by $\mathbb{E}\{x_{t+1}; x_s, s \leq t\}$. In the ARMA case this quantity depends on $\mu = \mathbb{E}\{x_t\}$ but in the FARMA case with $d > 0$, it does not. In the remainder of the chapter, unless stated to the contrary, the mean μ will be assumed equal to zero.

An interesting subset of the FARMA(p, d, q) family of processes is the FARMA$(0, d, 0)$ process which is referred to as the fractional differencing model. This model has been studied in some detail and expressions for the ACF, partial autocorrelations function (PACF), partial linear regression coefficients, and inverse autocorrelations are known (Hosking, 1981). One important fact about the stochastic behaviour of a FARMA$(0, d, 0)$ process is that all its autocorrelations are positive if d is positive, and they are negative otherwise. Also, all the partial autocorrelations of the FARMA$(0, d, 0)$ model have the same sign as the persistence parameter d, and their rate of decay to zero is of the same order as the inverse of the lag.

2.4 Parameter estimation

Orthogonal polynomials on the unit circle

In what follows, we will use some results on orthogonal polynomials of the complex unit circle C. We will review some selected results. Most of these results can be found in Szegõ (1935), Grenander and Szegõ (1955), or Geronimus (1961, 1977).

Given a weight function $F(\lambda)$, $\lambda \in [0, 2\pi)$, we consider the complex polynomials $\phi_k(z)$, of degree k that are orthonormal in the unit circle with respect to $F/2\pi$; i.e.

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_i(z)\overline{\phi_k(z)}\, dF(\lambda) = \delta_{ik}, \qquad z = e^{i\lambda}, \tag{2.11}$$

where δ_{ik} is the Kronecker delta symbol, and '$^{-}$' denotes the complex conjugate operator, (Grenander and Szegõ, 1955, p. 37). The connection of $\phi_k(z)$ with second order stationary stochastic processes is obvious if we see F as the spectral measure of the process $\{X_t\}$. We will assume that F is absolutely continuous, i.e. there is a function $f(\mu)$ such that

$$\frac{1}{2\pi}\int_{-\pi}^{\lambda} f(\mu)\, d\mu = F(\lambda). \tag{2.12}$$

From the point of view of time series, these polynomials can be written, except for a multiplicative constant, as $\phi_k(z) = \sum_{j=0}^{k} \phi_{kj} z^j$, where the ϕ_{kj} are the partial regression coefficients of $\{X_t\}$. Another way of looking at these polynomials is as follows: if $\underline{\phi_k}$ is the $1 \times k$ vector obtained from the first k elements of the kth row of the square root matrix that corresponds to the Cholesky decomposition of the inverse $N \times N$ autocovariance matrix Σ^{-1}, then

$$\phi_k(z) = \underline{\phi_k} \begin{pmatrix} 1 \\ z \\ \vdots \\ z^k \end{pmatrix}. \tag{2.13}$$

A very important relationship between the polynomials $\phi_n(z)$ is the following recurrent expression:

$$\alpha_{n+1}\phi_{n+1}(z) = \alpha_n z \phi_n(z) - l_{n+1}\phi_n^*(z), \tag{2.14}$$

where $\phi_n^*(z) = z^n\bar{\phi}(1/z)$, is the 'backward polynomial' associated with $\phi_n(z)$, α_n is the coefficient associated with z^n in the polynomial $\phi_n(z)$, and l_n is the constant term in $\phi_n(z)$ (Grenander and Szegő, 1955, p. 41). This relationship is just the Levison-Durbin algorithm, as can be seen by equating the coefficients of z^i in both sides of the equation (2.14). Also,

$$\begin{aligned} \alpha_n &= \left(\frac{\det(\Sigma_{n-1})}{\det(\Sigma_n)}\right)^{\frac{1}{2}} \\ &= \sigma_n^{-1}. \end{aligned} \tag{2.15}$$

The coefficient associated with the n-th power of z, α_n, and the constant l_n are related by the relationship

$$\alpha_{n+1}^2 - \alpha_n^2 = l_{n+1}^2. \tag{2.16}$$

Hence, if $\lim_{n\to\infty} \alpha_n$ exits, and is equal to α, then

$$\sum_{j=n+1}^{\infty} l_j^2 = \alpha^2 - \alpha_n^2, \tag{2.17}$$

and vice versa. By examining (2.7), we find that what we have said is that the one step ahead error converges to some nonzero quantity if and only if the product of the factors $1-\phi_{n,n}^2$ converges, where $\phi_{n,n}$ is the nth partial correlation of the process. A most basic result about this limit determinant was given by Szegõ (1921) and improved by Kolmogorov (1935).

THEOREM **2.4** (Szegõ, 1921; Kolmogorov, 1935) If $\ln f$ is $\boldsymbol{L}_1$ integrable i.e.

$$\int_{-\pi}^{\pi} \ln f(\lambda)\, d\lambda > -\infty \tag{2.18}$$

then $\lim_{n\to\infty} \sigma_n^2 = \sigma^2$, where σ^2 is given by

$$\sigma^2 = \exp\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(x)\, dx\}. \tag{2.19}$$

∇

REMARK **2.5** The important part of the theorem is not that the one step ahead error converges (it has to, as it is a positive decreasing sequence), but that the limit is given by (2.19). For the situations where the condition of the theorem fails, the theorem is still true if $f(\lambda)$ is positive. We see then that although the theorem is still valid for fractionally differencing processes, the case of antipersistence, that is when the differencing parameter is negative, may pose difficulties.

Another important quantity is the reproducing kernel function defined by

$$k_n(x,z) = \sum_{j=0}^{n} \overline{\phi_j(x)}\phi_j(z). \tag{2.20}$$

Note that

$$k_n(x,z) = (1,x,\ldots,x^n)\Sigma^{-1}(1,z,\ldots,z^n)'. \qquad (2.21)$$

Parzen (1962) has explored the connection of reproducing-kernel Hilbert spaces and time series. One reason we are interested in the kernel is because it satisfies the Christoffel-Darboux formulae

$$k_n(x,y) = \frac{\phi_n^*(x)\overline{\phi_n^*(y)} - x\bar{y}\phi_n^*(x)\overline{\phi_n^*(y)}}{1 - x\bar{y}}, \qquad (2.22)$$

and

$$k_n(x,y) = \frac{\phi_{n+1}^*(x)\overline{\phi_{n+1}^*(y)} - \phi_{n+1}^*(x)\overline{\phi_{n+1}^*(y)}}{1 - x\bar{y}}. \qquad (2.23)$$

Call $\varrho_n(x) = k_n(x,x)$.

If (2.18) is satisfied, the function

$$\pi(z) = \lim_{n\to\infty} k_n(0,z)$$

is well defined in the $\mathbb{L}_2(f)$ and has norm α. The best approximation to $\pi(z)$, in this norm, by a polynomial of degree n is given by $k_n(0,z)$; the $\mathbb{L}_2(f)$ norm of the difference is $\sum_n^\infty l_n^2 = \alpha^2 - \alpha_n^2$; also, $\alpha_n\phi_n^*(z) = k_n(0,z)$ and

$$\frac{1}{|\pi(e^{i\lambda})|^2} = f(\lambda) \qquad \text{almost everywhere.}$$

Hence,

$$|\alpha_n\phi^*(e^{i\lambda})|^2 = (f(\lambda))^{-1} + O(\alpha^2 - \alpha_n^2) \qquad (2.24)$$

Estimation of the mean

We will now consider the estimation of the mean. When d is positive the differencing operator ∇^d filters out constants. When d is negative, ∇^d is not defined for nonzero constant terms. Then, we have to see the mean of the stochastic process $\{X_t\}$ as a regression component, i.e.

$$X_t = \mu - Y_t,$$

where Y_t is a process that has a spectral density $f(\lambda)$ equal to $f(\lambda) = (2\sin(\lambda/2))^{-2d}g(\lambda)$ and $g(\lambda)$ is a continuous bounded and positive spectral density, also, Y_t has zero mean.

Because of the rate of convergence of the variance of the sample mean to zero is not the standard $O(1/N)$, as can be seen in (e) in section 2.3, it is important to find a more efficient estimator of the mean. The most obvious candidate is the maximum likelihood estimate of the mean, which is given by

$$\hat{\mu} = \mathbf{1}'\Sigma^{-1}X/\mathbf{1}'\Sigma^{-1}\mathbf{1}, \tag{2.25}$$

where Σ is the autocovariance matrix of the time series, $\mathbf{1}$ represents a column vector of ones, and X represents the vector of observations. The variance of the maximum likelihood estimate, $\hat{\mu}$, of μ is given by

$$\text{var}(\hat{\mu}) = \left(\mathbf{1}'\Sigma'\mathbf{1}\right)^{-1}.$$

This can be expressed as $\text{var}(\hat{\mu}) = (k_N(1,1))^{-1}$, as can be seen by looking at (2.21). Hence by (2.20)

$$\mathrm{var}(\hat{\mu}) = \left(\sum_{0}^{N} |\phi_k(1)|^2 \right)^{-1}$$

can be written in terms of the complex orthonormal polynomials; see Grenander and Szegö (1955, p. 209).

THEOREM **2.6** The variance of the m.l.e. mean of a $\{X_t\}$ stationary time series process with bounded spectral density satisfies asymptotically

$$\mathrm{var}(\hat{\mu}) = \frac{f(0)}{N} - O(\alpha^2 - \alpha_n^2). \tag{2.26}$$

▽

REMARK **2.7** Note that if $f(0) = 0$, the order of convergence is not $O(1/N)$.

REMARK **2.8** Theorem (2.6) has been known for a long time for the case of a positive bounded spectral density f.

Next, we investigate the case when f is unbounded at zero or when it has a zero at zero. Now, using (2.22) we see that

$$\lim_{r \to 1^-} \frac{1}{\varrho_n(r)} = \mathrm{var}(\hat{\mu}). \tag{2.27}$$

THEOREM **2.9** The variance of the m.l.e. mean $\hat{\mu}$ can be obtained from

$$\mathrm{var}(\hat{\mu}) = \left\{ \phi_N(1) \sum_{i=1}^{N} i[\phi_{N,N-i} - \phi_{N,i}] - \phi_N(1) \right\}^{-1}. \tag{2.28}$$

▽

Proof We just have to obtain the limit in (2.27), use L'Hospital rule and then simplify.

□

Because the spectral density of a FARMA process has a zero or it is unbounded at zero, standard results about the efficiency of the sample mean are not immediately applicable. For example, Grenander and Szegö (1955) show that in the case of a moving average process of order one

$$x_t = a_t - a_{t-1}$$

the variance of the sample mean is of order $0(1/N^2)$ whereas the variance of the m.l.e. estimate is of order $0(1/N^3)$. The problem with this example lies in the fact that the spectral density of the process has a zero at zero. But then it is of interest to establish for FARMA time series if the sample mean, $\bar{x}$, is an efficient estimate of the mean, μ. We proceed to do so. The variance of the sample mean is given by

$$\operatorname{var}(\bar{x}) = \frac{\gamma_0}{N} \sum_{k=-(N-1)}^{N-1} (1 - \frac{k}{N})\rho_k, \tag{2.29}$$

where γ_0 is the autocovariance of $\{X_t\}$. Hence, if the spectral density of X_t is given by $f(\lambda) = (2\sin(\lambda/2))^{-2d} g(\lambda)$, and g is a bounded and continuous function in $[-\pi, \pi)$ then

$$\begin{aligned}
\operatorname{var}(\overline{X}) &= \frac{1}{2\pi N^2} \int_{-\pi}^{\pi} \frac{(\sin(N\lambda/2))^2}{(\sin(\lambda/2))} f(\lambda)\, d\lambda. \\
&= \frac{2}{\pi N^2} \int_{-\pi}^{\pi} \frac{(\sin(N\lambda/2))^2}{(2\sin(\lambda/2))^{-2(1+d)}} g(\lambda)\, d\lambda. \\
&= \frac{1}{\pi N^2} \int_{-\pi}^{\pi} \frac{(1-\cos(N\lambda))}{(2\sin(\lambda/2))^{-2(1+d)}} g(\lambda)\, d\lambda.
\end{aligned}$$

$$\operatorname{var}(\overline{X}) = \frac{2}{N^2} \{\gamma_0(1-d) - \gamma_N(1-d)\}. \tag{2.30}$$

This proves the following result.

THEOREM 2.10 The variance of the sample mean of a stationary $\{X_t\}$ process with spectral density given by $f(\lambda) = (2\sin(\lambda/2))^{-2d} g(\lambda)$ where g is a bounded and continuous function in $[-\pi, \pi)$, is given by (2.30) where $\gamma_i(1+d)$ corresponds to the ith autocovariance of a process with spectral density $f(\lambda) = (2\sin(\lambda/2))^{-2(1+d)} g(\lambda)$. When d is positive, these autocovariances are understood in a formal way. ▽

Now, to study the asymptotic behaviour of the m.l.e mean we have, by (2.28), to study the behaviour of

$$\frac{1}{N} \sum_{i=1}^{N} i[\phi_{N,N-i} - \phi_{N,i}].$$

But, asymptotically $\phi_{N,N-i} - \phi_{N,i} = \phi_{N+1,N+1-i}$, using (2.14). Then, the Césaro sum

$$\frac{1}{N} \sum_{i=1}^{N} i\phi_{N,N-i}.$$

converges. Hence

$$\operatorname{var}(\hat{\mu}) = |\phi_N(1)|^{-2}/N + o(1/N). \tag{2.31}$$

Now, by (2.30) if $1 + d$ is less than $1/2$, i.e. the variances γ_N are well defined, then $\mathrm{var}(\overline{X}) = 2\gamma_0(1 + d)/N^2 + o(1/N^2)$. This is in agreement with the example about the noninvertible moving average process discussed above. If $1 + d > 1/2$, then the term $\gamma_N(1 + d)$ increases with N, and hence the same order of convergence is not achieved. To see how to estimate the order of convergence note that

$$\begin{aligned}\mathrm{var}(\overline{X}) &= \frac{2}{\pi N^2}\int_{-\pi}^{\pi} \frac{(\sin(N\lambda/2))^2}{(2\sin(\lambda/2))^{-2(1+d)}}\{g(\lambda) - g(0)\}\,. \\ &= \int_{(-\varrho,\varrho)} + \int_{(-\varrho,\varrho)^c}.\end{aligned} \tag{2.32}$$

Now consider the term given by

$$I = 1/N^2 \int_{-\pi}^{\pi} \frac{(\sin(N\lambda/2))^2}{(2\sin(\lambda/2))^{-2(1+d)}}\,d\lambda.$$

Then

$$I = 1/N^2(\gamma_0(\delta) - \gamma_N(\delta)), \quad \delta = 1 + d.$$

The γ's above are the autocovariances of a FARMA$(0,\delta,0)$ process. Using Hoskings' (1981) results, we find that

$$\gamma_N \sim \gamma_0 \frac{\Gamma(-1-d)}{\Gamma(d)N^{2d+1}}. \tag{2.33}$$

Hence, in the first integral in the (2.32) if ϱ is sufficiently small we have that

$$\int_{(-\varrho,\varrho)} \leq \epsilon/2,$$

because g is bounded. In the second integral in the (2.32), every term is bounded and hence by increasing N we can make it as small as we like it.

This together with the above estimate implies that

$$\mathrm{var}(\overline{X}) = \frac{2}{\pi N^2}\int_{-\pi}^{\pi} \frac{(\sin(N\lambda/2))^2}{(2\sin(\lambda/2))^{-2(1+d)}}\{g(\lambda) - g(0)\} \to 0, \qquad \text{as} \quad N \to \infty.$$

To assess how good these asymptotic results are, we computed for different values of the sample size N and of the persistence parameter d: $N = (30, 50, 100(10), 1000)$ and d, $(d = -0.499, -0.45, -0.3, -0.2, -0.1, 0.1, 0.3, 0.45, 0.499)$ the variance of the m.l.e. of the mean, see Table 2.1; the variance of the sample mean, see Table 2.2 and the efficiency ratio between the variance of the m.l.e. of the mean and the variance of the sample mean, see table 2.3.

	The differencing parameter d									
N	-0.499	-0.45	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.45	0.499
30	1.932	1.777	1.407	1.231	1.097	0.9339	0.8971	0.8908	0.9528	0.9989
50	1.957	1.797	1.416	1.235	1.099	0.9329	0.8957	0.8894	0.9522	0.9989
100	1.977	1.812	1.423	1.239	1.100	0.9321	0.8946	0.8883	0.9518	0.9989
200	1.986	1.820	1.426	1.240	1.101	0.9318	0.8941	0.8878	0.9516	0.9989
300	1.990	1.822	1.427	1.241	1.101	0.9316	0.8939	0.8876	0.9515	0.9989
400	1.991	1.823	1.428	1.241	1.101	0.9316	0.8938	0.8875	0.9515	0.9989
500	1.992	1.824	1.428	1.241	1.102	0.9315	0.8937	0.8875	0.9514	0.9989
600	1.993	1.825	1.428	1.242	1.102	0.9315	0.8937	0.8874	0.9514	0.9989
700	1.993	1.825	1.429	1.242	1.102	0.9315	0.8937	0.8874	0.9514	0.9989
800	1.994	1.825	1.429	1.242	1.102	0.9315	0.8936	0.8874	0.9514	0.9989
900	1.994	1.826	1.429	1.242	1.102	0.9315	0.8936	0.8874	0.9514	0.9989
950	1.994	1.826	1.429	1.242	1.102	0.9315	0.8936	0.8874	0.9514	0.9989
999	1.994	1.826	1.429	1.242	1.102	0.9315	0.8936	0.8874	0.9514	0.9989

Table 2.1

Variance of the m.l.e. of the mean of a FARMA(0, d, 0) process multiplied by N^{1-2d} and divided by the autocovariance of the process where N is the sample size.

The differencing parameter d

N	-0.499	-0.45	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.45	0.499
30	2.673	2.278	1.536	1.271	1.105	0.9374	0.9069	0.9047	0.9606	0.9991
50	2.926	2.437	1.572	1.283	1.107	0.9368	0.9063	0.9043	0.9605	0.9991
100	3.269	2.641	1.610	1.293	1.110	0.9364	0.9059	0.9041	0.9604	0.9991
200	3.612	2.831	1.638	1.301	1.111	0.9362	0.9058	0.9040	0.9604	0.9991
300	3.812	2.936	1.652	1.304	1.112	0.9362	0.9058	0.9040	0.9604	0.9991
400	3.954	3.008	1.660	1.305	1.112	0.9361	0.9057	0.9040	0.9604	0.9991
500	4.064	3.063	1.666	1.306	1.112	0.9361	0.9057	0.9040	0.9604	0.9991
600	4.153	3.107	1.670	1.307	1.112	0.9361	0.9057	0.9040	0.9604	0.9991
700	4.229	3.143	1.674	1.308	1.112	0.9361	0.9057	0.9040	0.9604	0.9991
800	4.295	3.174	1.677	1.308	1.112	0.9361	0.9057	0.9040	0.9604	0.9991
900	4.353	3.201	1.679	1.309	1.112	0.9361	0.9057	0.9040	0.9604	0.9991
950	4.380	3.213	1.680	1.309	1.112	0.9361	0.9057	0.9040	0.9604	0.9991
999	4.404	3.225	1.681	1.309	1.112	0.9361	0.9057	0.9040	0.9604	0.9991

Table 2.2

Variance of the sample mean of a FARMA$(0, d, 0)$ process multiplied by N^{1-2d} and divided by the autocovariance of the process where N is the sample size.

The differencing parameter d

N	-0.499	-0.45	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.45	0.499
30	0.7228	0.7801	0.9161	0.9682	0.9935	0.9962	0.9892	0.9847	0.9919	0.9998
50	0.6689	0.7372	0.9010	0.9631	0.9926	0.9958	0.9883	0.9835	0.9914	0.9998
100	0.6046	0.6861	0.8839	0.9577	0.9917	0.9954	0.9875	0.9826	0.9910	0.9998
200	0.5500	0.6427	0.8705	0.9538	0.9911	0.9952	0.9870	0.9821	0.9908	0.9998
300	0.5220	0.6206	0.8641	0.9520	0.9909	0.9952	0.9869	0.9819	0.9907	0.9998
400	0.5037	0.6061	0.8601	0.9510	0.9907	0.9951	0.9868	0.9818	0.9907	0.9998
500	0.4903	0.5956	0.8573	0.9503	0.9906	0.9951	0.9867	0.9817	0.9907	0.9998
600	0.4798	0.5874	0.8552	0.9498	0.9906	0.9951	0.9867	0.9817	0.9906	0.9998
700	0.4713	0.5807	0.8535	0.9494	0.9905	0.9951	0.9867	0.9817	0.9906	0.9998
800	0.4642	0.5751	0.8521	0.9491	0.9905	0.9951	0.9867	0.9817	0.9906	0.9998
900	0.4581	0.5703	0.8510	0.9488	0.9905	0.9951	0.9867	0.9816	0.9906	0.9998
950	0.4553	0.5682	0.8504	0.9487	0.9904	0.9951	0.9867	0.9816	0.9906	0.9998
999	0.4528	0.5662	0.8500	0.9486	0.9904	0.9951	0.9866	0.9816	0.9906	0.9998

Table 2.3

Efficiency Ratio Between Variance of the m.l.e. mean of a FARMA$(0,d,0)$ process and the sample mean for sample size N.

The above computations have established the following theorem.

THEOREM **2.11** If $d > -1/2$, the variance of both the sample mean and of the m.l.e. mean is of order N^{2d-1}. However, when d is negative the sample mean is an inefficient estimator of the mean. If $d < -0.5$, the rate of convergence are different, and the rate of decay of the variance of the sample mean to zero is at most of order $O(N^{-2})$. However, the rate of decay of the variance of the m.l.e mean is the same as before, N^{2d-1}. When $-0.5 < d < 0$, the sample mean is inefficient. The efficiency ratio is approximately $-d$. Finally, if $d > 0$ the sample mean is an efficient estimator of the sample mean. ∇

This agrees with the common knowledge that overdifferencing can lead to inefficient estimates. Although it is difficult to give a physical meaning to antipersistence, a negative value of d can be useful from a purely fitting point of view as it has been observed that sometimes FARMA models with negative d arise for example by overdifferencing, and, therefore, it is important in these cases to estimate the mean of the process using the maximum likelihood estimate as given by the formula (2.25).

Determinant and eigenvalues of the covariance matrix

We are going to present now some results concerning the eigenvalues of the Toeplitz forms associated with fractionally differenced time series. These results are useful for presenting an asymptotic expression for the logarithm of the determinant of the autocovariance function.

First we give an expression about the logarithm of the determinant of the autocovariance matrix of the fractionally differenced process, which is an extension of a result in Grenander and Szegō (1955) to the case of long-term dependence.

THEOREM **2.12** Suppose the spectral density $f(\lambda)$ of a stationary time series process has the following form

$$f(\lambda) = (h(\lambda))^d g(\lambda), \tag{2.34}$$

where h is continuous function in the interval $[-\pi, \pi]$, with a zero at $\lambda = 0$; $h(\lambda) = o(\lambda)$, as $\lambda \to 0$; $-1/2 < d < 1/2$; g is a continuous function in $[-\pi, \pi]$, and $0 < g(x) < \infty$. Then

$$\frac{\log(D_N)}{N} \to \frac{1}{2\pi}\{d\int_{-\pi}^{\pi} \log(h(x))\,dx + \int_{-\pi}^{\pi} \log(g(x))\,dx\}, \quad \text{as} \quad N \to \infty, \tag{2.35}$$

where $D_N = \log(\det(\Sigma_N))$, and Σ_N is the autocovariance matrix of order N associated with f. ▽

<u>Proof</u> To prove the theorem, we first recall some results about the determinant D_N. The one-step ahead error σ_N^2 is given in terms of the determinant D_N by (Grenander and Szegō, 1955 p. 38)

$$\sigma_N^2 = \frac{D_N}{D_{N-1}} \tag{a}$$

and

$$\lim_{N\to\infty} \sigma_N^2 = \exp\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log(f(x))\, dx\}. \qquad \text{(b)}$$

Then using (a) and (b) we obtain that

$$\frac{\log(D_N)}{N} \to \frac{1}{2\pi}\{d\int_{-\pi}^{\pi} \log(h(x))\, dx + \int_{-\pi}^{\pi} \log(g(x))\, dx\}.$$

□

COROLLARY **2.13** For the case of a FARMA(p, d, q) process:

$$\frac{\log(D_N)}{N} \to \log(\gamma_0) + \frac{d}{2} + \frac{1}{2\pi}\{\int_{-\pi}^{\pi} \log(g(x))\, dx\}, \qquad (2.36)$$

where g corresponds to the spectral density of the ARMA(p, q) filter of the process. The third term in (2.36) is the limit of an ARMA(p, q) process. ▽

REMARK **2.14** In order to obtain an approximate maximum likelihood estimation (m.l.e.) method based on a backforecasting procedure proposed by Box and Jenkins (1976), McLeod (1977) gave, based on Finch (1960), an approximation of order $1/N$ of the determinant of the covariance matrix divided by N. From a computational point of view, this approximation may be obtained very efficiently. Hence, it is of some interest to investigate if a similar approximation is still valid for FARMA processes.

The result of McLeod (1977) is based on the paper by Finch (1960) mentioned above which turn is based on the following result by Grenander and Szegö (1955).

THEOREM **2.15** (Grenander and Szegõ, 1955). Assume that the spectral density function $f(\lambda)$ is positive and continuous in $[-\pi, \pi]$. Also, assume that $f'(\lambda)$ is Lipshitzian. Then

$$\lim_{N\to\infty} \frac{D_N(f)}{G(f)^N} = \exp\left\{\frac{1}{\pi}\int \left|\frac{g'(z)}{g(z)}\right|^2 d\sigma\right\}, \tag{2.37}$$

where the integration is performed on the unit complex disk $\{z, |z| \leq 1\}$; $G(f) = g(0)$ is the geometric mean of f or the variance, σ^2, of the white noise process that generates the time series; and g is determined uniquely by

$$g(0)^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\lambda)\, d\lambda\right\},$$

$$|g(e^{i\lambda})|^{-2} = f(\lambda).$$

However, for FARMA time series the assumptions of this theorem are not valid because the term $\sin(\lambda/2)^{-2d-1}$ is not Lipshitzian when d is positive and f is nonpositive if d is negative.

In order to investigate an analogous result for the FARMA process, first note that Theorem (2.15) is best seen as the term of order $1/N$ of the limit of the logarithm of the determinant D_N divided by N:

$$\frac{\log D_N(f)}{N} = G(f) + \frac{1}{N}\left\{\frac{1}{\pi}\int \left|\frac{g'(z)}{g(z)}\right|^2 d\sigma\right\}.$$

Now, it is well known that

$$\frac{\log D_N(f)}{N} = \sum_{k=0}^{N-1}(1 - \frac{k}{N}) \log(1 - \phi_{k,k}^2).$$

For the case of a FARMA$(0,d,0)$ process the $\phi_{k,k}$ are known (Hoskings, 1981), and equal to

$$\phi_{k,k} = \frac{d}{k-d}.$$

Call $\alpha = \lim_{M\to\infty} \log D_M/M$. Then, the difference between the limit value and $\log D_N/N$ is equal to

$$\frac{\log D_N}{N} - \alpha = \lim_{M\to\infty} \sum_{k\geq N} (1 - \frac{k}{M}) \log(1 - \phi^2_{k,k}).$$

But, then

$$\begin{aligned}\frac{\log D_N}{N} - \alpha &= \lim_{M\to\infty} \sum_{k\geq N} (1 - \frac{k}{M}) \phi^2_{k,k}. \\ &= \lim_{M\to\infty} \sum_{k\geq N} (1 - \frac{k}{M}) \frac{d^2}{M^2(\frac{k}{M} - \beta_M)^2},\end{aligned}$$

where $\beta_M = \frac{d}{M}$. Thus,

$$\begin{aligned}\frac{\log D_N}{N} - \alpha &= \lim_{M\to\infty} \frac{d^2}{M} \int_{N/M}^{1} \frac{1-z}{(z-\beta_M)^2}\, dz. \\ &= \lim_{M\to\infty} \frac{d^2}{M} [\frac{1-\beta_M}{z-\beta_M} - \log(z-\beta_M)]^1_{N/M}. \\ &= -\frac{d^2}{N-d} \\ &= -d\phi_{N,N}.\end{aligned}$$

PROPOSITION 2.16 For the case of a FARMA$(0,d,0)$ process

$$\frac{\log D_N}{N} - \alpha = -\frac{d^2}{N-d}.$$

▽

Now, we consider the general FARMA(p,d,q) case. It is known that in case of an autoregressive process of order p, the expression (2.37) is valid, not only in the limit, but for all $N \geq p$,

$$\frac{D_N(f)}{G(f)^{N+1}} = \exp\left\{\frac{1}{\pi}\int |\frac{g'(z)}{g(z)}|^2\, d\sigma\right\}.$$

But, any covariance matrix of order N may be reproduced exactly by an autoregression of order N. Hence, for any stationary time series process

$$\frac{D_N(f)}{G(f)^{N+1}} = \exp\left\{\frac{1}{\pi}\int |\frac{g_N'(z)}{g_N(z)}|^2\, d\sigma\right\}. \tag{2.38}$$

Now, in case of a FARMA time series

$$\Phi(B)\nabla^d(B)X_t = \Theta(B)a_t,$$

we could use an approximation

$$\Phi(B)\phi_N X_t = \Theta(B)a_t,$$

of

$$\Phi(B)\nabla^d(B)X_t = \Theta(B)a_t.$$

Then we could approximate g_N by $\Phi(z)\phi(z)/\theta(z)$. But

$$\frac{g_N'(z)}{g_N(z)} = (\log g(z))',$$

and hence multiplicative factors enter the expression additively. Thus, the approximate long term differencing part factorizes from the ARMA part of $g_N(z)$. Hence, the term $\log D_N/N$ factorizes approximately into the corresponding quantity for a FARMA$(0, d, 0)$, for the ARMA(p, q) part and cross-products between the FARMA$(0, d, 0)$ and the ARMA(p, q) parts. The error committed by the FARMA$(0, d, 0)$ part is given in the preceding proposition. For the ARMA(p, q) expression we follow McLeod (1977), and finally, we neglect the cross products.

Now we need the definition of equally distributed sequences of H. Weyl (Polya and Szegõ,1945, p. 87 or Grenander and Szegõ, 1955, chapter 5). The definition is as follows:

DEFINITION **2.17** Two sequences uniformly bounded (by K) $\{x_\nu^i\}$ and $\{y_\nu^i\}$, where $\nu \in \mathbb{N}$ and $i = 1, \ldots, \nu$ are equally distributed in $[-K, K]$ if for any continuous function F in $[-K, K]$

$$\lim_{n \to \infty} \frac{\sum_{\nu=1}^{n+1} [F(x_\nu^i) - F(y_\nu^i)]}{n} = 0. \tag{a}$$

∇

REMARK **2.18** There are some families of functions such that if (a) in Definition 2.17 is satisfied for each of the functions in the family, then (a) is true for any continuous function. Two examples of these families are $F = \{h : h(x) = x^s\}$, and $F = \{h : h(x) = \log(1 + zx)\}$ (see Grenander and Szegõ, 1955). We will use the family of the log functions to prove equal distribution for the eigenvalues of the autocovariances of FARMA(p, d, q) processes.

THEOREM **2.19** **Assume that f is a spectral density function that satisfies the conditions of theorem (2.4). Then, there are two finite numbers α and**

β such that for any function that is continuous in the interval $[\alpha, \beta]$

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} F(\hat{\lambda}_i^n)}{n} = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(f(x))\, dx,$$

where

$$\hat{\lambda}_i^n = \frac{\lambda_i^n}{\max(1, n^{2d})}.$$

∇

Proof First, because of the long-term memory, the eigenvalues of Σ_n are of order $\min(1, n^2 d)$. Then, the standardized eigenvalues lie in some interval $[\alpha, \beta]$, and because of theorem (2.4) the result follows.

□

COROLLARY **2.20** (a) Let $F(\lambda) = \lambda^s$; then

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} (\lambda_i^n)^s}{n} = \frac{1}{2\pi} \{ \int_{-\pi}^{\pi} [f(x)]^s\, dx.$$

(b) $\lambda_\nu^n = O(\max(1, n^2 d)\alpha)$, $\lambda_{n+1-\nu}^n = O(\max(1, n^2 d)\beta)$, as $n \to \infty$.

∇

Proof This corollary can be proven extending Grenander and Szegö's (1955 p. 66) method for the case of a short memory process. However, because of the long term memory we have to substitute their limits by asymptotic expressions for the eigenvalues.

□

REMARK **2.21** When $d < 0$, the spectral density has a zero at zero. This means that the filter is a high-pass filter i.e. smooths out low frequencies, and that produces the nonefficient behaviour of the mean. Another consequence is that the lower order eigenvalues tend to zero. This implies that the autocovariance matrix tends to a noninvertible matrix. In fact using a theorem in Grenander and Szegö (1955, p. 72), the smallest eigenvalue is of order $O(1/N^2)$, and hence the condition number (where condition number (Golub and van Loan, 1983) is defined as the ratio between the smallest and the largest eigenvalue of the matrix, and it is a measure of the amplification of errors involved in the computation of quantities as the inverse of the matrix) of the autocovariance matrix is of order N^2, therefore, the autocovariance matrix for a FARMA process may be very badly conditioned. Hence, computational methods that use this inverse matrix will exhibit numerical problems. The same is true if instead of using the inverse matrix we use an approximation to it as in Taqqu and Fox (1986).

REMARK **2.22** The eigenvalues of the autocovariance matrix have an important application in that they set limits on the efficiency of ordinary least squares estimation (O.L.S.) versus the maximum likelihood method in the estimation of regression parameters with time dependent noise, through the Kantorovitch inequality. Specifically, if

$$y_t = p_t\beta + x_t.$$

Where $\{x_t\}$ is a time series process, $\{p_t\}$ is a process independent of x_t and

β is a constant parameter. The the O.L.S. regression estimate of β is given by

$$\bar{\beta} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{Y}.$$

Where $\mathbf{P}$ is the matrix $(p_t), t = 1, \ldots, N$, $\mathbf{Y}$ is the column vector of $\{y_t\}, t = 1, \ldots, N$ observations, and N is the number of observations. The minimum unbiased estimate is

$$\hat{\beta} = (\mathbf{P}'\Sigma^{-1}\mathbf{P})^{-1}\mathbf{P}'\Sigma^{-1}\mathbf{Y}.$$

Then, the ratio of the variances when β a one dimensional constant is given by

$$\mathrm{var}(\bar{\beta})\mathrm{var}(\hat{\beta})^{-1} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\Sigma\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}(\mathbf{P}'\Sigma^{-1}\mathbf{P}).$$

This ratio can be bounded, using the Kantorovitch inequality, when $\mathbf{P}$ is a one dimensional vector as

$$1 \geq \mathrm{var}(\hat{\beta})\mathrm{var}(\bar{\beta})^{-1} \geq \frac{4\lambda_{min}\lambda_{max}}{(\lambda_{min} + \lambda_{max})^2},$$

where λ_{min} is the minimum eigenvalue and λ_{max} is the maximum eigenvalue of the matrix Σ such that $\mathbf{P}$ is contained in the subspace generated by the eigenvectors of Σ associated to the eigenvalues λ_i of Σ, i from *min* to *max*. This result has an application to fractional times series in that if the differencing parameter d is negative then the minimum eigenvalues of Σ will tend to zero, and the maximum eigenvalues will tend to a constant. Hence, most of the regression sequences p_t will give nonefficient O.L.S. estimators. If d is positive, the minimum eigenvalues will tend to a constant and the maximum eigenvalues will tend to ∞, Therefore, most of the regression sequences will be efficient.

For the case when $\mathbf{P}$ is a matrix we can give the following bound to the matrix $\text{var}(\hat{\beta})\text{var}(\underline{\beta})^{-1}$.

THEOREM **2.23** The positive definite matrix $\text{var}(\hat{\beta})\text{var}(\underline{\beta})^{-1}$ is bounded above by the identity matrix and below by a diagonal matrix, Δ, with entries $4\lambda_i\lambda_M/(\lambda_i - \lambda_M)^2$, where λ_i are the eigenvalues of the covariance matrix Σ ordered in ascending order (the bounds on the matrix $\text{var}(\hat{\beta})\text{var}(\underline{\beta})^{-1}$ are understood in the ordering of the positive definite matrices). ∇

Proof We can assume that the columns of the matrix $\mathbf{P}$ are orthonormal, $\mathbf{P}'\mathbf{P} = I_M$. Also, we can diagonalize the covariance matrix $\Sigma = Q'DQ$, Q as an orthogonal $N \times N$ matrix. Then, if we define $\Omega := Q\mathbf{P}$,

$$\text{var}(\hat{\beta})\text{var}(\underline{\beta})^{-1} = \Omega' D^{-1}\Omega\Omega' D\Omega.$$

Hence, the problem reduces to find Ω such that

$$\Omega' D^{-1}\Omega\Omega' D\Omega$$

is a minimum in the ordering of the positive definite matrices. But this problem is equivalent to finding M orthonormal vectors ω_i, such that they solve the maximization problem

$$\omega_i' D-1\omega_i\omega_i' D\omega_i \doteq max!$$

$$\text{such that} \qquad \omega_i \perp \omega_j, \quad j < i, \|\omega_i\| = 1.$$

But from the Kantorovitch inequality, $\omega_i = \lambda_i - \lambda_M$ and the theorem follows.

Evaluation of $\hat{\mu}$

The evaluation of the maximum likelihood estimator of the mean can be performed efficiently using either Cholesky decomposition of the inverse of Σ given by the partial linear regression coefficients, $\{\phi_{i,t}\}$, which can be obtained easily by the Levison-Durbin algorithm (Durbin, 1960), or by the Trench algorithm for the inverse of a Toeplitz matrix (Trench, 1964). For the particular case of fractionally differenced noise, Hosking (1981) gave a closed expression for the reflection coefficients or partial linear regression coefficients Hence, in this case a closed expression for the maximum likelihood estimate of the mean is known. In terms of the partial linear regression coefficients the following expression could be used to evaluate the maximum likelihood estimate $\hat{\mu}$

$$\hat{\mu} = \frac{\sum_0^{N-1}(x_t - \phi_{1,t}x_{t-1} - .. - \phi_{t,t}x_0)}{\sum(1 - \phi_{1,t} - .. - \phi_{t,t})^2}. \tag{2.39}$$

Parameter estimation

There are two methods available to estimate the remaining parameters in the time domain: exact maximum likelihood estimation or an approximation of the filter ∇^d. Most of the maximum likelihood estimation algorithms depend on computing the one step ahead prediction errors, e_t, which can be computed in terms of the partial linear regression coefficients. These coefficients can be computed efficiently by the Durbin-Levinson algorithm. Finally, with these values of e_t the estimates of the parameters are obtained by minimizing the modified sum of squares function given by:

$$\log l = \sum (N - t + 1) \log(1 - \phi_{t,t}^2) + \sum e_t^2. \tag{2.40}$$

Although the computation of estimates by maximum likelihood is statistically attractive, the amount of computations involved in the above scheme makes algorithms having fewer numbers of computations competitive alternatives.

The algorithm proposed by Li and McLeod (1986) is computationally economical. The algorithm consists of approximating the filter ∇^d by the filter ∇_M^d where ∇_M^d is defined as the filter resulting by taking the first M terms of the filter ∇^d, i.e. by approximating the process by a 'long' autoregression. Then the algorithm minimizes the sum of the squared residuals, where the residuals are obtained as the output of the filters ∇_M^d and the ARMA filter. To compute the residuals, an algorithm such as the one given by McLeod and Sales (1983) could be used. Also, as recommended in Box and Jenkins (1976) the sum of squared residuals could be extended back in time by backforecasting. Note that the approximation of ∇^d by ∇_M^d is not the optimal approximation in a least squares sense. However, since the order of M is comparable with N it has to be very close to the optimal approximation. The order of approximation necessary to obtain consistent estimates can be shown to be of the order of $N^{1/2}$ and an ad hoc rule is to fit time series with at least 50 observations. The order of truncation M is chosen as a number between $N/4$ and $N^{1/2}$, by trying to balance the degree of approximation of the filter ∇_M^d to the filter ∇^d and the amount of computations involved. However, for N close to 50, M is taken as half the number of observations. The amount of computations using this algorithm is much smaller than that for the maximum likelihood approach. Moreover, estimates obtained in this form are asymptotically equivalent to the maximum likelihood estimates and it seems that the finite sample estimates are generally close enough to the maximum likelihood estimates.

N	MDS	MMSE for L&M	MMSE for exact m.l.e	MBias for L&M	MBias for exact m.l.e
33	0.0042	0.0524	0.0405	0.1279	0.1076
50	0.0019	0.0295	0.0253	0.0859	0.0796
75	0.0010	0.0148	0.0144	0.0510	0.0501

Table 2.4

N=number of observations.
L&M=Li and Mcleod (1986) approximate maximum likelihood algorithm.
MDS=Mean difference between L&M algorithm and exact m.l.e.
MMSE=Mean square error, assuming an uniform density for d on (−05 ,05)
MBias=Mean Bias.

To compare the performance of the Li and McLeod algorithm and the m.l.e algorithm, we used both algorithms to estimate the differencing parameter d of a FARMA$(0, d, 0)$ model, over an uniform grid of values where $-1/2 < d < 1/2$. As a measure of the performance of the Li and McLeod algorithm, we used the squared difference of the estimated parameter by both algorithms. The results are given in Table 2.4. It can be seen that Li and McLeod algorithm produces estimates very close to the m.l.e. estimates even for relatively small sample sizes.

Li and McLeod (1986) studied the asymptotic distributions of the estimates when the mean of the time series is known. They derived closed form expressions for the variances of the asymptotically normal distributions of the estimates. Estimation of the parameter d has been considered by several authors (Geweke and Porter-Hudak, 1983; Milhøl, 1984; Li and McLeod, 1985; Fox and Taqqu 1986). However, they assumed that the mean of the process, μ, is known. The standard argument to show that the unknown mean does not affect the normality of the parameter estimates depends on a rate of convergence of the estimate of the mean of the order of $N^{1/2}$. Hence, it is not applicable to fractionally differenced time series with positive parameter d. It can be demonstrated that the estimation of the mean does not affect the above asymptotic results. However, these results are not likely to hold for a finite sample size because of the long term persistence and the parameter d is constrained to lie in the open interval $(-1/2, 1/2)$. In practice, the interval is closed and it can be observed using simulation that if the persistence parameter is close to 1/2 there is a high probability for the estimate of d to be equal to 1/2. A similar phenomenon was observed by Cryer and Ledolter (1981) for the ARMA(0, 1) model. Hence, the rate of convergence of the estimates depends on the parameters even for relatively large sample sizes of more than 200. Additionally, it should be noted that above method is very similar to fitting an autoregressive process of order one if d is not close to 1/2, say less than 0.3.

A modification of Li and McLeod algorithm

In terms of its statistical properties it would be better to use exact maximum likelihood instead of a approximation to it. However, we can give an approximation to maximum likelihood that in practice is virtually indistinguishable from maximum likelihood estimation.

To obtain the likelihood, we could use the partial regression coefficients $\{\phi_{t,k}\}$. However, we know that: (a) as $t \to \infty$ then $\phi_{t,k} \to \pi_k$, where π_k is the coefficient associated with the k-th power of B in $\Pi(B) = (\Phi(B)/\Theta(B))\nabla^d(B)$; (b) as $k \to \infty$ then $\phi_{t,k} \to 0$. Then we expect that after some t_0,

$$e_{t,M} \approx e_t.$$

The problem is how to compute e_t, for $t < t_0$. This can be done by brute force. That is, obtain the first $t_0 - 1$ autocovariances of the process, and then use the Levinson-Durbin algorithm to obtain the $\{\phi_{t,k}\}$.

REMARK 2.24 (a) To specify the above algorithm completely, it remains to give some rules on how to choose, t_0, and M. This can be done in the following manner. Compute e_t by the Levinson-Durbin algorithm and $e_{t,M}$ using the approximation to the fractional differencing filter, $M = t$. When $e_t \approx e_{t,M}$, fix M. If for sufficiently large t, the difference between e_t and $e_{t,M}$ is small, then set $t = t_0$ and use the approximate maximum likelihood algorithm to compute e_t for future values of t. It is not necessary to find this value of t for d' if we have already found it for d and d is close to d'. This algorithm will produce a likelihood function virtually indistinguishable from the exact likelihood.

(b) The Li and McLeod algorithm will be also very similar to this algorithm, because the backforecasting procedure will tend to eliminate initial conditions.

Asymptotic behaviour

We are going to show that the estimates of $\underline{\beta} = (d; \underline{\phi}; \underline{\theta})$ are in the domain of attraction of the Gaussian probability. This result has been proven by Li and McLeod (1986) when the mean is known and the approximation ∇_M^d is used in place of ∇^d.
First, we need to eliminate the unknown mean from the log-likelihood function $l(\mu, \sigma^2; \underline{\beta})$.

To simplify the notation, let $\underline{\beta}_0$ be the true parameter,

$\delta\mu = \mu_0 - \hat{\mu}$, and $\underline{y} = \underline{x} - \mu_0\mathbf{1}$, the observations minus the true mean. Then, we have the following result.

LEMMA **2.25** (a) The log-likelihood function evaluated at $\hat{\mu}$ is given by

$$l_{|\hat{\mu}} = \underline{y}'\frac{\Sigma^{-1}}{\sigma^2}\underline{y} - (\delta\mu)^2\mathbf{1}'\Sigma^{-1}\mathbf{1} + \log(\det(\sigma^2\Sigma)). \tag{2.41}$$

(b) The expectation of the second term in (2.41) is given by

$$\mathbb{E}\{(\delta\mu)^2\mathbf{1}'\Sigma^{-1}\mathbf{1}\} = \frac{\mathbf{1}'\Sigma^{-1}\Sigma_0\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}. \tag{2.42}$$

▽

<u>Proof</u> To prove (a) first note that

$$\delta\mu = \frac{\mathbf{1}'\Sigma^{-1}y}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}, \tag{a}$$

then substituting (a) into the log-likelihood

$$l_{|\hat{\mu}} = \underline{y}'\frac{\Sigma^{-1}}{\sigma^2}\underline{y} - 2\delta\mu\mathbf{1}'\Sigma^{-1}\underline{y} - (\delta\mu)^2\mathbf{1}'\frac{\Sigma^{-1}}{\sigma^2}\mathbf{1} + \log(\det(\sigma^2\Sigma))$$

we obtain the result.

To obtain (b), just take the expectation.

□

REMARK **2.26** Part (b) implies that when $\Sigma = \Sigma_0$ then $\mathbb{E}(\delta\mu)^2\mathbf{1}'\Sigma^{-1}\mathbf{1} = 1$. That is, this term is $O_p(1)$, while the other terms in the log-likelihood are of order $O_p(\sqrt{N})$. Hence, we can eliminate this term for the purpose of obtaining the asymptotic behaviour of the estimates.

The concentrated log-likelihood $l|_{\hat{\mu},\hat{\sigma}^2}$ is given by

$$l|_{\hat{\mu},\hat{\sigma}^2} = \log(\hat{\sigma}^2) + \log(\det(\Sigma)) + \text{const}.$$

Hence, it is asymptotically equivalent to minimizing the the sum of squares of the innovations. We have the following theorem

THEOREM **2.27** The asymptotic distribution of the m.l.e. estimates of $\underline{\beta}$ is given by

$$\sqrt{N\mathcal{I}}(\hat{\underline{\beta}} - \underline{\beta}_0) \rightarrow \mathcal{N}(0, I), \qquad \text{in law}.$$

Where $\mathcal{I}$ corresponds to the information matrix of the observations, and the $\sqrt{\mathcal{I}}$ is a matrix square root obtained by Cholesky decomposition. ∇

Proof The proof follows standard procedures. It consists of using a martingale central limit theorem, after we have shown that the $\hat{\underline{\beta}} - \underline{\beta}_0$ may be written as a sum of martingale differences (Hall and Heyde, 1980). $\square$

REMARK **2.28** Li and McLeod (1986) found the inverse of the information matrix, $\mathcal{I}^{-1}$, by using their approximation to the differencing operator. This matrix is the same as in the above theorem.

An investigation of the finite sample properties of the m.l.e. algorithm

Even if maximum likelihood estimation produces estimates that are asymptotically efficient and, hence, in a sense optimal, the finite properties of m.l. estimators maybe inferior to those of some other estimators. A related discussion was given by Barnard et. al. (1962) and investigated in a time series context by Copas (1966). In summary, they proposed that instead of using m.l.e. for estimating the parameter θ one should use an average of the parameter θ with respect to the likelihood function.

$$\hat{\theta} = \frac{\int \theta L(\theta)\, d\theta}{\int L(\theta)\, d\theta}.$$

where θ is the parameter that we want to estimate and $L(\theta)$ is the likelihood of θ. In many situations this estimate has better finite sample properties than the m.l. estimator. As can be seen, this amounts to the use of a Bayesian estimate with quadratic loss function and an uniform a priori density on θ. Copas (1966), used this procedure to estimate the regression parameter of an Gaussian autoregressive process of order one with known variance. He found that the average likelihood estimator was superior to the m.l. estimator for small sample sizes. Hence, because in long memory models almost every sample may be considered small, it maybe interesting to compare the performances of m.l.e and the average likelihood estimates.

We applied the average likelihood procedure to estimate the unknown differencing parameter d in a FARMA$(0, d, 0)$ time series. However we used the two $L(\theta)$ functions given by

$$L_1 \approx e^{ss/2}$$

and

$$L_2 \approx (1 + ss)^{-N/2},$$

where ss denotes the innovation sum of squares. L_1 assumes that the variance is known and equal to one. While L_2 assumes that the variance is unknown. For the variance in L_2 we assume a diffuse a priori density; and integrate the variance out of the joint likelihood of the parameters. The results can be seen in Tables 2.5–2.8, where the MSE and the BIAS of the different estimates are compared for series of sizes 33,50,75,100, over an uniform grid in the interval $-1/2 < d < 1/2$. The tables show that if the parameter does not lie too close to the boundaries it is better to use the average likelihood procedure using L_2. Also, assuming an uniform distribution on the parameter d, we computed the overall MSE with the similar results.

d	M.S.E $\hat{d}$	M.S.E $\hat{d}_1$	M.S.E $\hat{d}_2$	Bias $\hat{d}$	Bias $\hat{d}_1$	Bias $\hat{d}_2$
-0.4900	0.0042	0.0097	0.0179	-0.0096	-0.0818	-0.1211
-0.4400	0.0060	0.0053	0.0110	0.0362	-0.0415	-0.0837
-0.3900	0.0124	0.0038	0.0062	0.0760	-0.0041	-0.0500
-0.3400	0.0200	0.0069	0.0065	0.0891	0.0177	-0.0253
-0.2900	0.0374	0.0126	0.0070	0.1633	0.0755	0.0281
-0.2400	0.0486	0.0212	0.0133	0.1636	0.0881	0.0439
-0.1900	0.0552	0.0296	0.0206	0.1290	0.0694	0.0325
-0.1400	0.0569	0.0342	0.0241	0.1201	0.0782	0.0471
-0.0900	0.0747	0.0464	0.0326	0.1724	0.1262	0.0943
-0.0400	0.0995	0.0618	0.0427	0.2557	0.2031	0.1647
0.0100	0.0680	0.0477	0.0358	0.1425	0.1160	0.0970
0.0600	0.1095	0.0743	0.0556	0.1910	0.1529	0.1321
0.1100	0.0948	0.0707	0.0560	0.1680	0.1445	0.1322
0.1600	0.0550	0.0458	0.0381	0.0873	0.0891	0.0901
0.2100	0.0426	0.0374	0.0329	0.0908	0.0956	0.1012
0.2600	0.0495	0.0431	0.0396	0.1041	0.1132	0.1225
0.3100	0.0529	0.0450	0.0410	0.1068	0.1182	0.1305
0.3600	0.0648	0.0592	0.0579	0.1451	0.1646	0.1778
0.4100	0.0722	0.0680	0.0682	0.1505	0.1807	0.1964
0.4600	0.0524	0.0523	0.0562	0.1418	0.1765	0.1948

Figure 2.5

M.S.E. and bias of the estimators of the differencing parameter d for a FARMA$(0, d, 0)$ model using m.l. estimation and Bayesian with loss functions L_1, L_2. The number of observations is 33.

d	M.S.E $\hat d$	M.S.E. $\hat d_1$	Bias $\hat d$	Bias $\hat d_1$
-0.49	0.343e-02	0.719e-02	0.198e 00	-0.733e-01
-0.44	0.703e-02	0 651e-02	0.162e 00	-0.475e-01
-0 39	0.848e-02	0.277e-02	0.130e 00	0.569e-02
-0.34	0.139e-01	0.640e-02	0.101e 00	0.144e-01
-0.29	0.253e-01	0.133e-01	0.753e-01	0.499e-01
-0.24	0.280e-01	0.163e-01	0.530e-01	0.745e-01
-0.19	0.310e-01	0.217e-01	0.346e-01	0.941e-01
-0.14	0.241e-01	0.178e-01	0.193e-01	0.667e-01
-0.09	0.437e-01	0.332e-01	0.951e-02	0.857e-01
-0 04	0.320e-01	0.262e-01	0.365e-02	0.962e-01
0.01	0.326e-01	0.261e-01	0.150e-02	0.565e-01
0.06	0.362e-01	0.318e-01	0.544e-02	0.960e-01
0.11	0.460e-01	0.417e-01	0.133e-01	0.919e-01
0.16	0.252e-01	0.238e-01	0.257e-01	0.770e-01
0.21	0.268e-01	0.243e-01	0.433e-01	0.796e-01
0.26	0.241e-01	0.209e-01	0.648e-01	0.379e-01
0 31	0.370e-01	0.343e-01	0.941e-01	0.126e 00
0 36	0.213e-01	0.176e-01	0.125e 00	0.817e-01
0 41	0.446e-01	0.413e-01	0.163e 00	0.129e 00
0.46	0.354e-01	0.350e-01	0.206e 00	0.125e 00
0 49	0.259e-01	0.285e-01	0.233e 00	0.133e 00

Figure 2.6

M.S.E. and bias of the estimators of the differencing parameter d for a FARMA$(0, d, 0)$ model using m.l. estimation and Bayesian with loss functions L_1, L_2. The number of observations is 50.

d	M.S.E $\hat{d}$	M.S.E. $\hat{d}_1$	M.S.E. $\hat{d}_2$	Bias $\hat{d}$	Bias $\hat{d}_1$	Bias $\hat{d}_2$
-0.49	0.00322	0.18400	0.01720	-0.02440	-0.42900	-0.12500
-0.44	0.00510	0.14800	0.01110	0.01100	-0.38500	-0.09490
-0.39	0.00661	0.11500	0.00673	0.02110	-0.33900	-0.06540
-0.34	0.00908	0.08420	0.00303	0.05550	-0.29000	-0.02610
-0.29	0.01500	0.06040	0.00532	0.06300	-0.24500	-0.00824
-0.24	0.01340	0.04170	0.00577	0.05400	-0.20400	-0.00179
-0.19	0.01950	0.02470	0.00815	0.08310	-0.15600	0.03180
-0.14	0.01550	0.01520	0.01040	0.03990	-0.12100	0.00358
-0.09	0.02190	0.00590	0.01330	0.08120	-0.07380	0.04270
-0.04	0.02520	0.00132	0.01570	0.09650	-0.02970	0.06210
0.01	0.01480	0.00056	0.01160	0.04210	0.00175	0.01750
0.06	0.01760	0.00263	0.01360	0.06600	0.04150	0.04240
0.11	0.01160	0.00629	0.00822	0.03870	0.07330	0.02090
0.16	0.02160	0.01620	0.01660	0.07780	0.12100	0.06220
0.21	0.02530	0.02520	0.01970	0.08180	0.15300	0.06960
0.26	0.01250	0.03270	0.00937	0.04480	0.17600	0.04330
0.31	0.01110	0.04890	0.00878	0.05660	0.21800	0.06360
0.36	0.01330	0.06520	0.01140	0.05610	0.25200	0.07910
0.41	0.01310	0.08260	0.01410	0.05840	0.28400	0.09680
0.46	0.01120	0.09920	0.01600	0.05220	0.31200	0.11200
0.49	0.00959	0.11700	0.01930	0.05540	0.33900	0.12900

Table 2.7a

M.S.E. and bias of the estimators of the differencing parameter d for a FARMA$(0, d, 0)$ model using m.l. estimation and Bayesian with loss functions L_1, L_2. The number of observations is 75.

	d	$\hat{d}_1$	$\hat{d}_2$
M.S E	0.0141	0.0560	0.0117
Bias	0.0529	-0.0143	0.0264

Table 2.7b

Average M.S.E. and average bias of the estimators of the differencing parameter d for a FARMA(0,d,0) model using m.l. estimation and Bayesian with loss functions L_1, L_2. The number of observations is 75.

d	M.S.E $\hat{d}$	M.S.E $\hat{d}_1$	M.S.E. $\hat{d}_2$	Bias $\hat{d}$	Bias $\hat{d}_1$	Bias $\hat{d}_2$
-0.49	0.00287	0.18500	0.01250	-0.02330	-0.43000	-0.10700
-0.44	0.00291	0.14700	0.00688	-0.00058	-0.38300	-0.07590
-0.39	0.00539	0.11400	0.00559	0.00035	-0.33800	-0.05660
-0.34	0.00885	0.08550	0.00454	0.02750	-0.29200	-0.02760
-0.29	0.01110	0.06070	0.00457	0.05040	-0.24600	0.00135
-0.24	0.01150	0.04220	0.00725	0.03100	-0.20500	-0.00458
-0.19	0.00934	0.02630	0.00629	0.04090	-0.16100	0.01180
-0.14	0.01470	0.01410	0.00981	0.05810	-0.11700	0.03030
-0.09	0.00774	0.00651	0.00595	0.03870	-0.07930	0.01560
-0.04	0.00884	0.00204	0.00728	0.04230	-0.04050	0.02100
0.01	0.01160	0.00078	0.00977	0.03970	-0.00359	0.01970
0.06	0.00942	0.00191	0.00784	0.04490	0.03400	0.02610
0.11	0.01070	0.00543	0.00917	0.03120	0.06630	0.01480
0.16	0.00831	0.01110	0.00692	0.03110	0.10100	0.01620
0.21	0.01030	0.01950	0.00783	0.03570	0.13500	0.02550
0.26	0.00909	0.02980	0.00640	0.03050	0.16900	0.02780
0.31	0.00850	0.04320	0.00657	0.04220	0.20400	0.04460
0.36	0.00978	0.05830	0.00796	0.04780	0.23800	0.06210
0.41	0.01150	0.07940	0.01100	0.04710	0.27900	0.08130
0.46	0.00861	0.09270	0.01200	0.04390	0.30200	0.09630
0.49	0.01000	0.11100	0.01660	0.06670	0.33200	0.12000

Table 2.8a

M.S.E. and bias of the estimators of the differencing parameter d for a FARMA(0, d, 0) model using m.l. estimation and Bayesian with loss functions L_1, L_2. The number of observations is 100.

	d	$\hat{d}_1$	$\hat{d}_2$
M.S.E.	0.0091	0.0541	0.0082
Bias	0.0346	-0.0207	0.0163

Table 2.8b

Average M.S.E. and average bias of the estimators of the differencing parameter d for a FARMA$(0,d,0)$ model using m.l. estimation and Bayesian with loss functions L_1, L_2. The number of observations is 100.

2.5 A multivariate generalization

We are interested in generalizing fractionally differenced models to the multivariate case. More specifically, we want to extend the class of contemporaneous models to models that exhibit persistence and are related through the present time values. We present two generalizations.

First generalization to contemporaneous processes

We are going to assume in what follows that the means are known and equal to zero.

DEFINITION **2.29** We assume that the innovations given by (2.5) are contemporaneously correlated. That is, the model is as follows. The k processes $\{x_{t,i}\}, i = 1, \ldots, k$ are generated as in (2.5), but the multivariate innovations $\underline{e} = \{e_{t,i}\}, i = 1, \ldots, k$ are such that $\underline{e}_t \sim \mathcal{N}(O, \Delta_t)$; the vectors $\underline{e}_t$'s are independent; where $\Delta_t = \Phi_t \Delta \Phi_t$, $\Phi_t = \mathrm{diag}(\sigma_{t,1}, \ldots, \sigma_{t,k})$, the $\sigma_{t,i}$'s are the variances of the innovations (2.5). ▽

The likelihood function

The log-likelihood l is given in terms of the innovations by

$$l = \sum_t \underline{e}_t \Delta_t^{-1} \underline{e}_t' + \sum_k \sum_t \log(\sigma_{t,k}^2) + N \log(\det(\Delta)). \tag{2.43}$$

The m.l.e. of Δ is given by

$$\hat{\Delta} = \frac{1}{N} \sum_t \underline{e}_t \underline{e}_t',$$

that is the sum of cross-products of the innovations $\underline{e}_t$. The concentrated log-likelihood function with respect to Δ is written as

$$\hat{l} = l|_{\hat{\Delta}} = \sum_k \sum_t \log(\sigma_{t,k}^2) + N \log(\det(\hat{\Delta})) + \text{const.} \tag{2.44}$$

A m.l.e. algorithm

An algorithm can be obtained if we realize that the concentrated log-likelihood $\hat{l}$ just contains the univariate innovations $\{e_{t,i}\}$. Hence, if we couple the univariate minimization algorithms for each of the k time series we can compute the concentrated log-likelihood.and, using a minimization algorithm, estimate the parameters. Also, as we are using the univariate algorithms, an easy choice of the starting values for the multivariate algorithm could be the univariate m.l.e. estimates.

Asymptotic theory

As for the univariate model, asymptotically we are only concerned with the minimization of $\det(\hat{\Delta})$, call it Q. Note that $\Delta(\underline{\beta}) = (\sigma_{ij})$, where $\sigma_{ij} = \sum_t e_{t,i} e_{t,j}$. Then, the derivative of Q with respect to α, where α is one of the parameters of $\underline{\beta}$, is given by

$$\frac{\partial Q}{\partial \alpha} = \sum_{i,j} \Delta_{ij} \frac{\partial \sigma_{ij}}{\partial \alpha}, \tag{2.45}$$

and

$$\frac{\partial \sigma_{ij}}{\partial \alpha} = \sum_t e_{t,i} \frac{\partial e_{t,j}}{\partial \alpha} + e_{t,j} \frac{\partial e_{t,i}}{\partial \alpha}, \tag{2.46}$$

where Δ_{ij} is the ij cofactor of the matrix $\Delta(\underline{\beta})$. A multiplicative factor has been discarded because it will not enter in future computations as we will equate the derivatives to zero, thus discarding nonzero multiplicative factors.

Now, if $\alpha = \beta_{lk}$, where β_{lk} is the k-th parameter in the l-th process, we find that

$$\frac{\partial \sigma_{ij}}{\partial \beta_{lk}} = \left\{ \begin{array}{ll} 0, & \text{if } l \neq j; \\ \sum_t e_{t,i} \partial e_{t,j} / \partial \beta_{lk}, & \text{if } l = i. \end{array} \right. \tag{2.47}$$

Then

$$\frac{\partial Q}{\partial \beta_{lk}} = \left\{ \begin{array}{l} \sum_j \Delta_{lj} \frac{\partial \sigma_{lj}}{\partial \beta_{lk}} \\ \sum_t \frac{\partial e_{t,l}}{\partial \beta_{lk}} (\sum_j \Delta_{lj} e_{t,j}). \end{array} \right.$$

Let

$$u_{t,l} := \sum_j \Delta_{lj} e_{t,j}.$$

Note that $\underline{u}_t$, is a collection of independent vectors, when the parameter $\underline{\beta}$ corresponds to the true parameter. Then

LEMMA **2.30** The derivative $\partial Q / \partial \beta_{lk}$ is a sum of martingale differences.
∇

Let $\partial Q / \partial \underline{\beta}$ be the vector of derivatives with respect to the parameters $\underline{\beta}$. The Hessian matrix $H_N = (\partial^2 Q / \partial \beta_{rs} \partial \beta_{uv})$ is given by

$$\frac{\partial^2 Q}{\partial \beta_{rs} \partial \beta_{uv}} = \sum_t \frac{\partial^2 e_{t,r}}{\partial \beta_{rs} \partial \beta_{uv}} u_{t,l} + \sum_t \frac{\partial e_{t,r}}{\partial \beta_{rs}} \frac{\partial u_{t,l}}{\partial \beta_{uv}} \tag{2.48}$$

but

$$\begin{aligned}\frac{\partial u_{t,l}}{\partial \beta_{uv}} &= \sum_j e_{tj} \frac{\partial \Delta_{lj}}{\partial \beta_{uv}} - \sum_j \Delta_{lj} \frac{\partial e_{tj}}{\partial \beta_{uv}} \\ &= v_{t,luv} - \Delta_{lu} \frac{\partial e_{tu}}{\partial \beta_{uv}},\end{aligned}$$

then

$$\frac{\partial^2 Q}{\partial \beta_{rs} \partial \beta_{uv}} = \sum_t w_{tluv} - \sum_t \frac{\partial e_{t,r}}{\partial \beta_{rs}} \Delta_{lu} \frac{\partial e_{t,u}}{\partial \beta_{uv}} \tag{2.49}$$

and $\sum_t w_{tluv}$ is a sum of martingale differences that converges to zero almost surely (a.s.), if it is divided by N. The second term in (2.49) also converges a.s. and, finally, H_N converges absolutely to a positive definite matrix Σ, after dividing it by N. To see this, note that $e_{t,i} \to a_{t,i}$, as $t \to \infty$ where $\{a_{t,i}\}, i = 1, \ldots, k$ are k white noise process that can be thought of the input of the FARMA process in the definition (2.4). For the purpose of convergence, we can substitute e's by a's in all the above formulae and then classic results on convergence of a sum of i.i.d. random quantities implies that .

$$H_N \to E\{\frac{\partial a_{t,r}}{\partial \beta_{rs}} \Delta_{ru} \frac{\partial a_{t,u}}{\partial \beta_{uv}}\}. \tag{2.50}$$

The derivatives, $\partial a_{t,r}/\partial \beta_{rs}$, are the processes that are obtained in the univariate m.l.e. of the r process. Expressions for them are given by Li and McLeod, (1986). Now, by Taylor's expansion about the true parameters $\underline{\beta}_0$ we can write, discarding terms $O_p(1/\sqrt{N})$,

$$\sqrt{NH_N}(\hat{\underline{\beta}} - \underline{\beta}_0) \approx \sqrt{(NH_N)} \frac{\partial Q}{\partial \underline{\beta}}|_{\underline{\beta}_0}. \tag{2.51}$$

Now, applying a martingale central limit theorem (Hall and Heyde 1980) to the right hand side of (2.51), we obtain the following theorem.

THEOREM **2.31** The m.l.e. estimates of the contemporaneous FARMA model defined in definition (**2.29**) are in the domain of the multivariate Gaussian law

$$\sqrt{NI}(\hat{\underline{\beta}} - \underline{\beta}_0) \rightarrow \mathcal{N}(0, I).$$

▽

Second Multivariate generalization.

We can work with the white noise processes $\{a_{t,i}\}, i = 1, \ldots, k$, instead of with the innovation processes $\{e_{t,i}\}, i = 1, \ldots, k$.

DEFINITION **2.32** Assume that each subprocess $x_{.,i}$ of the multivariate process $x_t = (\{x_{t,i}\}), i = 1, \ldots, k$, satisfies (2.4) independently, i.e. they are long-term processes. Also, generally we will assume that $\underline{a}_t \sim \mathcal{N}(0, \Delta)$, where Δ is a positive definite matrix. ▽

REMARK **2.33** (a) Both definitions of contemporaneous FARMA processes are asymptotically identical. The main difference is in their treatment of the starting values. The first definition is not stationary, although it is asymptotically stationary. The second definition is stationary. The first definition assumes that the new information about the processes, that is the innovations, are dependent through the present, and that the way the new information is related to the past information does not change with time.

(b) Because both processes are asymptotically the same, the asymptotic results which are valid for the first definition are still true for the second definition.

(c) To obtain the m.l.e.'s of the parameters for the second definition is more complicated then with the first definition. First, if we want exact m.l.e.'s we have to compute the likelihood function and this implies that we have to obtain the Cholesky decomposition of the covariance matrix corresponding to the multivariate process. This means that we have to compute the autocovariance, and the crosscovariances using the moving average representation of the univariate processes. After we have computed the covariance matrix we may use a Levinson-Durbin type algorithm to factorize the inverse covariance matrix and to compute the determinant term. This could be done as in Morf et al. (1978). Finally, we compute the values of the parameters that minimize the log-likelihood function. This approach to estimation is very time consuming. A second approach to the m.l. estimation of the parameters may use an approximate maximum likelihood as done by Li and McLeod (1986) for the univariate case. We could approximate the fractional differencing operator ∇^d by its first M terms, ∇^d_M, and assume the use of an ARMA fitting procedure described by McLeod and Sales (1984), for each univariate processes. This produces approximations of $\hat{a}_{t,i}, t = 1, \ldots, N$, where the '^' means that they are values of the white noise process $\{a_t\}$ conditional on the available observations. Finally, we could then minimize the sum of squares function. A third approach for obtaining approximate m.l. estimates follows from the approach of Camacho et al. (1987). This approximate m.l. procedure consists in finding the univariate estimates and then going through an iteration of the Newton-Raphson minimization procedure. This assures the asymptotic efficiency of the estimates.

Treatment of series with different length and/or missing observations.

For handling series contemporaneously related, but of different length and/or having missing observations, we proceed as follows. Assume that we are using the first generalization of FARMA models to contemporaneously related processes. In this case, a missing observation is translated into a zero innovation. We are still assuming that the mean of the processes is zero. Then, we can apply the algorithm and theorems to this case without any problem. Camacho et al. (1987) present a similar procedure for a regular contemporaneous ARMA process. The ideas are the same for the second generalization.

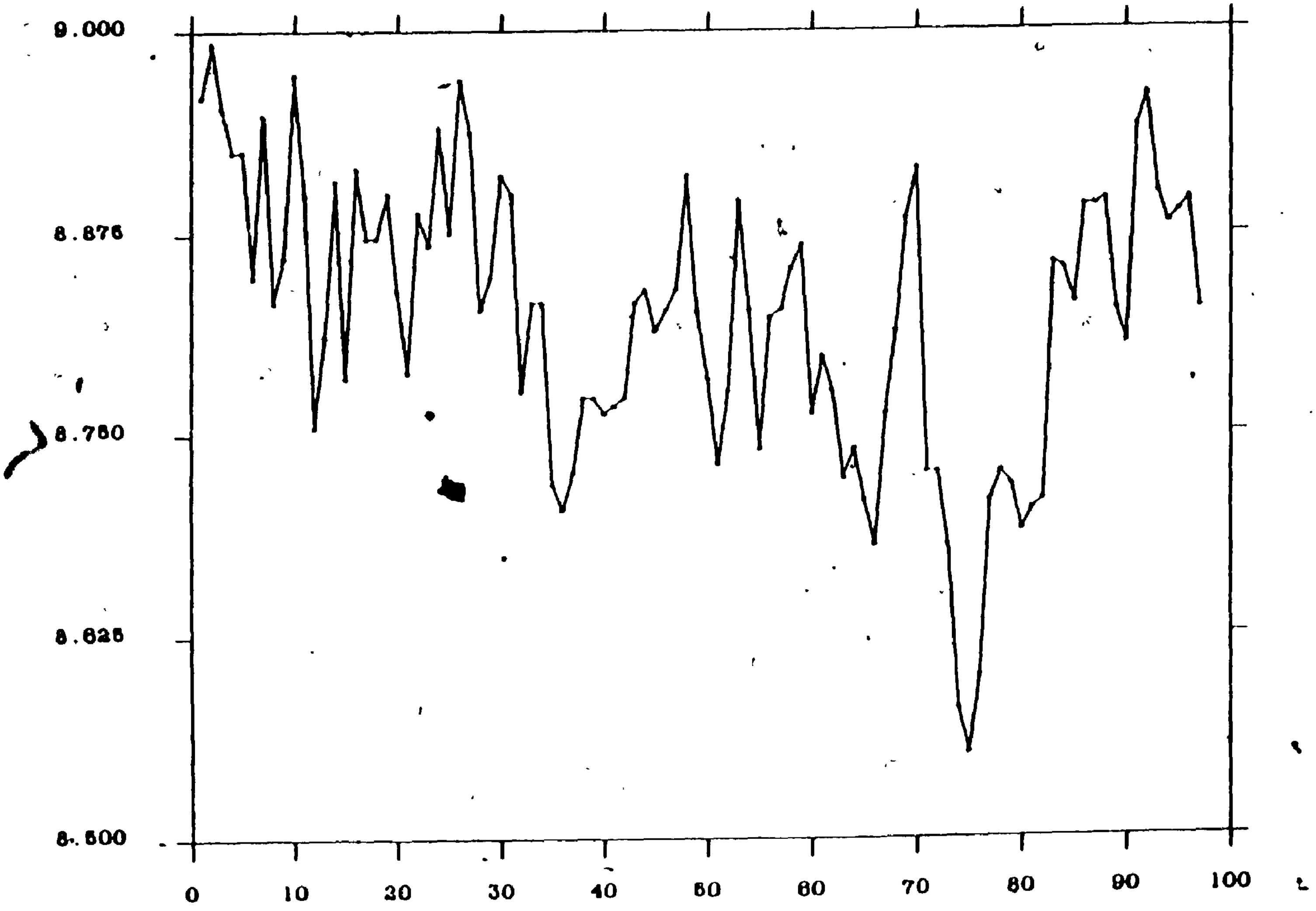


Figure 2.1. St. Lawrence River average annual yearly flow in m**3/s.

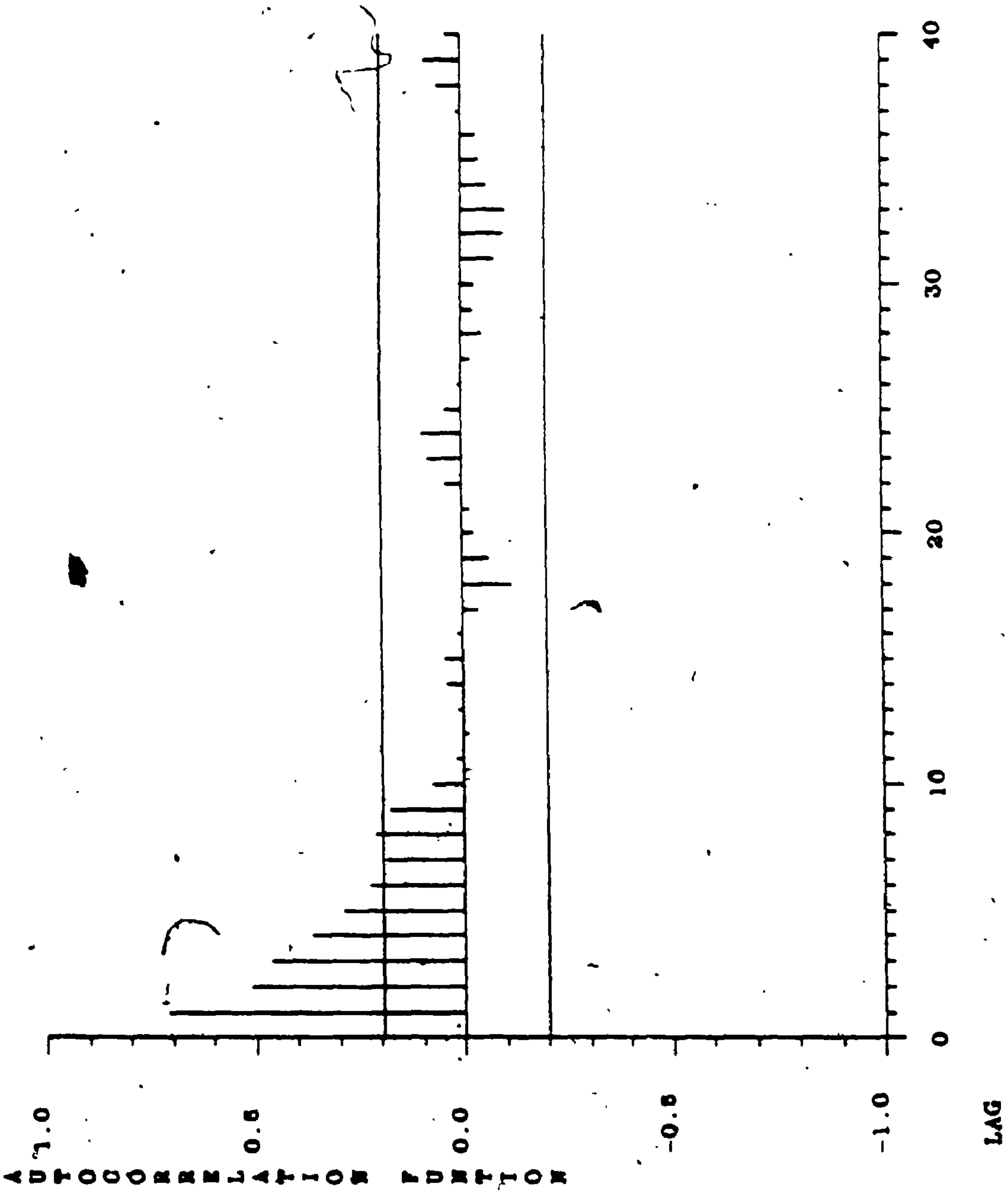


Figure 2.2 Sample ACF for the St. Lawrence River.

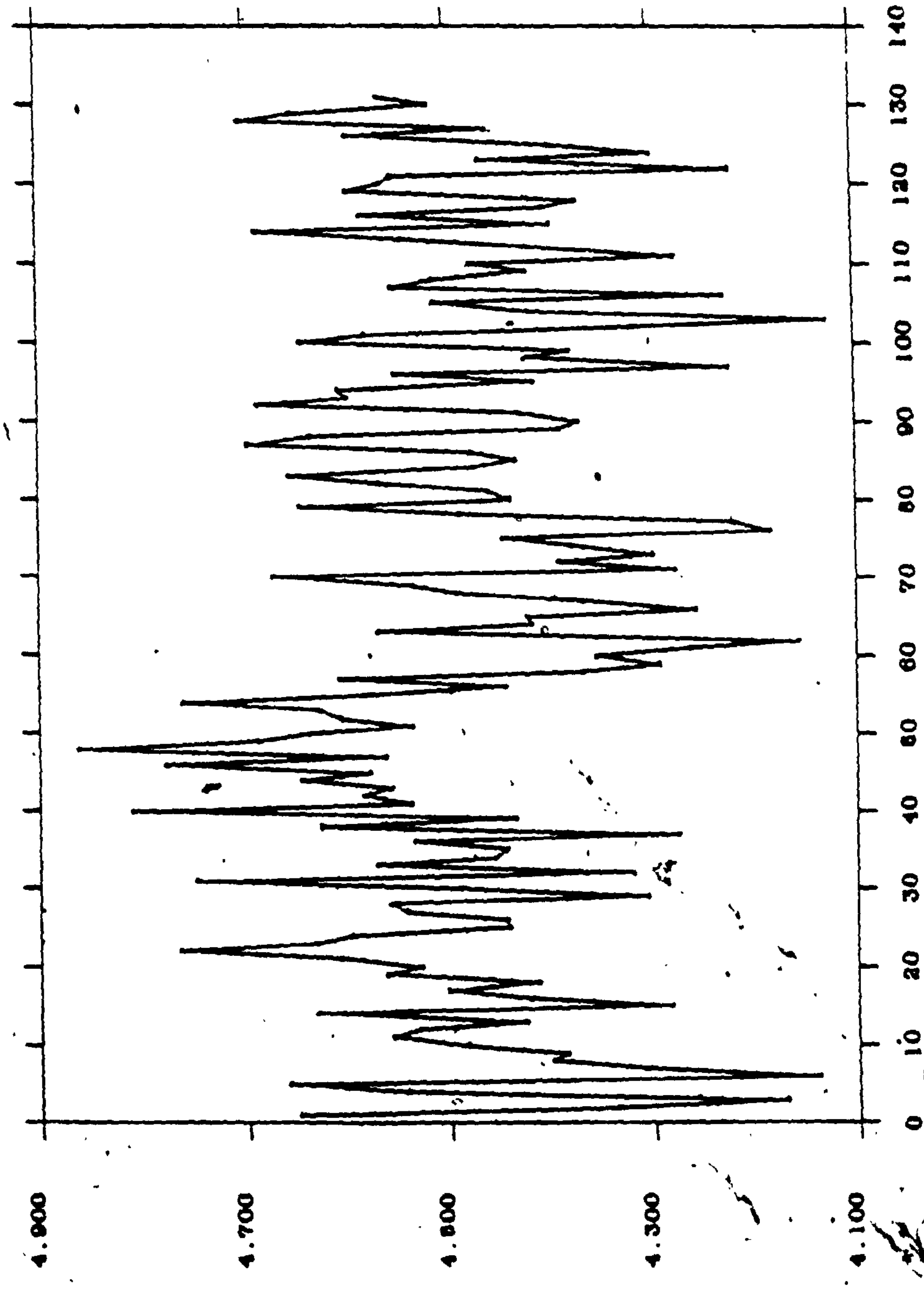


Figure 2.3. Philadelphia's average annual rainfall in mm.

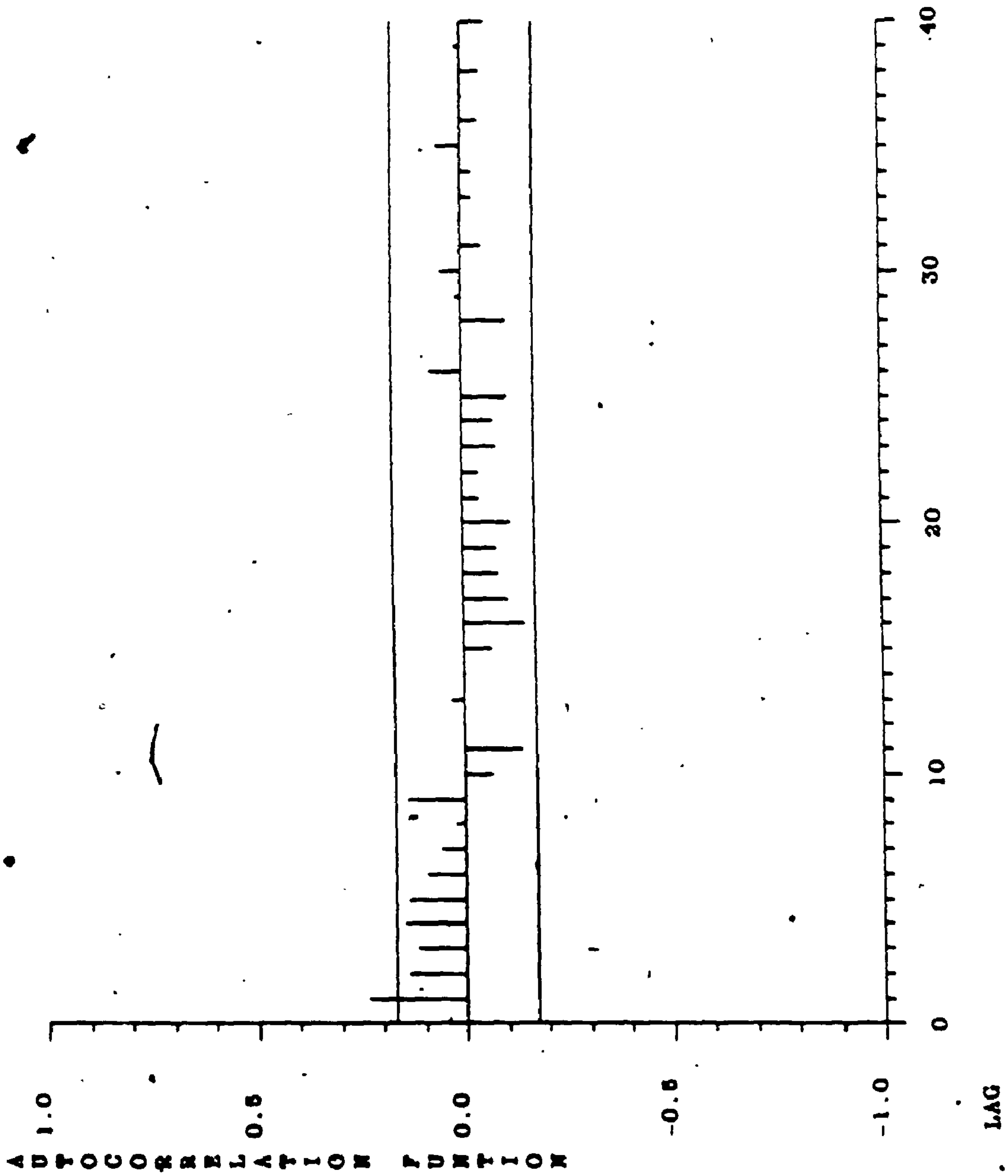


Figure 2.4 Sample ACF for Philadelphia's average annual rainfall

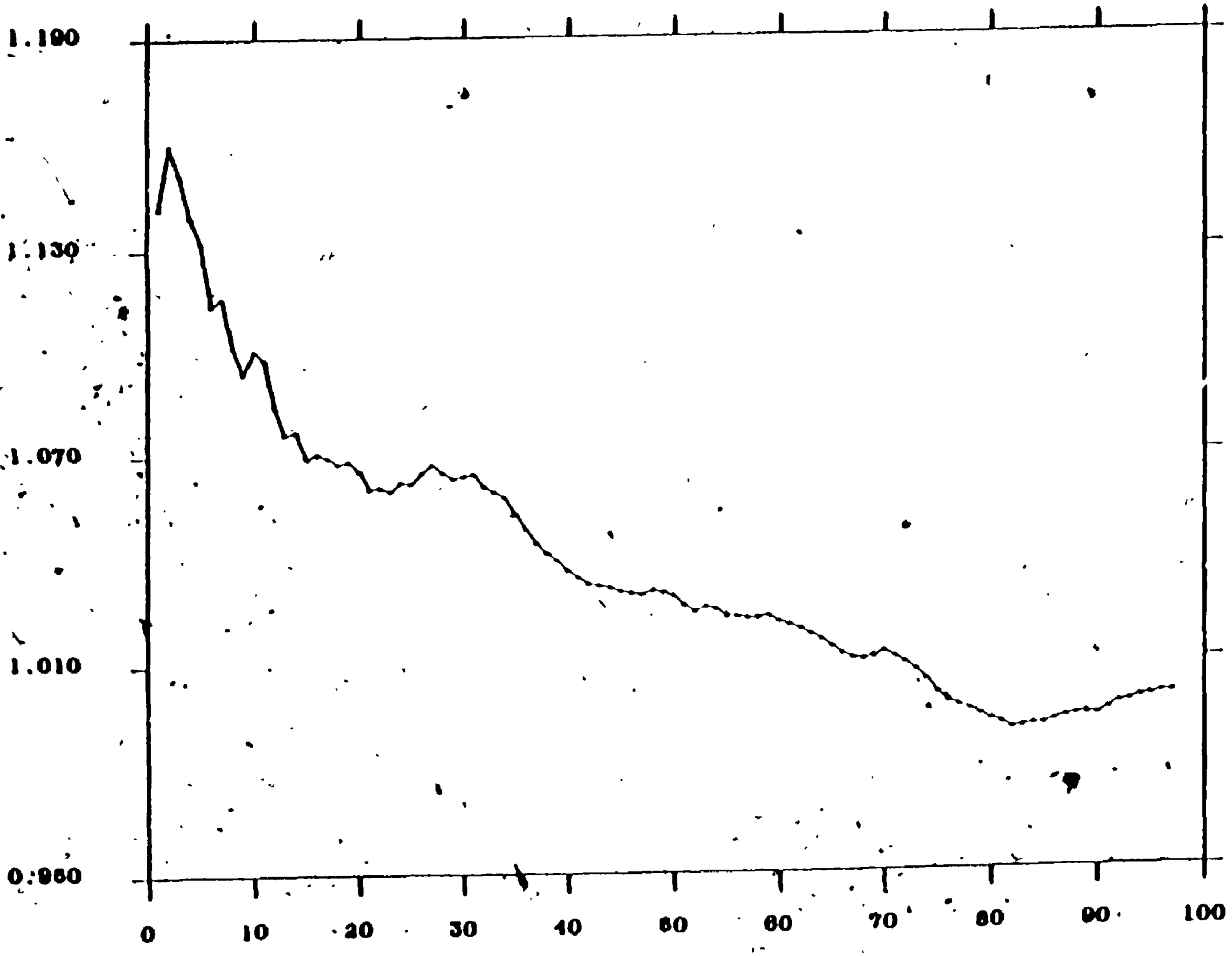


Figure 2.6. Partial means of the average annual m**3/s for the St. Lawrence River.

2.6 Applications

To demonstrate the applicability of FARMA models, the fourteen data sets shown in Table 2.9 were analyzed. The data consists of eleven annual river flows in m^3/s from different parts of the world, two records of average annual rainfall in mm, and an annual temperature series in degrees Celsius. The graph of the average annual St. Lawrence riverflow against time and its sample ACF are shown in Figure 2.1 and 2.2, respectively, while the same graphs for the average annual rainfall at Philadelphia are displayed in Figures 2.3 and 2.4.

In practice, the definition of long-term memory in terms of $M = \infty$ is difficult to check and instead we could use the persistence criterion given in (2.2) to decide if the time series has long term memory. For those time series whose sample ACF decays to zero with an hyperbolic rate, the possibility of modelling them by FARMA models is considered. Within the fourteen data sets, the St. Lawrence riverflows and the Philadelphia Rainfall series show an estimated ACF that seems to decay to zero hyperbolically. Therefore, these two data sets seem to present some evidence suggesting the use of FARMA models. For other records such as Saugeen river and rainfall at Fortaleza, the evidence, as given by the estimated ACF, in favor of a persistence parameter is not so strong, but it is a possibility. However, it should be remarked that if the persistence parameter d is close to zero, between 0 and 0.2, detection of long term memory by visual inspection of the autocorrelations can be difficult. Moreover, because Bartlett's formula needs to be multiplied by a factor of order N^{-1+4d} if $d \geq 0.25$, in case of a FARMA(p, d, q) process, visual inspections of the sample ACF should be used with care when it is suspected that the process under analysis could have long term memory.

If the process belongs to the FARMA family of models the PACF should decay to zero at a hyperbolic rate. This rate is independent of the degree of persistence. However, for the case of a FARMA$(0, d, 0)$ process, long term memory implies that all the autocorrelations should be positive This behaviour of the PACF of the FARMA$(0, d, 0)$ process suggests that to detect persistence not only a hyperbolic decay of the PACF is of interest, but also the behaviour of the signs of the PACF. This suggests the use of a nonparametric sign test to test the signs of the estimated PACF. None of the estimated PACF's of the fourteen data sets show strong evidence of a hyperbolic rate of decay to zero. However, some of them, like the St. Lawrence riverflows and Philadelphia rainfall show PACF structures that are generally positive. A sign test of these PACF's gives further support for the conjecture that these time series exhibit signs of persistence.

Another characteristic of a time series that could indicate the presence of persistence is the behaviour of the partial sample means, $\bar{x}_k, \quad k = 1, \ldots, N$, of the process that are defined as

$$\bar{x}_k = (\frac{x_{k-1} + \cdots + x_0}{k}). \qquad (2.52)$$

For the case of a short memory time series, a plot of $\bar{x}_k$ against k should show great stochastic variability for the first values of k, but after k reaches a moderate value the graph should decay to an almost constant value and should show small stochastic variability. However, for the situation of a long memory time series, the plot of $\bar{x}_k$ against k should show great stochastic variability for the first few values of k. For moderate values of k the graph should display a gentle trend that should oscillate around a constant value as k increases. After k reaches a very large value, which depends on the degree of persistence, $\bar{x}_k$ should reach a constant value. Furthermore, because the present values of the time series are correlated with the past, the present values of $\bar{x}_k$ are highly correlated with the past and, therefore, a plot of $\bar{x}_k$ against k could show local trends. To detect persistence, the rate of decay towards a constant value of the local trends is of interest, as is also the presence of an overall gentle trend. However, the presence of local trends in the plot of $\bar{x}_k$ against k by itself does not indicate the presence of persistence. The graph of $\bar{x}_k$ against k of the St. Lawrence riverflows is displayed in Figure 2.5. The St. Lawrence riverflows present an overall decreasing trend. This trend is gentle enough to assume that it could be due to the presence of persistence in the time series and not due to nonstationarity.

Table 2.9

River flow records

Data set Identification	Geographical position	Time span	Length
Saugeen	Walkerton, Canada	1915-1976	62
Dal	near Norslund, Sweden	1852-1922	70
Danube	Orshava, Romania	1837-1957	120
French Broad River	Asheville, N. Carolina, USA	1830-1900	70
Gota	near Sjotop-Vannersburg, Sweden	1807-1957	150
Mckenzie	Mckenzie Bridge, Oregon, USA	1900-1956	55
Missisipi River	St. Louis, Missouri, USA	1861-1957	96
Neumunas	Smallininkai, Lithuania, USSR	1811-1943	132
Rhine River	Basle, Switzerland	1807-1957	150
St. Lawrence	Ogdensburg, New York, USA	1800-1930	131
Thames	Teddington, England	1883-1954	71
Rain Philadelphia	Philadelphia,USA	1820-1950	131
Rain Fortaleza	Fortaleza, Brasil	1849-1979	131
Average Temperature	Central England	1723-1970	248

For some of the other data sets local trends in $\bar{x}_k$ seemed to be present even at the end of the series. Finally, for most of the data sets the behaviour of the partial means is consistent with what could be expected in time series with a short term memory, which is a rapid decay of the graph to a constant value.

All of the FARMA models considered for fitting to the series in Table 2.9 are subsets of the FARMA(2, d, 1) model given by

$$(1-\phi_1 B-\phi_2 B^2)\nabla^d(B)(x_t-\mu)=(1-\theta_1 B)a_t, \qquad (2.53)$$

where ϕ_i is the ith autoregressive parameter and θ_1 is the first moving average parameter. For the St. Lawrence riverflows, the additional AR(3) model given by

$$(1-\phi_1 B-\phi_3 B^3)(x_t-\mu)=a_t, \qquad (2.54)$$

was considered, because this was the model selected by Hipel and McLeod(1977), within the class of ARMA models. The most appropriate FARMA model from (2.53) to fit each series was selected according to the minimum Akaike information criterion (AIC) (Akaike, 1974) considering only those models that passed tests for whiteness of the fitted residuals. The maximum likelihood estimates of the model parameters for each series along with the standard errors given in brackets are displayed in Table 2.10. Those time series where the estimate of d given in Table 2.10 is positive show persistent behaviour. Also, as the degree of persistence depends on the magnitude of d, for those series with higher values of d the degree of persistence is higher than the degree of persistence in time series having d's of smaller magnitudes. For example, the model for the St. Lawrence river was estimated as a FARMA$(0, d, 0)$ with $d = 0.4999$. This indicates that the flows of the St. Lawrence are persistent in a strong sense, Therefore, the memory M is infinite and the far away past strongly influences the present. A consequence of this influence is the slow rate of convergence of the sample mean to the true value. For the case of the St. Lawrence river this rate of decay is of order $O(N^{-0.0001})$, where N is the number of observations. This order of convergence is also true for the forecasting function and the estimated ACF of the St. Lawrence riverflows. An interesting feature of the St. Lawrence river is that it is associated with great masses of water which perhaps suggests a model with a reservoir term whose time step is larger than the time step used to measure the series.

All the models that exhibit persistence in Table 2.10 are FARMA$(0, d, 0)$. This model is used for Mckenzie river, Thames river, and Philadelphia rainfall series. There were other data sets where the AIC selected an ARMA model but the differences between the minimum AIC for the ARMA models and the AIC for FARMA$(0, d, 0)$ were very small. Finally, note that some rivers do not show any sign of second order correlation structure as the optimal model according to the AIC was simply the mean. This model was used for Dal, Danube, and Rhine rivers. Note that most of these findings are consistent with the models suggested by the ACF, PACF and behaviour of the partial means.

Table 2.10

Best FARMA models

Data set identification	ϕ_1	ϕ_2	d	θ_1
Saugeen	-	-	-	-
Dal	-	-	-	-
Danube	-	-	-	-
French Broad R.	-0.234 (0.12)	-	-	-
Gota	0.587 (0.08)	-0.27 (0.08)	-	-
Mckenzie	-	-	- (0.1)	-
Missisipi River	0.29 (0.1)	-	-	-
Neumunas	-	-	-	-0.19 (0.08)
Rhine River	-	-	-	-
St. Lawrence	-	-	0.499 (0.08)	-
Thames	-	-	0.12 (0.1)	-
Rain Philadelphia	-	-	0.23 (0.08)	-
Rain Fortaleza	0.240 (0.08)	-	-	-
Average Temperature	0.120 (0.06)	0.20 (0.06)	-	-

Standard deviations of the estimates are given in parenthesis under the estimated value.

2.7 Summary

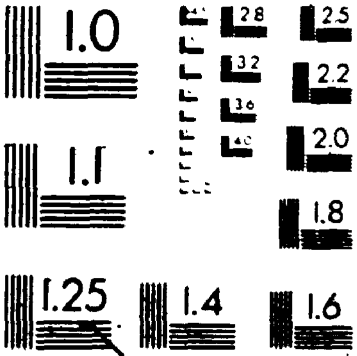
The fact that at zero the spectral density function a of FARMA time series is either infinite or zero, poses unique problems with respect to its definition, identification and parameter estimation. In this chapter, we have found that the sample mean of an antipersistent FARMA process is not an efficient estimator (Theorem 2.11). Also, for the estimation of regression parameters when the observations are contaminated by fractionally differenced antipersistent noise the estimates by O.L.S. are generally not efficient (Theorem 2.23). However, when the noise is a fractionally differenced long term memory process the estimation of the regression parameters by O.L.S is generally consistent. The extreme eigenvalues of the autocovariance matrix of a FARMA process approach 0 and ∞. This implies the nonefficiency of O.L.S. for the regression parameters discussed above.

McLeod (1977) obtained a good approximation for the determinant of the autocovariance matrix of an ARMA process. We have developed a similar approximation for the case of fractionally differenced noise, FARMA$(0, d, 0)$ (Proposition 2.16). Now, asymptotically the logarithm of the determinant of the autocovariance function of a FARMA(p, d, q) process decomposes into the sum of the fractionally differenced part and the ARMA part (Theorem 2.19). We have argued, based on the preceding fact that the above approximation of the determinant of a FARMA$(0, d, 0)$

process should be reasonable also in the case of a FARMA(p, d, q) process.

We have compared Li and McLeod's (1986) approximate maximum likelihood estimation with exact maximum likelihood using the mean square error. Both algorithms are comparable when the number of observations is about 75. Also, we have observed that there is high probability of the exact maximum likelihood to choose the 1/2 value if the true parameter is close to this boundary. Then, based on the results of Barnard et al. (1962) and the simulation results of Copas (1966), it seemed reasonable to use an average maximum likelihood method for estimating the parameters. We have compared two loss functions, one based on the normal density function, L_1, the other based on the t-distribution, L_2. We found that the estimates based on the t-distribution outperform the estimates obtained by maximum likelihood and the average likelihood based on the normal density when the number of observations is small to medium. The rate at which the relative performance of the L_2 loss function, with respect to the other two estimation approaches, deteriorates, is faster the closer to the 1/2 boundary is the fractionally differencing parameter d. It seems sensible to use the L_2 loss function when the sample size is small, if d is not suspected to be close to 1/2. However, for a full fledged FARMA process having autoregressive and moving average parameters to estimate, the computational burden of the average likelihood estimation methods is much higher than with the m.l. method for estimation.

2



We have given, based on the partial autocorrelation coefficients, a definition (Remark 2.2) of processes that may exhibit long term dependence but with a richer autocovariance structure than the FARMA(p, d, q) models.

Finally, two multivariate generalizations of FARMA processes have been proposed. We gave these generalizations for the case of a contemporaneous time series process. Both generalizations are similarly defined except that in one we worked with the innovations and in the other with the white noise process. We have presented the asymptotic theory of thee maximum likelihood estimates. Also, we have given an algorithm for computing the likelihood.

References

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC–**19**, 716–723.

2. Bartlett, M. S. (1946). On the theoretical specification of sampling properties of autocorrelated time series.

3. Barnard G. A., Jenkins, G. M., Winsten, C. B. (1962). Likelihood inference and time series. *J. Roy. Statist. Soc. Ser. A*, **125**, 321–373.

4. Ballerini R. and Boes D. C. (1985). Hurst behavior of shifting level processes. *Water Resources Research*, **21**, 1642–1648.

5. Bhattacharya, R. N., Gupta, V. K., Waymire, E. (1983). The Hurst Effect Under Trends *J. Appl. Prob.*, **20**, 649–662.

6. Boes D. C. and J. D. Salas (1973). On the expected range of partial sums of exchangeable random variables. *J. Appl. Prob.*, **10**, 671–677.

7 Boes D. C. and J. D. Salas (1978). Nonstationarity of the mean and the Hurst phenomenom. *Water Resourcees Researeh*, **14**, 135–145.

8. Box, G.E.P., and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control* (Second Edition). Holden–Day, San Francisco, California.

9. Brass, R. and I. Rodriguez–Iturbe, (1984). *Random Functions and Hydrology*. Addison–Wesley.

10. Brillinger, D. (1976). *Time series: Data Analysis and Theory*, Holden–Day, San Francisco, California (Second Edition).

11. Burton, R. and Waymire, E. (1983). Limit theorems for point random fields *J. Math. Anal.*.

12. Camacho, F., McLeod, I. A., Hipel, K. W. (1987). Contemporaneous bivariate time series. *Biometrika*, **74**, 691–694.

13. Copas J. B. (1966). Monte Carlo results for estimation in stable Markov time series. *J. Roy. Statist. Soc. Ser. A*, **129**, 110–116.

14. Cox, D. R. (1984). Long range dependence: a review. *In Statistics: An Appraisal*, H. A. David and H. T. David (editors). Proceedings of the 50th Anniversary Conference, Iowa State Statistical Laboratory, Iowa State University Press, Ames, Iowa; 55–74.

15. Cryer, J. D. and Ledolter J. (1981). Small-sample properties of the maximum likelihood estimator in the first–order moving average model. *Biometrika*, **68**, 691–694.

16. Durbin, J. (1960). The fitting of time–series models. *Rev. Inst. Int. Statist.*, **28**, 233–244.

17. Finch, P. D. (1960). On the covariance determinant of moving-average and autoregressive models. *Biometrika*, **47**, 194–196.

18. Geweke, J. and Porter-Hudak S. (1983). The estimation and application of long memory time series models. *J. Time Series Analysis*, **4**, 221–238.

19. Geronimus, Ya. L. (1960). *Polynomials Orthogonal on a Circle and Interval.* Pergamon Press, Oxford.

20. Geronimus, Ya. L. (1977). Orthogonal polynomilas. *American Math. Soc. Transl., Ser. 2*, **108**, 37–121.

21. Golub G .H. and van Loan C. (1983). *Matrix computations.* John Hopkins University Press, Baltimore, Maryland.

22. Granger, C. W. J. (1980). Long-memory relationships and the aggregation of Dynamic models. *Journal of Econometrics*, **14**, 227–238.

23. Granger, C. W. J. and R. Joyeux (1980). An introduction to long memory time series models and fractional differencing. *J. Time Series Analysis*, **1**, 15–29.

24. Grenander, U. and Szegö, G. (1958). *Toeplitz Forms and their Applications.* Chelsea Publishing Co., New York.

25. Hall, P., Heyde, C.C. (1980). *Martingale Limit theory and its application.* Academic Press, New York.

26. Hannan, E. J. (1970). *Multiple Time Series.* John Wiley, New York, New York.

27. Hipel, K. W. (1981). Geophysical model discrimination using the Akaike information criterion. *IEEE Transactions on Automatic Control*, **AC–26(2)**, 358–378.

28. K. W. Hipel, McLeod, A. I. (1978). Preservation of the rescaled adjusted range, part two, simulation studies using Box-Jenkins models. *Water Resources Research*, **14**, 509–516.

29. K. W. Hipel, McLeod, A. I., Lennox, W. C. (1977). Advances in Box-Jenkins modelling. *Water Resources Research*, **13**, 567–586.

30. Hipel, K. W. and A. I. McLeod, (1989). *Time Series Modelling for Water Resources and Environmental Engineers.* Elsevier, Amsterdam (in press).

31. Hosking, J. R. M, (1981). Fractional differencing. *Biometrika*, **68**, 165–176.

32. Hosking, J. R. M. (1984). Modeling persistence in hydrologic time series using fractional differencing. *Water Resources Bulletin.*, **20**, 1898–1908.

33. Hosking, J. R. M. (1985). Fractional differencing modeling in hydrology. *Water Resources Res.*, **21**, 677–682.

34. Hurst, H. E. (1951). Long–term storage capacity of reservoirs *Trans. Amer. Soc. Civil Eng.*, **116**, 770–808.

35. Hurst, H. E., Black, R. P., Simaika, Y. M. (1965). *Long term storage: An Experimental Study.* Constable Press, London.

36. Klĕmes, V. (1974). The Hurst phenomenom- A puzzle?. *Water Resources Res.*, **10**, 675–688.

37. Kolmogorov, D. (1935). Sur interpolation et extrapolation des suites stationaries. *C.R. Acad. Sci., Paris*, **208**, 2043–2045.

38. Kottegoda, N. T. (1980). *Stochastic water resources technology.* MacMillan Press, London, England.

39. Leopold, L. B., Wolman, M. G., Miller, J. P. (1964). *Fluvial Processes in Geomorphology*. W. H. Freeman, San Francisco.

40. Li, W. K. and A. I. McLeod, (1986). Fractional time series modelling. *Biometrika*, 73, 165–176.

41. Mandelbrot, B. B., and Wallis, J. R. (1968). Naoh, Joseph, and operational Hydrology. *Water Resources Research*, 4. 909–918.

42. Mandelbrot, B. B., and Wallis, J. R. (1969a). Computer experiments with fractional Gaussian noises, 1: Averages and variances. *Water Resources Research*, 5, 228–241.

43. Mandelbrot, B. B., and Wallis, J. R. (1969b). Computer experiments with fractional Gaussian noises, 2: Rescaled ranges and spectra, *Water Resources Research*, 5, 242–259.

44. Mandelbrot, B. B., and Wallis, J. R. (1969c). Computer experiments with fractional Gaussian noises, 3, Mathematical appendix, *Water Resources Research*, 5, 260–267.

45. McLeod, A. I. (1977). Improved Box–Jenkins estimators. *Biometrika*, **64**, 531–534.

46. McLeod, A. I. and K. W. Hipel, (1978a). Preservation of the rescaled adjusted range, part one. A reassesment of the Hurst phenomenon. *Water Resources Research*, **14**, 491–508.

47. McLeod, A. I. and K. W. Hipel, (1978b). Simulation procedures for Box–Jenkins models. *Water Resources Research*, **14**, 969–975.

48. McLeod, A. I. and Sales, P. R. (1983). An algorithm for the approximate likelihood calculations of ARMA and seasonal ARMA models. *Appl. Statistics*, **32**, 211–223.

49. Mejia, J. M., I. Rodriguez–Iturbe, and D. R. Dawdy (1972). Streamflow simulation 2,The broken line process as a potential model for hydrologic simulation. *Water Resources Res.*, **8**, 931–941.

50. Milhøj, A. (1984). Multiplicative exponential models for stationary time series. *J. Time Series Analysis*, **5**, 19–35.

51. Morf, M., Viera, A., Kailath, T. (1978). Covariance characterization by partial autocorrelation matrices. *Annals of Mathematical Statistics*, , **6**, 643–648.

52. Newman, C. (1980). Self–similar random fields in mathematical physics *Proc. Measure Theory Conference*, Delkab, Il.

53. Parzen, E. (1982). ARARMA models for time series analysis and forecasting. *J. Forecasting*, **1**, 67–82.

54. Pólya, G. and Szegö, G. (1972). *Problems and Theorems in Analysis. vol 1.*, Springer–Verlag.

55. Rosenblatt, M. (1961). Independence and dependence. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, 431–443.

56. Rosenblatt, M. (1979). Some limit theorems for partial sums of quadratic forms in stationary Gaussian variables. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 49, 125–132.

57. Rosenblatt, M. (1981). Limit theorems for Fourier transforms of functional of Gaussian sequences. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55, 123–132.

58. Yevjevich, and G. G. S. Pegram (1979). Hurst phenomenom as a pre–asymptotic behaviour *J. Hydrology*, 44, 1–15.

59. Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane (1980). *Applied Modelling of Hydrologic Time Series*. Water Resources Publications, Littleton, Colorado.

60. Szegõ, G. (1921). Über die Entwicklung einer analylischen Funktion nach dem polynomen eines orthogonal systems. *Math. Ana.*, **82**, 121–145.

61. Trench, W. F. (1964). An algorithm for the inversion of finite Toeplitz matrices. *S.I.A.M. J. Appl. Math.*, **12**, 515–521.

62. Taqqu, M. (1975). Weak convergence to fractional Brownian motion

and the Rosenblatt process. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **31**, 287–302.

63. Taqqu, M. (1979). Self–similar processes and related ultraviolet and infrared catastrophes. *International Conference of Statisticians.*

Chapter 3

PERIODIC MODELS

3.1 Introduction

Many time series processes exhibit periodicity or seasonality. That is, although the mechanism that drives the process seems to evolve over time, it repeats itself after some period of time. Two important examples of seasonal time series are riverflows and many time series related to the economy. For example, for many rivers the average monthly flows are different in winter than in summer, or in the wet season compared to the dry season. However, the river flows in winter seem to behave similarly across different years. The same may be true for the summer flows. In general, geophysical processes exhibit, for some time scales, some form of seasonality. Many economic processes also exhibit a strong cyclic behaviour. In many instances, it is the modelling of this periodic component that is one of the main reasons to analyze the data. Hence, as the above examples show there is, evidently, a need to work with models that can exhibit seasonal behaviour similar to that exhibited by the real life processes under study.

Because of the importance of modelling seasonality many different stochastic models have been proposed to account for the periodicity. Perhaps the most popular modelling methodology has has been the Box-Jenkins (1976) fitting procedures based on seasonal integrated autoregressive-moving average models, or simply SARIMA models. Another important modelling approach is the so called components method, of which we will explain more in a later section.

We will briefly consider Box-Jenkins models in order to motivate the definition of periodic ARMA models (PARMA). SARIMA models are defined by

$$\Delta^l(B^S)\Phi(B^S)\Delta^k(B)\phi(B)(x_t - \mu) = \Theta(B^S)\theta(B)a_t, \tag{3.1}$$

where S, the number of seasons; B is the backshift time operator; $\Delta(B) := 1 - B$ the difference operator; μ is the mean of the process x_t $\{a_t\}$ is a white noise sequence, and ϕ, θ are polynomials in B of degree p and q, respectively, whose zeros lie outside the complex unit circle and they do not share a common zero. Similarly, Φ, Θ are polynomials in B^S of degree P and Q, respectively; the zeros of Φ and Θ lie outside the complex unit circle, and they do not have common zeros. In a signal processing context, the process described by (3.1) corresponds to a process that must be filtered by an ARMA filter plus an ARIMA filter with lag size corresponding to the the seasonal period in order to obtain white noise. Also, Box-Cox transformations (Box and Cox, 1964) given by

$$x_t^\lambda = \begin{cases} \dfrac{x_t^\lambda - c}{\lambda} \\ \log(x_t) \end{cases}$$

are often used in connection with (3.1) to extend the applicability to the case of nonnormal or nonlinear time series processes.

The Box-Jenkins approach assumes that the deterministic trends that are present in the data may be removed by differencing, i.e. it assumes that the trends obey a polynomial equation in time. And, also, in a sense, the model is time constant because the defining equation (3.1) is constant for all time. For the SARIMA models in (3.1), the mean μ_t and the variance σ_t^2 of x_t are constant functions of time and hence the dependence between two points in time is the same no matter where these points are situated in the time scale. For the case of a monthly time series the above assumptions for model (3.1) means that the relationship between, say, January and March is the same as that between July and October. Although SARIMA models seem to be very popular among people who have to fit a model to periodic data, it is not difficult to encounter time series for which the SARIMA model does not successfully fit the time series. For example, monthly river run-offs with a raining season that produces random inputs having a higher variance than those in the dry season, cannot be properly modelled by the SARIMA family of models. Usually, these situations have been dealt with by "deseasonalizing" the time series by subtracting out seasonal means and dividing by the seasonal standard deviations. After this has been done, to obtain a white noise process, an ARMA model is fitted to the deseasonalized time series. This method for deseasonalizing seems to be widely used by time series modelers.

It is not unreasonable to presume that if the mean and variance of the process are seasonally dependent so should be their autocorrelation functions, or equivalently the transfer functions of the process. Thus, we arrive naturally to a generalization of stationary models to models with periodically varying second (and first) order properties (Gladeshev 1961, 1963). In particular, we obtain an extension of ARMA models to models with seasonally varying parameters, i.e., PARMA models. (Note that ARMA models describe only the second order properties of the time series process.)

To define PARMA models, assume that the seasonal period is S, i.e., in case of monthly time series $S = 12$. Also we will use the following notation $T = (t, i)$ to indicate that the time point T belongs to the the set with tag i and it is the t-th value within this set; also there are S different sets. For example, in a regular monthly periodic model $(t, i) = (t-1)S + i$, this is almost mod notation except for the fact that sometimes we allow i to to be outside the set $0, ..., S-1$. However, the general notation is useful for more complicated situations such as random seasons, seasons with unequal lengths, etc. PARMA models are defined by

$$\phi_i(B)(x_{(t,i)} - \mu_i) = \theta_i(B)a_{(t,i)}, \qquad 0 \le i \le S; \tag{3.2}$$

$\sigma_i^2 := \text{var}(a_{(k,i)})$, $\phi_i(B)$ and $\theta_i(B)$ are polynomials in B that are relative prime between them and that satisfy some additional joint constraints in order to obtain a well defined time series.

Note that this extension of ARMA models could also be done with other time series models. For example, we could extend transfer function models to accommodate seasonal behaviour to obtain periodic transfer function models. The transfer function could be seasonal or nonseasonal and, moreover, it could have a different seasonal period than the ARMA part of the model. Also, it is clear how we could extend the scope to intervention analysis for periodic models. Multivariate periodic models, MPARMA, are a natural generalization to model interrelated seasonal quantities, such as two monthly riverflows, and rainfall series. An important subclass of multiple PARMA time series could be contemporaneous periodic time series models, i.e., multivariate time series with dependence just in the present random input and with periodic second order structure. Another interesting subclass of models could be the periodic casual models.

Periodic structure could also be built in nonlinear models. For example, models like exponential autoregressive, (EAR), (Lawrence and Lewis, 1980), new exponential autoregressive, (NEAR), (Lawrence and Lewis, 1985), state-dependent models (Priestley 1980) and threshold models (Tong and Lim, 1980) could be extended to the case of periodically varying parameters. Finally, discrete count models of the type considered by McKenzie (1986) could also be easily extended to the case of periodically varying parameters.

Hence, PARMA models seem to be a natural generalization of ARMA and related models, with a very important and large field of application. They are almost a "necessary" generalization.

The principal drawback of these models seems to be the increase in the number of parameters. For example, an AR(1) process needs three parameters to be specified: mean, variance and first order correlation. The analogous AR process with monthly data needs thirty-six parameters This is almost half or more of the number data points usually available.

One way of solving this "curse of the dimensionality" is to restrict the models in (3.2) to some particular subfamily. A interesting subclass of PARMA models are models where the parameters are thought to vary according to some trend. For example, the means and variances may slowly increase, shoot-up and then, decrease and level off, as is the case with many river flows. In particular, for the case of PAR(1) models the parameters could easily be visualized as following a trend. Higher order models could perhaps be modeled by using insight into the autocorrelations function, partial autocorrelation and spectral density.

3.2 History

Time series analysts traditionally have been interested in the cyclic or seasonal aspects of the process under study. Spectral frequency ..

analysis illustrates this point. However, models with first and second order time dependent characteristics were studied first by Gladeshev (1961, 1963). He defined seasonally varying models as models where the means and the autocorrelation function are periodic. He used a spectral frequency domain approach for describing these models. He proved that the ACF of a seasonally varying process can also been seen as the ACF of a S-dimensional time series process. Also, he showed that for each season, it is possible to obtain a spectral measure $F_i(\lambda)$ such that the ACF for this season corresponds to the Fourier coefficients of this measure.

Another early paper is by Jones and Brelsford (1968). In their paper, they introduced the class of seasonal varying autoregressive and moving average models. They consider the estimation of the parameters for the autoregressive case. Their method is based on the following ideas. Because of the seasonality, use the first few terms of the Fourier expansion of the periodic autoregressive coefficients for estimating the coefficients. They use ordinary least squares for the estimation. Their approach is then equivalent to regressing the current time value x_t onto the first few Fourier components of the past given by
$(x_{t-j}, x_{t-j}\sin(2k\pi/S), x_{t-j}\cos(2k\pi/T), k = 1,..,K; j = 1,..,p)$.

Pagano (1978) defined periodic autoregressive (PAR) models and showed that the PAR models and S-dimensional multiple autoregressive time series are equivalent. However, there are many situations, (i.e. monthly riverflows) where we may use a family of PAR processes that needs fewer parameters to be specified than a corresponding multivariate model. For example a PAR(1) process needs S parameters, apart from the means and the variances, a multivariate AR(1), needs S^2 parameters. He also shows that the natural extension of the estimator for the ACF to PAR models is consistent and asymptotically normal. He shows that the ACF satisfies a Yule-Walker type of difference equation. This fact is used to propose an estimate of the parameters, which are asymptotically efficient and normally distributed. Finally, he shows that the information matrix is block diagonal.

Troutman (1979) used the alternative characterization of PAR models as multivariate AR models to give an infinite moving average representation of PAR models. He derived some constraints that the parameters must satisfy in order to obtain an infinite moving average representation or, equivalently, in order that the associated multivariate AR model be stationary. Moreover, he showed that the estimates obtained by Pagano's generalization of Yule-Walker equations automatically satisfy the stationarity constraints and that the estimated variance matrix is positive-definite.

Sakai (1982) gave a Levinson-Durbin type of recursive algorithm for solving the periodic Yule-Walker equations, and showed that the process has a circular lattice structure that can be exploited to obtain a Burg-type algorithm to estimate by maximum-entropy the spectral function of the process. Other papers concerned with moment estimators are Salas et al. (1982) and Salas et al. (1985). In this last paper the authors consider moment estimators for periodic autoregression with a moving average parameter.

Maximum likelihood estimation of PARMA models with gaussian inputs was discussed by Vecchia (1985). The approach he uses to write the likelihood is the same as Newbold (1974) for the univariate case and Hillmer and Tiao (1979) for the multivariate autoregressive moving average case. Moreover, he proved that PARMA models and multivariate autoregressive moving average models are equivalent. His algorithms seem to be feasible only when the number of seasons is small. Estimation of parameters has also been considered by McLeod and Hipel (1978) for the PAR case.

Identification issues have been studied by Cleveland and Tiao (1979). In their paper they extend the Box-Jenkins approach to the PARMA case. Jones and Brelsford (1967) used a spectral density approach for estimation together with Gladeshev's (1961) result that the spectral density concentrates along parallel lines. Tiao and Gruppe (1980) considered the consequences of misspecification of a seasonal process when the seasonality is not modeled using a time varying model. McLeod and Hipel (1978) extend Box-Jenkins methodology to PAR models. In particular they developed the Portmanteau diagnosis techniques. Also, they developed an automatic technique for model selection based on an extension of the AIC to the PAR models. To obtain a computationally feasible algorithm they use the algorithm of Morgan and Tartar (1974) to choose the subset autoregression with the minimum AIC. Their guiding building principle is parsimony. Another paper interested in parsimony is Thompsone et al. (1984a) where they are interested in obtaining a parsimonious PAR model but with the additional feature that they group seasons that are not very different. They developed a test to decide which seasons are similar.

Asymptotic distributions results are given, as already mentioned, by Pagano (1978) and Troutman (1979). McLeod and Hipel (1978) give asymptotic distribution results about residuals and Portmanteau statistics. Dunsmuir (1981) studies asymptotic distributions in the case of a process with periodic means and periodic standard deviations. Results for aggregated PAR process are given in Tiao and Gruppe (1980) and by Vecchia (1984).

An interesting application of PAR models is given by Newton (1982). He estimates the spectral density of a multivariate process by using a multivariate time series approximation in the spirit of Parzen (1964) and using the equivalent PAR parametrization of the same process. The PAR specification of the process needs less parameters than the multivariate specification the process. Other papers in the same spirit are by Cipra (1985a,b), Cipra and Tlustý (1987) who use the equivalence between periodic models and multivariate ARMA models to fit the latter. The fitting procedure that they use is based on the Yule-Walker equations together with the generalization of these equations to moving average models as in Durbin (1964).

Forecasting using PAR models was considered by Noakes et al. (1985) who performed a forecasting competition. Several methods were considered such as Box-Jenkins models, exponential smoothing, seasonal means etc.. The PAR models provided a feasible alternative to these other models due to their superior forecasting capabilities for monthly riverflows series.

Finally, time varying periodic models have been intensely used by researchers interested in hydrological time series process, specially monthly river flows. Models that fall in this category are given by Yevjevich (1972); Rao and Kashyap (1974); Delleur et al. (1976); Croley and Roa (1978).

3.3 A basic property of PARMA models

A fundamental property is that the PARMA process $X_{(t,i)}$ is related to the multivariate ARMA process $\underline{X}_t$ by the following equations.
First we write (3.2) as

$$\begin{aligned} x_{(t,i)} = \phi_{i1}x_{(t,i)-1} & + \cdots + \phi_{(t,i)-p(i)}x_{(t,i)-p(i)} + \\ & \theta_{i1}a_{(t,i)} + \cdots + \theta_{(t,i)-q(i)}a_{(t,i)-q(i)}. \end{aligned} \tag{3.3}$$

Then, define the $S \times S$ matrices α_k and Ω_k as:

$$\begin{aligned} \alpha_k &:= (-\phi_{(k,i)-j}), k = 0, \ldots, p(k). \\ \Omega_k &:= (\theta_{(k,i)-j}), k = 0, \ldots, q(k). \end{aligned} \tag{3.4}$$

Note that α_0 is a lower triangular unit matrix. With these definitions we can rewrite the process $X_{(t,s)}$ as

$$\alpha_0 \underline{X}_t - \alpha_1 \underline{X}_{(t-1)} - \cdots - \alpha_p \underline{X}_{(t-p)} = \Omega_0 \underline{a}_t - \cdots - \Omega \underline{a}_{t_q}. \tag{3.5}$$

i.e., $\underline{X}_t$ is a multivariate ARMA$(\underline{P}, \underline{Q})$ process that satisfies the following equation:

$$\underline{X}_t = \alpha_1^{\#} \underline{X}_{(t-1)} + \cdots + \alpha_p^{\#} \underline{X}_{(t-p)} + e_{(t-1)} - \cdots - \Omega_q^{\#} e_{(t-q)}. \tag{3.6}$$

where $\alpha_i^{\#} := \alpha_0^{-1}\alpha_i$; $\Omega_i^{\#} := \alpha_0^{-1}\Omega_i$; $\text{var}(e_t) = \alpha_0^{-1}\Omega_0 \text{diag}(\sigma_a^2, \ldots, \sigma_S^2)\Omega_0^T \alpha_o^{-T}$. Also, a multivariate ARMA process can be represented as a PARMA process following the above reasoning in the reverse order. This result has been proven by Pagano (1978) for the PAR case and Vecchia (1985) for the PARMA case.

REMARK 3.1 (a) The above correspondence can be extended to multivariate PARMA processes simply viewing the $\phi's$ and $\theta's$ as matrices.

(b) The above result could be seen as implying that PARMA models are a meaningless generalization as they are just multivariate models. However, in many situations it is physically more meaningful to specify PARMA models, instead of multivariate models. Moreover, one of the most interesting uses of PARMA models is precisely as an alternative parametrization of multivariate ARMA models. They require less parameters in general.

(c) The correspondence between multivariate and PARMA models could be exploited as in Pagano (1978), Troutman (1979), and Vecchia (1985) to derive conditions that the parameters should satisfy in order to assume that the multivariate time series is stationary and thus has a moving average representation. In that case the PARMA model could be represented as an infinite moving average model but with periodic coefficients. Hence, PARMA parametrizations of multivariate processes seem to be an interesting and, in many situations have a physical justification for issues concerning identification, estimation, forecasting and control.

3.4 State-space representation

The state-space representation is most important for estimation and forecasting of time series, even if there is not a physical model with which we could relate the state variables.

We could write the PARMA model as

$$\begin{aligned} x_{(t,i)} - a_{(t,i)} &= \phi_{(i,1)}x_{(t,i)-1} - \theta_{(i,1)}a_{(t,i)-1} + \\ &\quad + \phi_{(i,2)}x_{(t,i)-2} - \theta_{(i,2)}a_{(t,i)-2} + \\ &\quad \vdots \\ &\quad + \phi_{(i,r(i))}x_{(t,i)-r(i)} - \theta_{(i,r(i))}a_{(t,i)-r(i)}, \end{aligned} \tag{3.7}$$

where $r(i) = \max\ (p(i), q(i)+1)$, and we make the convention that parameters not defined are equal to zero. We can write (3.7) as

$$\begin{aligned} x(t,i) &= \phi_{(i,1)}x_{(t,i)-1} + a_{(t,i)} + z_{(t,i)-1} \\ z_{(t,i)-1} &= \phi_{(i,2)}x_{(t,i)-2} + \theta_{(i,1)}a_{(t,i)-1} + z_{(t,i)-2} \\ \vdots &= \vdots \\ z_{(t,i)-r(i)+1} &= \phi_{(i,r(i))}x_{(t,i)-r(i)} + \theta_{(i,r(i)-1)}a_{(t,i)-r(i)}. \end{aligned} \tag{3.8}$$

Denote by $S_{(t,i)}$ the state variable at time (t,i), which satisfies:

$$S(t,i) = F_{(i)}S_{(t,i)-1} + G_{(i)}a_{(t,i)}, \tag{3.9}$$

where

$$F_{(i)} = \begin{pmatrix} \phi_{(i,1)} & 1 & 0 & \dots & 0 \\ \phi_{(i+1,2)} & 0 & 1 & \dots & 0 \\ \vdots & 0 & \dots & \ddots & 1 \\ \phi_{(i+r(i)-1} & 0 & 0 & \dots & 0 \end{pmatrix}. \qquad (3.10)$$

and

$$G_{(i)} = \begin{pmatrix} \theta_{(i,0)} \\ \theta_{(i+1,1)} \\ \vdots \\ \theta_{(i+r(i)-1,r(i)-1)} \end{pmatrix}. \qquad (3.11)$$

Then $x_{(t,i)}$ is the first coordinate of $S_{(t,i)}$,

$$x_{(t,i)} = hS_{(t,i)}; \qquad h = (1,0,\dots,0)'. \qquad (3.12)$$

Equations (3.9–3.11) give one of the possible alternative representations of the state space equations for PARMA models. These are the so called observability equations. Other forms could be obtained from the multivariate representation of PARMA models. For example, we could stack up the state vectors to form

$$S_{(t)} = (S'_{(t,1)}|, S'_{(t,2)}| \dots |S'_{(t,s)})' \qquad (3.13)$$

and then $S_{(t)}$ would evolve according to

$$\begin{aligned} S_{(t)} &= FS_{(t-1)} + G\underline{a}_t, \\ \underline{X}_{(t)} &= HS_{(t)}, \end{aligned} \qquad (3.14)$$

where F is a block diagonal matrix with the diagonal formed by the matrices $F_{(i)}$. H is formed by glueing together S h-vectors. Note that F, G and H are constant matrices.

Another useful state space representation can be obtained using the state vector $\underline{r}_{(t,i)}$ that is defined by the following evolution equation:

$$\begin{aligned} \underline{r}_{(t,i)} &= U_{(i)}\underline{r}_{(t,i)-1} - V_{(i)}a_{(t,i)} \\ x_{(t,i)} &= h\underline{r}_{(t,i)}, \end{aligned} \tag{3.15}$$

where the matrices $U_{(i)}$ and $V_{(i)}$ are defined as

$$\begin{aligned} U_{(i)} &= \begin{pmatrix} \phi_{(i,1)} & \cdot & \cdots & & \phi_{(i,p(i))} \\ 1 & 0 & \cdots & & 0 \\ 0 & 1 & & & 0 \\ \vdots & \cdots & \ddots & & 0 \\ 0 & \cdots & & 1 & 0 \end{pmatrix} \\ V_{(i)} &= \begin{pmatrix} 1 & -\theta_{(i,1)} & \cdots & -\theta_{(i,q(i))} \end{pmatrix} \end{aligned} \tag{3.16}$$

and

$$\underline{a}_{(t,i)} = \begin{pmatrix} a_{(t,i)} \\ a_{(t,i-1)} \\ \vdots \\ a_{(t,i-q(i))} \end{pmatrix}.$$

Finally, referring to (3.7) we could define the following state space equations:

$$\begin{aligned} \underline{\alpha}_{(t,i)} &= F_{(i)}\underline{\alpha}_{(t,i-1)} + R_{(i)}\underline{a}_{(t,i-1)}, \\ x_{(t,i)} &= h\underline{\alpha}_{(t,i)} + a_{(t,i)} \end{aligned} \tag{3.17}$$

where $r(i) = \max(p(i), q(i))$, $F_{(i)}$ is defined as before and

$$R_{(i)} = \left(\phi_{(i,1)} - \theta_{(i,1)}, \ldots, \phi_{i+r(i)-1,r(i)} - \theta_{i+r(i)-1,r(i)}\right). \tag{3.18}$$

These state space representations are useful to obtain conditions that imply seasonal stationarity and invertibility. To derive some of these conditions we will assume that $\boldsymbol{E}(log_{+}|a_{(t,i)}|) < \infty$. Then we have the following result.

THEOREM **3.2** The series $\{x_{(t,i)}\}$ is S-dimensional stationary if and only if one of the following equivalent conditions is satisfied:

(a) The roots of $\det(\alpha_0 - \alpha_1 B - \ldots - \alpha_p B^p)$, considered as a polynomial in B, are outside the complex unit circle. (Refer to (3.5)).

(b) The right eigenvalues of

$$M = F_{(S)} \ldots F_{(1)} \tag{3.19}$$

are strictly less than 1 in modulus. (Refer to (3.9)).

(c) The right eigenvalues of

$$N = U_S \ldots U_1. \tag{3.20}$$

are strictly less than 1 in modulus. (Refer to (3.4)).

∇

Proof S-dimensional stationarity is equivalent to the existence of coefficients $(\psi_{(k,i)}, k \in \mathbb{N}, i \in S)$, such that

$$x_{(t,i)} = \sum_{k \in \mathbb{N}} \psi_{(k,i)} a_{(t,i)-k}. \tag{3.21}$$

To prove that the above representation is valid we proceed as follows. For each of the state representations (3.5),(3.9),(3.17) after L recursive substitutions the process $x_{(t,i)}$ can be written as

$$x_{(t,i)} = \sum_{k=0}^{L} \psi_{(k,i)} a_{(t,i)-k} \quad + + \sum_{k=L+1}^{N+p} \psi_{(k,i)}(L) x_{(t,i)-k} + \sum_{k=L+1}^{N+q} \pi_{(k,i)}(L) a_{(t,i)-k} \tag{3.22}$$

Now, the coefficients $\psi_{(k,i)}(L)x_{(t,i)-k}$ and $\pi_{(k,i)}(L)a_{(t,i)-k}$ arise for example from representation (3.9) as coefficients of matrices involving powers of the matrix M above. Therefore, these coefficients will tend geometrically fast to zero if the right eigenvalues of the matrix M are less than one in modulus. Similarly, for the other representations. Now, if

$E(log_+|a_{(t,i)}|) < \infty$, then

$$\sum_{j=0}^{\infty} t^j a_{(t,i)-j} \tag{3.23}$$

is almost sure convergent (Stout 1974) and hence

$$x_{(t,i)} = \sum_{k=0}^{\infty} \psi_{(k,i)} a_{(t,i)-k} \qquad \text{a.s.} \tag{3.24}$$

□

REMARK 3.3 (a) Note that seasonal stationarity is present even when the roots of $\phi_i(B)$ are not within the unit circle for some i. If this fact is compensated for by subsequent small roots of $\phi_j(B)$, this could be seen as an explosive process for some seasons and as a damping process for others. An example where this behaviour could occur is monthly river flows.

(b) Similar conditions to the ones given above are required for invertibility of the process.

(c) As remarked before, the extension to the multivariate case is just a matter of regarding each quantity as a matriv instead of as a scalar.

PROPOSITION 3.4

(a) After repeated substitution, using the state representation (3.5) we obtain:

$$\begin{aligned} S_{(t,i)} &= F_{(i)} S_{(t-1,i)} - v_{(t,i)} \\ v_{(t,i)} &= \textstyle\sum_{l=0}^{S-1} (\prod_{k=0}^{l} F_{(i-k)}) \theta_l a_{(t,i)-l}. \end{aligned} \tag{3.25}$$

where

$$F_{(i)} = F_i \ldots F_{i-S+1}.$$

This gives the state space equations for the subseries corresponding to the i season with a lag interval equal to the number of seasons.

(b) If $F_{(i-1)}$ is invertible then:

$$\begin{aligned} F^{-1}_{(i+1)} S_{(t,i+1)} &= F^{-1}_{(i+1)} S_{(t-1,i+1)} + F^{-1}_{(i+1)} v_{(t,i+1)}. \\ &= F^{-1}_{(i+1)} S_{(t-1,i+1)} - W_{(t,i+1)}. \end{aligned} \tag{3.26}$$

Define $U_{(t,i+1)} = F^{-1}_{(i+1)} S_{(t,i+1)}$, then

$$U_{(t,i+1)} = F_{(i)} U_{(t-1,i)} - W_{(t,i+1)}. \tag{3.27}$$

▽

This result shows the equation for the state variable corresponding to the $i + 1$ season has the same autoregressive component as the equation for the state variable corresponding to the i season. The moving average components are different and the observational equations are different too. This implies that the autoregressive part of the ARMA equation implied by the state space model is the same for both seasons. Note that $F_{(i+1)}$ is invertible if $r(i + 1) = r(i)$ and $\phi_{i+r(i+1)-1,r(i+1)-1} \neq 0$. This result could be useful in identification (Vecchia 1985).

THEOREM 3.5 Let $\{x_t\}$ be a time series process. Assume that x_t is governed by the following evolutionary equations: $\phi_t(B)x_t = \theta_t(B)a_t$. Then, if $\{x_t\}$ can be represented by an infinite moving average $x_t = \Psi_t(B)a_t$, the coefficients $\{\psi_{t,i}, 0 \leq i < \infty\}$ are given by the following recursive equations:

$$\psi_{t,k} = \phi_{(t,1)}\psi_{t-1,k-1} - \cdots - \phi_{(t,p(t))}\psi_{t-p(t),(k-p(t))} - \theta_{t,k}. \qquad (3.28)$$

The terms not defined by the evolution equations are taken to be zero A similar result holds for the infinite autoregressive representation. ∇

Proof The result follows immediately from the formal equations $\Psi_t(B)\phi_t(B) = \theta_t(B)$ after equating the coefficients of the same powers of B on both sides of the equation. However note that the shift operator B in $\Psi_t(B)$ has to be applied both to B and t in $\phi_t(B)$.

□

REMARK 3.6 The $\{\psi_{t,i}, 0 \leq i < \infty\}$ are useful to find the autocovariances because the latter can be written as

$$\boldsymbol{E}(x_t, x_s) = \sum_{0}^{\infty} \psi_{t,k}\psi_{s,k+s}. \qquad (3.29)$$

3.5 Autovariance Structure

The autocovariance function $\gamma_{(t,t+d)}$ of PARMA models is periodic, and hence $\gamma_{(t,t+d)} = \gamma_{(t+S,t+d+S)}$. Also, it satisfies the following

Yule-Walker equations

$$\gamma_{(k,k-i)} - \sum_{0}^{p(k)} \psi_{(s(k)j)} \gamma_{(k-j,k-i)} = \mu_{(k,k-i)}, \tag{3.30}$$

and

$$\mu_{(k,k-i)} = \begin{cases} 0 & \text{if } i >_s s(k) \\ \sum_0^{q(k)} \theta_{(s(k)j)} \psi_{(s(k)-i,j-i)}, & \text{otherwise.} \end{cases} \tag{3.31}$$

The $\psi_{(i,j)}$'s could be found from the infinite moving average expansion

$$x_i = \sum_{0}^{\infty} \psi_{(i,j)} a_{(i-j)}. \tag{3.32}$$

To obtain $\psi_{(i,j)}$ we equate the coefficients in

$$\Psi_i(B)\phi_i(B) = \theta_i(B) \tag{3.33}$$

where $\Psi_i(B), \phi_i(B), \theta_i(B)$ are the generating functions associated the infinite moving average, the autoregressive part and the moving average component of the ith-season ARMA model. The $\Psi_i(B)$'s can be found recursively from (3.28). These equations are well known for the ARMA case (McLeod, 1975).

The relations (3.31) may be used ~~to~~ estimate the order of the process. For each season ν and for each value of $(p(\nu), q(\nu))$, we solve the system of linear equations given by (3.31) with $i = \nu, \ldots, \nu + p(\nu)$ using the usual estimates for the covariances. Another posibility is to use relation (3.31) with $i \leq \nu$ and solve the resulting overdetermined system of linear equations. Also, we could filter the sample autocorrelation function as in Menard and Roy (1986).

The estimates of μ are called the sample extended autocovariance function of the process. If $(p(\nu), q(\nu))$ goes from 0 to m a table that should exhibit a characteristic triangular cut-off pattern is obtained, similar to the sample extended autocovariances as defined by Tsay and Tiao (1984). The basic problem is the unknown variance of the estimates.

Another important identification technique relies on the use of the partial autocorrelation of the process. Sakai (1982) gave the appropriate extension of the partial autocorrelations to the periodic case. It is as follows. Define

$$\Gamma_{(t,r)} = (\gamma_{(i,j)})_{i,j=t}^{i,j=r}. \tag{3.34}$$

Next, let $\alpha_{(t,r)}$ be equal to the first column of the inverse of $\Gamma_{(t,r)}$, normalized to have its first coefficient equal to one. Similarly, define $\beta_{(r,t)}$ to be the first row of the inverse of Γ normalized to have its last coefficient equal to one. Then we define the backwards innovations $\epsilon_{(t,r)}$ equal to

$$\epsilon_{(t,r)} = (x_t, \ldots, x_r)\alpha_{(t,r)} \tag{3.35}$$

the forward innovations $\xi_{(r,t)}$ of the process are equal to

$$\xi_{(r,t)} = (x_t, \ldots, x_r)\beta_{(r,t)}. \tag{3.36}$$

But $\alpha_{(t,r+1)}$ and $\beta_{(r+1,t)}$ may be found recursively in terms of a linear combination of $(\alpha_{(t,r)}, 0)$ and $(0, \beta_{(r+1,t+1)})$. Call Ω the product of $(\alpha_{(t,r)}, 0)$ and the last row of $\Gamma_{(t,r+1)}$. Then Ω is also equal to the product of $\Gamma_{(t,r+1)}$ and $(\beta_{(r+1,t+1)})$. Call $\sigma^2(t,r)$ the variance of $\epsilon_{(t,r)}$. Let $\delta^2(r,t)$ be the variance of $\beta_{(r,t)}$. Then the following relations are valid

$$\begin{aligned} \sigma^2(t, r-1) &= \sigma^2(t,r) - \Omega^2\delta^2(r-1, t-1) \\ \delta^2(r-1, t) &= \delta^2(t-1, r-1) - \Omega^2\sigma^2(t,r), \end{aligned} \tag{3.37}$$

and

$$\begin{aligned} \epsilon_{(t,r-1)} &= (\epsilon_{(t,r)}, 0) - \Omega)/\delta(r-1, t-1)\sigma^2(r-1, t-1) \\ \xi(r-1, t) &= (\epsilon_{(t,r)}, 0) - \Omega^2\sigma^2(t,r)\alpha((t,r), 0). \end{aligned} \tag{3.38}$$

Equations (3.37) to (3.38) are the periodic version of the Levinson-Durbin recursions to obtain the Cholesky decomposition of the inverse of the autocovariance matrix of the process. They have been given by Sakai (1982). They can be used as an aid to identify the process under consideration. Note that if $\sigma^2(t, r-1) = \sigma^2(t,r)$ then $\Omega(t, r+1) = 0$, and hence $\delta(r+1, t) = \delta(r-1, t+1)$ and
$\alpha(t, r+1) = \alpha(t,r)$, $\beta(r+1, t) = \beta(r+1, t-1)$ and $\Omega(t+S, r+1+S) = 0$. Hence we have a periodic cutoff to indicate an autoregression.

REMARK 3.7 Define the seasonal partial correlations by

$$\rho(t,r) = \frac{\Omega(t,r)}{\sigma(t,r)\delta(r,t)}. \tag{3.39}$$

These analogues to the univariate partial correlations are the correlation between $\epsilon_{(t,r+1)}$ and $\xi_{(r+1,t)}$. They can be used in identification because they are periodic with the cut-off property just described.

The Levinson-Durbin recursions are also useful for obtaining an alternative algorithm to the Kalman filter described in section 3.4. The idea is that the innovations, either forward or backward innovations, are in a one to one correspondence with the observations and by construction they are orthogonal to each other. Hence, they are independent because we are assuming normality. Thus, the construction of the likelihood corresponding to the innovations is quite simple. However, to use this algorithm the autocovariance function need to be computed. To obtain this autocovariance matrix, we could use the analog to the algorithms given by McLeod (1975; 1977). An important point to speed up the construction of the likelihood is to observe that the autocovariances will either damp-off exponentially fast or exhibit a cut-off point (this follows easily from the multivariate representation of PARMA processes). In the first case, we could after some point extrapolate the damping constant to the limit. In the second case, we could set the remaining autocovariances to-zero. Similar comments are applicable to the Levinson-Durbin algorithm.

Another important consideration is that the nonlinear minimization algorithm that has to be used to estimate the parameters is iterative. This means that we could use the last evaluation of the autocovariance as the initial value for an iterative algorithm to evaluate the autocovariance. One possibility to make this idea feasible could be to change the parameters a season in each iteration. Then, the autocovariance matrix corresponding to the new parameters will differ from the old autocovariance matrix by only the rows and columns that corresponds to the season whose parameters has been changed. Finally, note that the algorithms that have been presented give the true likelihood and that is not true for the algorithms that use an approach similar to the one of Hillmer and Tiao (1979).

3.6 Algorithms to compute the autocovariance function

We will give two algorithms for computing the autocovariance matrix of a PARMA process. The first algorithm is similar to Mcleod's (1975) algorithm for computing the autocovariance matrix of a ARMA process.

First algorithm

We will write the PARMA process equations as

$$x_t + \phi_{t,1}x_{t-1} + \cdots + \phi_{t,p}x_{t-p} = a_t - \theta_{t,1}a_{t-1} - \cdots + \theta_{t,q}a_{t-q}. \tag{3.40}$$

where we have assumed for simplicity that all seasons have an autoregressive component of order p and a moving average component of order q.

Multiplying both sides of (3.40) by x_{t-k} and taking expectations we obtain

$$\begin{aligned}\gamma(t,t-k) + \phi_{t,1}\gamma(t-1,t-k) + \cdots + \phi_{t,p}\gamma(t-p,t-k) = \\ \gamma^{ay}(t,t-k)\theta_{t,1}\gamma^{ay}(t-1,t-k) + \cdots + \theta_{t,q}\gamma^{ay}(t-q,t-k),\end{aligned} \tag{3.41}$$

for $k = 0,\ldots,p$, and $t = S,\ldots,1$, where $\gamma(t,t-k) = \boldsymbol{E}\{y_t y_{t-k}\}$, and $\gamma^{ay}(t,t-k) = \boldsymbol{E}\{a_t y_{t-k}\}$. This gives a system of linear equations once the right hand side in (3.41) is known. To obtain the right hand side of (3.41) multiply both sides of equation (3.40) by a_{t-k}, and take expectations

$$\gamma^{av}(t,t-k)+\phi_{t,1}\gamma^{av}(t-1,t-k)+\cdots-\phi_{t,p}\gamma^{av}(t-p,t-k)=\theta_{t,k}\sigma^2_{t-k}, \quad (3.42)$$

for $k=0,\ldots,p$, and $t=S,\ldots,1$. Many times these systems of linear equations are sparse, i.e. the associated matrix is full of zeros. We could use then algorithms for solving sparse systems of linear equations.

Second algorithm

First, we will consider a PAR model and assume that for all seasons the order of the autoregression is the same. If necessary set the ϕ's equal to zero. Write the process equations as

$$\begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \ldots & \phi_p \\ 1 & \ldots & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ . \\ \vdots \\ y_{t-p} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} a_t \quad (3.43)$$

and let $S_t=(y_t\ y_{t-1}\cdots y_{t-p})'$. Then, we can write (3.43) as

$$S_t=\Phi_t S_{t-1}+G_t a_t. \quad (3.44)$$

For $k>1$

$$\boldsymbol{E}\{S_t S^T_{t-k}\}=\Phi_t\boldsymbol{E}\{S_{t-1}S^T_{t-k}\}, \quad (3.45)$$

and

$$\boldsymbol{E}\{S_t S^T_t\}=\Phi_t\boldsymbol{E}\{S_{t-1}S^T_{t-1}\}\Phi^T_t+\sigma^2_t G_t G^T_t \quad (3.46)$$

Using (3.46) recursively we obtain

$$\boldsymbol{E}\{S_t S^T_t\}=\Phi\boldsymbol{E}\{S_{t-s}S^T_{t-s}\}\Phi^T+\sum_{k=0}^{s-1}\left(\prod\Phi_{t-k}\right)g_t G^T_t\left(\prod\Phi_{t-k}\right)'. \quad (3.47)$$

But using the periodic recurrence of the process we obtain

$$\Gamma_t = \Phi \Gamma_t \Phi^T - \Omega_t \tag{3.48}$$

where $\Gamma_t = \mathbb{E}\{S_t S_t^T\}$ and

$$\Omega_t = \sum_k (\prod \Phi_i) G_k G_k^T (\prod \Phi_i)^T. \tag{3.49}$$

REMARK 3.8 (a) Note that once Γ_t is known then $\mathbb{E}\{S_t S_{t-k}^T\}$ can be found using (3.37) recursively. Also, we do not need to perform the complete multiplication. We simply have to find recursively the first rows for each matrix.

(b) Vectorize the equation (3.48) as

$$\text{vec}(\Gamma_t) = (\Phi \otimes \Phi^T)\text{vec}(\Gamma_t) + \text{vec}(\Omega_t). \tag{3.50}$$

Then

$$(I - \Phi \otimes \Phi)\text{vec}(\Gamma_t) = \text{vec}(\Omega_t). \tag{3.51}$$

To solve (3.51) involves $O((p+1)(p+2)/2^3)$ operations. However, we just need to solve this system for $\mathbb{E}\{S_t S_t^T\}$ in order to find $\mathbb{E}\{S_{t+1} S_{t+1}^T\}$. We can use (3.46) and the fact that the submatrix that consists of the last $(p-1)$ rows and $(p-1)$ columns of $\mathbb{E}\{S_{t+1} S_{t+1}^T\}$ is equal to the submatrix. This is the form obtained from the first (p-1) rows and (p-1) columns of $\mathbb{E}\{S_t S_t^T\}$. Hence, we need to employ (3.46) just to find the first row of Γ_{t+1}.

For the case of a PARMA system we have instead of (3.44)

$$S_t = \Phi_t S_{t-1} + G_t \underline{a}_t, \tag{3.52}$$

where S_t andΦ_t are defined as before but $\underline{a}_t = (a_t, a_{t-1}, \ldots, a_{t-q})'$ and

$$G_t = (1 - \theta_{t1} \cdots - \theta_{tq}). \tag{3.53}$$

Define Γ_t as before, $\Gamma_{t,r} = \mathbb{E}\{S_t S_r^T\}$ and $\Psi_{t,r} = \mathbb{E}\{\underline{a}_t y_r^T\}$. Then

$$\Gamma_t = \Phi_t \Gamma_{t-1} \Phi_t^T + G_t D_t G_t^T + \Phi_t \Psi_{t,t} + \Psi_{t,t}^T \Phi_t^T, \tag{3.54}$$

and for $k > 1$

$$\Gamma_{t,t-k} = \Phi_t \Gamma_{t-1,t-k} - G_t \Omega_{t,t-k}. \tag{3.55}$$

As before, using (3.54) recursively we obtain

$$\Gamma_t = \Phi \Gamma_{t-s} \Phi + \alpha_t, \tag{3.56}$$

but $\Gamma_t = \Gamma_{t-s}$ by periodicity. Then

$$\Gamma_t = \Phi \Gamma_t \Phi^T + \alpha_t, \tag{3.57}$$

and we can proceed as before. However, to obtain α_t we need to compute $\Psi_{t,r}$. Note that its components can be obtained from the infinite moving average expansion using the recursion (3.28).

3.7 Inverse autocorrelations and inverse partial autocorrelations

The inverse autocorrelation function, IACF, of periodic models will be defined as the autocorrelation function of the 'dual' process

$$\theta_t(B) Y_t = \phi_t(B) \alpha_t, \tag{3.58}$$

i.e., interchange the roles of ϕ_t and θ_t for each season. Note that this is not precisely the generalization of the univariate IACF to periodic models, because the IACF in the univariate case is defined as the sequence of coefficients of the powers of z of $\Xi^{-1}(z)$, where $\Xi(z)$ is the autocorrelation generating function of the process. The interest of the IACF in the ARMA case is that it will exhibit 'inverse' properties to the ACF, i.e. it will cut-off for an autoregressive process and it will damp-off for a moving average one. We will investigate the properties of the IACF for periodic models. Assume that for the tth season we are dealing with an autoregressive process

$$\phi_t(B)y_t = a_t. \tag{3.59}$$

We will consider then the process

$$z_t = \phi_t(B)a_t. \tag{3.60}$$

We assume that the order of autoregression is $p(t)$, then

$$\mathbb{E}\{z_t z_{t-k}\} = \mathbb{E}\{\phi_t(B)a_t)z_{t-k}\} = 0, \qquad \text{for } k > p(t). \tag{3.61}$$

Because z_{t-k}, $k > p(t)$, is an infinite moving average, then

$$z_{t-k} = \sum_{j=0}^{\infty} \psi_{t-k,j} a_{t-k-j}, \tag{3.62}$$

and hence the orthogonality of a_t sequence implies the result.

Note that we have not assumed anything about the type of model other seasons follow. Now, assume the process $\{y_t\}$ follows a moving average

$$y_t = \theta(B)a_t. \tag{3.63}$$

We will consider

$$\theta_t(B)z_t = a_t. \tag{3.64}$$

Then $\boldsymbol{E}\{z_t z_{t-k}\}$ does not follow a well defined pattern except that for each season $\boldsymbol{E}\{z_t z_{t-k_s}\}$ decays exponentially to zero. Now, consider the periodic partial autocorrelation function for this model. Assume the order of the moving average is $q(t)$. Then because of Remark (3.7) the inverse partial autocorrelation function exhibits the cut-off property after lag $q(t)$.

Estimation of the inverse autocorrelation

If we approximate the evolution equations by a 'long' autoregression

$$\Pi_t(B)y_t = a_t, \tag{3.65}$$

we obtain an estimate $\hat{\Pi}_t(B)$ based on the data. Then sample IACF and sample IPACF are the ACF and PACF, respectively of

$$z_t = \hat{\Pi}_t(B)a_t. \tag{3.66}$$

Then

$$\boldsymbol{E}\{z_t z_{t-k}\} = \sum \hat{\Pi}_{t,j}\hat{\Pi}_{t-k,j-k}. \tag{3.67}$$

3.8 Estimation

We will assume that the random input $a_{(t,i)}$ is Gaussian noise with periodic variance σ_i^2. Consider the innovations

$$\epsilon_{(t,i)} = x_{(t,i)} - \hat{x}_{(t,i)}(Past), \tag{3.68}$$

where *Past* refers to the finite available observations up to the time (t, i), and $\hat{x}_{(t,i)}(Past)$ is the projection onto the linear space generated by *Past*. Because of the Gaussian assumption, $\{\epsilon_{(t,i)}\}$ is a collection of normal and independent Gaussian random variables. Also, the innovations and the observations are in a one-to-one correspondence with unit Jacobian. Hence, the loglikelihood $\mathcal{L}$ of the observations is given up to a constant by

$$\mathcal{L} = \sum_{(t,i)}^{N} \frac{\epsilon_{(t,i)}^2}{\gamma_{(t,i)}\sigma_i^2} + \sum_{(t,i)}^{N} log\gamma_{(t,i)}\sigma_i^2. \tag{3.69}$$

Generally, we work with the concentrated log likelihood obtained by

$$\hat{\sigma}_i^2 = \frac{1}{y} \sum_{(t,i)}^{N} \frac{\epsilon_{(t,i)}^2}{\gamma_{(t,i)}}. \tag{3.70}$$

Then,

$$\mathcal{L} = \sum log\gamma_{(t,i)} + \sum log\hat{\sigma}_i^2. \tag{3.71}$$

To obtain $\epsilon(t, i)$ and $\gamma_{(t,i)}$ several methods are possible. For example, we could obtain the autocovariance matrix and the Cholesky decomposition of its inverse; or we could use a Box-Jenkins type recursion. However, it seems that there is a growing agreement that methods based on Kálman filtering (KF) are the most suitable (Morf et al. 1974; Gardner et al.

1980; Perlman, 1980; Ansley and Kohn 1983; Kohn and Ansley 1982 ; Shea, 1987). Kalman filtering has many desirable properties. One of them is that it is simple to modify its standard implementation in order to work with missing data. We remark that if we work with the multivariate representation of PARMA models we could work with an algorithm as given in Shea (1987). However, we believe that it is more suitable to work directly with one of the state-space representations given in the former section. The reason is that it is possible to take advantage of fact that for each season PARMA models behave just like ARMA processes with respect to computation, except for the fact that we have to change for each season the "evolution" matrix $F_{(i)}$. Instead, when we use the multivariate form the matrices lose many of their zeroes. Note also that it is simpler to obtain the initial conditions in the PARMA state space representation. The main drawback is that we cannot use the Chandrashekar equations as is usually done for constant coefficient ARMA models. Instead, we have to work with the traditional KF. It seems to us that this is not a great disadvantage in this case. Note also the large number of parameters involved in a full multivariate model.

Algorithm:

We consider the state-space representation given by

$$\begin{aligned} S_{(t,i)} &= F_{(i)}S_{(t,i)-1} + G_{(i)}a_{(t,i)} \\ x_{(t,i)} &= hS_{(t,i)} \\ h &= (1,0,\ldots,0)'. \end{aligned} \tag{3.72}$$

where $F_{(i)}, G_{(i)}$ and h are given by (3.10), (3.11) and (3.12), respectively.

The KF recursive equations are given by Jaswinski (1970)

$$
\begin{aligned}
R_k &= hP_k^{(k-1)}h' \\
\epsilon_k &= x_k - hS_k^{(k-1)} \\
K_k &= P_k^{(k-1)}h'R_k^{-1} \\
S_k^k &= S_k^{(k-1)} + K_k\epsilon_k \cdot \\
P_k^k &= P_k^{(k-1)} - K_k h P_k^{(k-1)} \\
S_{k+1}^k &= F_k S_k^k \\
P_{k+1}^k &= F_k P_k^k F_k' + G_k G_k .
\end{aligned}
\tag{3.73}
$$

With initial conditions S_1, P_1^1, all the indices are understood in modulus S arithmetic, when applicable. The special structure of periodic models implies that

$$
\begin{aligned}
R_k &= P_k^{k-1}(1,1) \\
\epsilon_k &= x_k - S_k^{k-1}(1) \\
K_k &= P_k^{k-1}(1,.)/P_k^{k-1}(1,1) \\
S_k^k &= S_k^{k-1} + K_k\epsilon_k \\
P_k^k &= \begin{pmatrix} & O & \\ O & & A_k \end{pmatrix} \\
\epsilon_k(i,j) &= P_k^{k-1}(i+1,j+1) - P_k^{k-1}(1,j+1)P_k^{k-1}(1,i+1)P_k^{k-1}(1,1) \\
S_{k+1}^k &= \begin{pmatrix} \phi_{(s(k),1)}S_k^k(1) + S_k^k(2) \\ \vdots \\ \phi_{(s(k),r-1)}S_k^k(1) + S_k^k(p-1) \\ \phi_{(s(k),1)}S_k^k(1) \end{pmatrix} \\
P_{k+1}^k &= \begin{pmatrix} A_k & O \\ O & \end{pmatrix} + G_{s(k)}G'_{s(k)} .
\end{aligned}
\tag{3.74}
$$

REMARK 3.9 Evaluation of ϵ_k needs $2r^2 + r$ multiplications and $2r^2 + 3r$ additions. These can be reduced if not all the modes are of the same order for each new iteration as our initial guess. It has been our experience that, at most, five iterations should generally suffice to achieve convergence.

Initial conditions.

The initial values are obtained as follows. Let $S_1^1 = 0$, and P_1^1 is the solution of the following equation

$$P = FPF' + A \tag{3.75}$$

where $F = F_S \ldots F_1$, and

$$A = \sum_{j=0}^{S-2} (\Pi_{k=S+1-j}^{S} F_k) G_j G_j' (\Pi_{k=S+1-j}^{S} F_k)' \sigma_j^2. \tag{3.76}$$

To prove (3.76) we could use (3.72) S times and then make use of the periodicity of the process.

REMARK 3.10 The solution of (3.75) is given by

$$P = \sum F^i A(F^i)'. \tag{3.77}$$

However, in our computer implementation we used the interaction

$$P^{k+1} = FP^k F' + A. \tag{3.78}$$

until a convergence criterion based on the size of the Frobenious norm of the matrix $P_{(k)}$ was satisfied. The reason for using this recursion is that the minimization procedure is iterative and if the parameters between successive iterations are reasonably close we could expect also the matrices P_1^1 to be reasonable close. Hence, we used the value of P_1^1 obtained in the last iteration as the initial value of the matrix $P_{(k)}$ in the iteration (3.78).

3.9 The information matrix

To obtain the information matrix

$$I = \frac{1}{N}\mathrm{var}\left(\frac{\partial \log \mathcal{L}}{\partial \underline{\gamma}}\right), \tag{3.79}$$

we will assume the approximation

$$I \approx \frac{1}{S}\boldsymbol{E}\frac{1}{\sigma_\nu^2}\left(\sum_{\nu=1}^{S}\left(\frac{\partial a_{(t,\nu)}}{\partial \underline{\gamma}}\right)\left(\frac{\partial a_{(t,\nu)}}{\partial \underline{\gamma}}\right)'\right). \tag{3.80}$$

But

$$\begin{aligned} \partial a_{(t,\nu)}/\partial\phi_{(\nu,j)} &= -x_{(t,\nu-j)} + \theta_\nu(B)^\cdot \partial a_{(t,\nu-1)}/\partial\phi_{(\nu,j)} \\ \partial a_{(t,\nu)}/\partial\theta_{(\nu,j)} &= -a_{(t,\nu-j)} + \theta_\nu(B)^\cdot \partial a_{(t,\nu-1)}/\partial\theta_{(\nu,j)}, \end{aligned} \tag{3.81}$$

where $\theta_\nu(B)^\cdot = (\theta_\nu(B) - 1)/B)$, and

$$\begin{aligned} \theta_{\nu'}(B)\partial a_{(t,\nu')}/\partial\phi_{(\nu,j)} &= 0 \\ \theta_{\nu'}(B)\partial a_{(t,\nu')}/\partial\theta_{(\nu,j)} &= 0 \end{aligned} \tag{3.82}$$

if $\nu \neq \nu'$. If moving average terms are absent the well known result (Pagano 1978; Vecchia, 1985) that the information matrix is block diagonal follows easily from these equations.

Another simplification results when

$$\prod_{i=1}^{S}\theta_i(B) \tag{3.83}$$

is small in the unit circle. In this case, the derivatives can be approximated by

$$\begin{aligned} \frac{\partial a_{(t,\nu)}}{\partial \phi_{(\nu,j)}} &= -x_{(t,\nu-j)} \\ \frac{\partial a_{(t,\nu-1)}}{\partial \theta_{(\nu,j)}} &= -a_{(t,\nu-j)} \\ \theta_{\nu'}(B)\frac{\partial a_{(t,\nu')}}{\partial \phi_{(\nu,j)}} &= 0 \\ \theta_{\nu'}(B)\frac{\partial a_{(t,\nu')}}{\partial \theta_{(\nu,j)}} &= 0. \end{aligned} \tag{3.84}$$

If we assume further that

$$\frac{\partial a_{(t,\nu')}}{\partial \phi_{(\nu,j)}} \approx 0 \tag{3.85}$$

we obtain as in the AR case that the information matrix I is close to a block diagonal matrix which is easily computable.

In the general case, observe that because of (3.82), using (3.81) recursively we obtain that

$$\begin{aligned} \partial a_{(t,\nu')}/\partial \phi_{(\nu,j)} &= \textstyle\sum_{k=0}^{\infty} \pi(\nu,k;\nu',j)x_{(t-k,\nu-j)} \\ \partial a_{(t,\nu')}/\partial \theta_{(\nu,j)} &= \textstyle\sum_{k=0}^{\infty} \pi(\nu,k;\nu',j)a_{(t-k,\nu-j)}, \end{aligned} \tag{3.86}$$

where $\pi(\nu,k;\nu',j)$ is obtained recursively using the following equations

$$\theta^{(l+1)}_{\nu,i+1} = \theta^{(l)}_{\nu,i}\theta^{(1)}_{\nu-l+1,i} + \theta^{(l)}_{\nu,i+1}, \tag{3.87}$$

and

$$\pi(\nu,k;\nu',j) = -\theta^{([k-1]S+1)}_{\nu,[k-1]S+1}. \tag{3.88}$$

Another approach for the computation of the information matrix is as follows. First, note that

$$I \approx \frac{1}{S}\boldsymbol{E}\frac{1}{\sigma^2_\nu}\left(\sum_{\nu=1}^{S}\left(\frac{\partial^2 a_{(t,\nu)}}{\partial \underline{\gamma}\partial \underline{\gamma}'}\right)\right). \tag{3.89}$$

Next, multiply both sides of the equation (3.40) by a_t and differentiate with respect to the parameters γ_1 and γ_2. We obtain the following equation

$$\frac{\partial^2 a_t^2}{\partial\gamma_1\partial\gamma_2} + \theta_{t,1}\frac{\partial^2 a_t a_{t-1}}{\partial\gamma_1\partial\gamma_2} + \cdots + \theta_{t,q}\frac{\partial^2 a_t a_{t-q}}{\partial\gamma_1\partial\gamma_2} + \sum_{i=1}^{p}\frac{\theta_{t,i}}{\partial\gamma_1}\frac{a_t a_{t-i}}{\partial\gamma_2} = \\ (x_t + \phi_{t,1}x_{t-1} + \cdots + \phi_{t,p}x_{t-p})\frac{\partial^2 a_t}{\partial\gamma_1\partial\gamma_2} + \sum_{i=1}^{p}\frac{\phi_{t,i}}{\partial\gamma_1}\frac{a_t x_{t-i}}{\partial\gamma_2}. \tag{3.90}$$

The first term in the right hand side of (3.90) is, after taking expectations, negligible. Next, we consider the last term in the left hand side of (3.90). It satisfies the following equations

$$\boldsymbol{E}\frac{\partial a_{t-j}a_{t-i}}{\partial\gamma_2} + \theta_{t,1}\boldsymbol{E}\frac{\partial a_{t-i}a_{t-j-1}}{\partial\gamma_2} + \cdots + \theta_{t,q}\boldsymbol{E}\frac{\partial a_{t-j}a_{t-q}}{\partial\gamma_2} \\ + \sum_{i=1}^{p}\frac{\theta_{t,i}}{\partial\gamma_1}\boldsymbol{E}a_{t-j}a_{t-i} = \sum_{l=1}^{p}\frac{\phi_{t,l}}{\partial\gamma_2}\boldsymbol{E}a_{t-j}x_{t-i}. \tag{3.91}$$

Both, $\boldsymbol{E}a_{t-j}a_{t-i}$, and $\boldsymbol{E}a_{t-j}x_{t-i}$ are known. Hence, we have for each i a system of linear equations with unknowns $\boldsymbol{E}\partial a_{t-j}a_{t-i}/\partial\gamma_2$ for $j = 1,\ldots,p$, and $t = 1,\ldots,S$. Note that we have to use the periodicity of the process. An analogous system of linear equations for each i can be obtained for $\boldsymbol{E}\partial x_{t-i}a_{t-j}/\partial\gamma_2$,

$$\boldsymbol{E}\frac{\partial x_{t-i}a_{t-j}}{\partial\gamma_2} + \theta_{t,1}\boldsymbol{E}\frac{\partial a_{t-j}a_{t-i-1}}{\partial\gamma_2} + \cdots + \theta_{t,q}\boldsymbol{E}\frac{\partial a_{t-j}a_{t-i-q}}{\partial\gamma_2} = \sum_{l=1}^{p}\frac{\phi_{t,l}}{\partial\gamma_2}\boldsymbol{E}x_{t-l-j}x_{t-i}. \tag{3.92}$$

Finally, taking expectations in (3.90) and using (3.91), and (3.92), we obtain a system of linear equations for $\boldsymbol{E}\partial^2 a_t a_{t-i}/\partial\gamma_1\partial\gamma_2$

$$\boldsymbol{E}\frac{\partial^2 a_t a_{t-i}}{\partial\gamma_1\partial\gamma_2} + \theta_{t,1}\boldsymbol{E}\frac{\partial^2 a_{t-i}a_{t-1}}{\partial\gamma_1\partial\gamma_2} + \cdots + \theta_{t,q}\boldsymbol{E}\frac{\partial^2 a_{t-i}a_{t-q}}{\partial\gamma_1\partial\gamma_2} - \sum_{l=1}^{p}\frac{\theta_{t,l}}{\partial\gamma_1}\boldsymbol{E}\frac{a_{t-l}a_{t-i}}{\partial\gamma_2} =$$
$$\boldsymbol{E}(x_t + \phi_{t,1}x_{t-1} + \cdots + \phi_{t,p}x_{t-p})\boldsymbol{E}\frac{\partial^2 a_{t-i}}{\partial\gamma_1\partial\gamma_2} + \sum_{i=1}^{p}\frac{\phi_{t,i}}{\partial\gamma_1}\boldsymbol{E}\frac{\partial a_t x_{t-i}}{\partial\gamma_2}. \tag{3.93}$$

It is important to note that all systems of linear equations in (3.91), (3.92), and (3.93) involve the same left hand side. Hence, once we factorize the associated matrix into an upper triangular matrix times a lower triangular matrix, we may use this factorization for all the remaining systems of linear equations. This speeds up algorithm considerably.

Another approach to obtain the information matrix could be the estimation of the variance matrix of the vector of derivatives by simulating using the equations (3.78-3.82). Also, we could use the multivariate representation of the process, to obtain the covariance matrix. Finally, we could proceed as in McLeod (1984) and use results about dual models to obtain the information matrix. We proceed to do this.

3.10 Duality

Consider first a multiple time series that depends on R parameters Ψ, where the parameters are in vector form, and Ψ is an R-vector. Then, under general conditions the information matrix of the parameters is obtained as follows. Define

$$\mathcal{L} = \sum \hat{a}_t^2. \tag{3.94}$$

Then, the information matrix can be obtained from

$$I \rightarrow -\frac{1}{N}\boldsymbol{E}\{\frac{\partial^2 \mathcal{L}}{\partial\Psi\partial\Psi'}\}. \tag{3.95}$$

Also, under very general conditions we have that

$$I = \lim_N \frac{1}{N}\boldsymbol{E}\left(\frac{\partial \mathcal{L}}{\partial\Psi}\right)\left(\frac{\partial \mathcal{L}}{\partial\Psi}\right)'. \tag{3.96}$$

Now, consider a one to one transformation of the parameters to Ψ^*

$$\Psi^* = \Psi^*(\Psi). \tag{3.97}$$

Observe that

$$\frac{\partial \mathcal{L}}{\partial\Psi} = \left(\frac{\partial \mathcal{L}}{\partial\Psi^*}\right)\left(\frac{\partial\Psi*}{\partial\Psi}\right)'. \tag{3.98}$$

Call J the Jacobian matrix of the transformation $\Psi = \Psi^*(\Psi)$, then

$$I(\Psi) \approx JI(\Psi^*)J' \tag{3.99}$$

where $I(\Psi)$ is the information matrix with respect to the parametrization Ψ, and $I(\Psi^*)$ is the information matrix with respect to the parametrization Ψ^*.

This result can be used with the stochastic process given by

$$\Phi(B)\underset{\sim}{X}_t = \Theta(B)\underset{\sim}{a}_t. \tag{3.100}$$

We naturally set $\Psi = (\Phi; \Theta)$, and $\Psi^* = \alpha$, where α is the vector of the coefficients of the polynomial $\alpha(B) = \Phi(B)\Theta(B)$, corresponding to a multivariate stochastic process that satisfies the equations

$$\alpha(B)\underline{Y}_t = \underline{a}_t. \tag{3.101}$$

Observe that both stochastic processes have the same white noise input. These two models are said to be adjoint to each other. In this case,

$$\begin{aligned} \frac{\partial \alpha_i}{\partial \phi_j} &= -\theta_{i-j} \\ \frac{\partial \alpha_i}{\partial \theta_j} &= -\phi(i-j). \end{aligned} \tag{3.102}$$

Note that we are reordering the matrix derivatives to obtain an expression similar to the univariate case. But if the polynomials ϕ and θ satisfy the well known conditions for stationarity and invertibility, then, because both processes have the same input noise, the estimated residuals from N observations of each stochastic process will be within $O_p(N^{-1})$, where O_p denotes the probability big O symbol. Then using (3.99) we may express the information matrix of the multivariate ARMA process in (3.100) in terms of the information matrix of the multivariate AR process defined in (3.101).

We could use (3.99) and the equivalence between PARMA models and multivariate time series models to obtain a symbolic expression for the information matrix of the general PARMA time series process. However, it is also of interest to find a symbolic formula for the information matrix using the special structure of PARMA models, especially taking advantage of the fact that the information matrix of a PAR process is block diagonal. This is useful because in the PARMA case we can work with univariate time series directly, hence, we do not need the special ordering of the matrix derivatives that we used for multiple time series. Now, we could follow the same line of reasoning for PARMA models as with univariate models (McLeod, 1984). If the PARMA model is given by

$$\phi_t(B)x_t = \theta_t(B)a_t, \tag{3.103}$$

we work with the following adjoint

$$\phi_t(B)\theta_t(B)y_t = a_t. \tag{3.104}$$

Then, if $\hat{a}_{t,x} = \hat{a}_{t,y} + O_p(n^{-1})$

$$I(\phi_t, \theta_t, \quad t = 1, \ldots, S) = J'I(\alpha_t, t = 1, \ldots, S)J \tag{3.105}$$

or

$$I_x = J'I_yJ. \tag{3.106}$$

3.11 A result about the mse of estimated forecasts

We will consider in this section a model given by

$$x_{(t+1)} = \alpha_{(t+1)}x_t + \sigma_{(t+1)}\epsilon_{t+1}, \tag{3.107}$$

where $\{\epsilon_t\}$ is a collection of independent random variables with mean 0 and variance 1. The $\alpha_t = \alpha_t(\underline{\phi})$ is a function of t with that depends on some parameter $\underline{\phi}$.

We are interested in the one step ahead prediction $\hat{x}_{N+1|N}$ of the process $\{x_t\}$ based on the observations $x_1, \ldots, x_N$. As is well known, this is given by

$$\hat{x}_{N+1|N} = \alpha_{N+1}(\underline{\hat{\phi}})x_N, \tag{3.108}$$

where $\underline{\hat{\phi}}$ is an estimate of $\underline{\phi}$ based on the observations. Then,

$$\hat{x}_{N+1|N} - x_{N+1} = (\alpha_{N+1}(\underline{\phi}) - \alpha_{N+1}(\underline{\hat{\phi}}))x_N - \sigma_{(N+1)}\epsilon_{N+1}. \tag{3.109}$$

Hence, the mean square error (MSE) of the forecast is given by

$$\mathrm{MSE}(\hat{x}_{N+1|N} - x_{N+1}) = \sigma^2_{N+1} + \mathrm{MSE}((\alpha_{N+1}(\underline{\phi}) - \alpha_{N+1}(\underline{\hat{\phi}}))x_N). \tag{3.110}$$

As a first approximation to the variance, drop the variance term for now and assume that

$$\underline{\hat{\phi}} = \underline{\phi} + O_p(N^{-1/2}). \tag{3.111}$$

Then, we have

$$\mathrm{MSE}(\hat{x}_{N+1|N} - x_{N+1}) \approx \boldsymbol{E}\left(x_N^2((\alpha_{N+1}(\underline{\phi}) - \alpha_{N+1}(\underline{\hat{\phi}}))^2\right) \tag{3.112}$$

$$\mathrm{MSE}(\hat{x}_{N+1|N} - x_{N+1}) \approx \boldsymbol{E}\left(x_N^2\right)\mathrm{MSE}((\alpha_{N+1}(\underline{\phi}) - \alpha_{N+1}(\underline{\hat{\phi}}))), \tag{3.113}$$

and

$$\mathrm{MSE}((\alpha_{N+1}(\underline{\phi}) - \alpha_{N+1}(\underline{\hat{\phi}})) \approx \frac{\partial\alpha_{N+1}(\underline{\phi})}{\partial\underline{\phi}}. \tag{3.114}$$

Then

$$\begin{aligned}\mathrm{MSE}(\hat{x}_{N+1|N} - x_{N+1}) &= \sigma^2_{N+1} \\ &+(\boldsymbol{E}x_N^2)\mathrm{var}((\alpha_{N+1}(\underline{\phi}) - \alpha_{N+1}(\underline{\hat{\phi}}))\frac{\partial\alpha_{N+1}(\underline{\phi})}{\partial\underline{\phi}}\end{aligned} \tag{3.115}$$

This result can be easily being extended to higher lags

$$\begin{aligned} \mathrm{MSE}(\hat{x}_{N+1|N} - x_{N+1}) &= \sigma^2_{N+1} \\ &+ (\mathbb{E}x^2_N)\mathrm{var}((\alpha_{N+1}(\underline{\phi}) - \alpha_{N+1}(\hat{\underline{\phi}}))\frac{\partial \alpha_{N+1}(\underline{\phi})}{\partial \phi}. \end{aligned} \tag{3.116}$$

The above equation shows that the curvature of the parametrization of the model is important so far as expansions up to order N^{-1} are considered. This result can be extended easily to the case of higher lags. The same equation (3.115) is valid but where the variance is

$$\sigma^2_{N+1} = \sum_{j=N+l}^{N-1} \prod_{i=N+l}^{j} \alpha_i \sigma^2_j \tag{3.117}$$

and

$$\alpha_N = \prod_{j=N+l}^{N+1} \alpha_j. \tag{3.118}$$

Hence, as the lag increases the curvature of α_N decreases, and the first term dominates.

For the case of PAR(1) processes, because the information matrix is block diagonal we could estimate the parameters individually by the Yule-Walker equations and hence

$$\hat{\phi}_i = \hat{\gamma}_{(i,1)}) = \phi_i + \sum \frac{\sigma_{(t(i)+1)}\sigma_{(t(i))}\epsilon_{(t(i)+1)}\epsilon_{t(i)}}{\hat{\sigma}^2_i}, \tag{3.119}$$

and then as before

$$\mathrm{MSE}(\hat{x}_{N(i)+1|N} - x_{N(i)+1}) = \sigma^2_{n(i)+1}\sigma^2_{n(i)} / \ \sigma^2_i. \tag{3.120}$$

3.12 Trends, smoothing and component models

Apart from seasonality another important characteristic of a time series process is its long term behaviour. The analysis of the long term behaviour of the time series is obscured by the presence of a strong cyclic component, or by any other process that blurs the long term trend. The removal of these components is of prime importance for the accurate analysis of long term trends. The strong emphasis in separating seasonal components from long term components has resulted in that the models that have been used traditionally have been restricted to models where the seasonal behaviour, the long term behaviour, and so on, are considered as independent elements. Moreover, these components appeared in time series process either in additive form, the most usual case, or in multiplicative form. Models that satisfy the above conditions will be called component models. The type of components that they could include will depend on the particular stochastic model under consideration. However, generally a long term trend component and a seasonal model component will be present.

Because of the economic and scientific importance of the providing acceptable models for seasonal stochastic behaviour, many mathematical models have been proposed for this purpose. A short review, limited to the components model, will follow.

Historically, the first class of models that was considered is what could be called, the deterministic trend and seasonal component models. Models in this class assume that the time series X_t can be described by,

$$X_t = T_t - S_t - N_t, \tag{3.121}$$

where T_t is a deterministic trend function, S_t is a deterministic periodic wave function, and N_t corresponds to the stochastic part of the model, generally assumed to be white noise (independent and identically distributed Gaussian random variables, with mean zero). The estimation of T_t, S_t, and of the probability distribution function of N_t depends on the level of detail that is assumed for each of the different components of (3.121). We use maximum likelihood estimation if the probability density function is known up to a few parameters. Alternatively, we use least squares if the probability density is not known to belong to a particular parametric family. Generally, it is assumed that the trend is linear or polynomial, and that the periodic function S_t can be adequately represented by a linear combination of sines and cosines. The noise is generally thought to be Gaussian.

The analysis of many economic time series has been done without assuming a parametric model for the global trend or the seasonal component, and the noise has generally been assumed to be Gaussian. In this nonparametric situation, the use of moving-average filters has been very popular. Indeed, the most popular procedure among government agencies and industry in the U.S. has been the X-11 methodology developed by The Bureau of Census of U.S. (Shiskin, et. al., 1967). The methodology to obtain estimates of the different values of T_t, S_t, and N_t can be described as consisting of:

(a) Estimate the long term trend by a particular moving-average filter of forty-three terms. The weights of this moving average filter were developed after careful consideration of many economic time series.

(b) After the estimation of the trend, the seasonal component is estimated by the use of a nine term trimmed mean of the seasonal values of the detrended time series The trimming consists of dropping the largest and the smallest of the different seasonal values and averaging the remaining values.

(c) The noise component is the end result of eliminating the seasonal component of the detrended time series.

The X-11 methodology has been proven very popular with the government and the industry in the U.S., and a variant of it with the government of Canada (Dagum, 1978). The reason for this popularity seems to be that the moving-average filters that the X-11 ARIMA uses were chosen after careful consideration of many time series that are similar to the time series where X-11 ARIMA is usually applied. Another important reason for its popularity is that the filter that X-11 uses includes some robustness procedures that eliminate the influence of outliers in the estimation of the parameters.

Useful extensions of the classical trend model include the generalization of the stochastic part, N_t, to a autoregressive moving average model, and the use of the Box-Cox (Box-Cox, 1964) family of power transformations:

$$X^{(\lambda)} = (X^{\lambda} - 1)/\lambda. \tag{3.122}$$

As explained in section 3.1 in the past decades the most important family of models for time series analysis has been the ARMA family. There are several approaches to fit these models, but the most important approach has been the fitting philosophy of Box and Jenkins.

Box-Jenkins methodology sometimes does not seem appropriate in certain situations, as already mentioned in section 3.1. Apart from examples given in section 1, another shortcoming of SARIMA models is that (3.1) does not show explicitly the different components of the model such as a global term, and seasonal component, etc.. A further example of a situation where it seems that the family of models in (3.1) is not appropriate is the following. Sometimes the time series modeler encounters this situation. Using the Box-Jenkins modelling approach a certain seasonal integrated autoregressive moving average model is tentatively chosen. But, by checking the residuals the modeler realizes that seasonal variation has not been removed adequately by the chosen model. Also, by looking at the autocorrelation function of the process, computed for each season, the modeler sees that the second moment behaviour of the process seems to vary with the season. For example, the first order autocorrelation of some monthly river flows seems to have a different sign in winter than in the summer. However, the SARIMA family of models is not able to exhibit this type of behavior. These models are static in time, although they use periodic information. Moreover, for some investigators static models do not make sense. They feel that the models should be different for different seasons.

These two situations indicate different generalizations of the Box-Jenkins methodology. The first one is the combination of the classical trend model with Box-Jenkins autoregressive and moving average models. That is, instead of considering a deterministic trend function, either for the global trend or for the seasonal trend, one could consider that these component are stochastic in nature but independent between them, and assume that they could modeled by an ARMA model (Kitagawa and Gersch, 1984; Shiller, 1979; Akaike, 1980). Then the process that we are considering satisfy (3.121) plus,

$$\begin{aligned} \phi_T(B)T_t &= \theta_T(B)\omega_t(T), \\ \phi_S(B)S_t &= \theta_S(B)\omega_t(S), \\ \phi_N(B)N_t &= \theta_N(B)\omega_t(N). \end{aligned} \tag{3.123}$$

Each of the processes defined in (3.123) belongs to the family of ARMA processes, and the noise inputs $\omega_t(.)$ are assumed independent, or at least uncorrelated, among each other. Generally, the noises $\omega_t(.)$ are assumed to have the same distribution function but with different variances. That is,

$$\omega_t(.) = \lambda_{(.)}\epsilon_t(.). \tag{3.124}$$

By substituting (3.123) back into (3.121), one obtains,

$$X_t = \frac{\theta_{\mathcal{T}}(B)}{\phi_{\mathcal{T}}(B)}\omega_t(\mathcal{T}) + \frac{\theta_{\mathcal{S}}(B)}{\phi_{\mathcal{S}}(B)}\omega_t(\mathcal{S}) - \frac{\theta_{\mathcal{N}}(B)}{\phi_{\mathcal{N}}(B)}\omega_t(\mathcal{N}). \tag{3.125}$$

Hence, the extension of the classical trend model in (3.121) by (3.123) is also an extension of the Box-Jenkins models in (3.1). However, obtaining a model for the time series X_t using (3.123) is totally different from obtaining a model using (3.1). Because the model in (3.125) given in Box-Jenkins form could have been produced by several models of type (3.123), the identification of the different components is easier working with (3.123) than with (3.125).

The model defined by (3.121) and (3.123) can also be written in state-space form. To accomplish this representation any of the well known canonical forms of ARMA models in the state-space form (Harvie 1984a,b) could be used for each of the processes appearing in (3.123). For example, state-space variables of each component in (3.123), $S_t(\mathcal{T})$, $S_t(\mathcal{S})$, $S_t(\mathcal{N})$, respectively, could be defined as follows. Given initial values of $S_0(\mathcal{T})$, $S_0(\mathcal{S})$, $S_0(\mathcal{N})$, and their covariance matrix, future values of $S_t(\mathcal{T})$, $S_t(\mathcal{S})$, $S_t(\mathcal{N})$, satisfy the first order matrix difference equation given by,

$$\begin{aligned} S_{t+1}(.) &= \mathbf{F}(.)S_t(.) + \mathbf{G}\lambda_{(.)}\epsilon_t(.) \\ S_t(.) &= \mathbf{H}S_t(.), \end{aligned} \tag{3.126}$$

where $(\mathbf{F}, \mathbf{G}, \mathbf{H})$ is the canonical observability form corresponding to (3.123). The symbol (.) is used to denote the different components, that is $\mathcal{T}$, $\mathcal{S}$, or, $\mathcal{N}$. These equations could be referred to as the process equations. They update the internal state of the process. Finally, the resulting output is put together by the use of (2.121), the observability equation.

The representation of (2.123) in state space form provides us with efficient algorithms to estimate recursively the state variables, like the Kalman filter (Kalman, 1960) or any of its stable numerical representations (Morf et al., 1974). The state space representation it is also important, because, as we will see later, it generalizes easily to other models.

Finally, we will make two other points about the state variable representation of (2.123). The first observation is that, the definition of a stochastic global trend is based on the variance of this trend. The global trend could be thought as a process with much smaller variance than the noise process. But,in many situations it is a matter of the subjective judgement for the definition of this trend. That is, in many cases it is up to the modeler to decide how much variance, or how *smooth*, the global trend process.is. The second observation is that state space representation allows that information, a priori or external information, to be entered into the model equations and to be used in the estimation of the state space. Box-Jenkins procedures are not so flexible in this respect as the state space models are. Indeed, the Kalman filter could be seen as based on the recursive estimation of the state variable using Bayes' theorem and the a priori estimate of the state. This has been investigated by Harrison and Stevens (1976) and we will investigate it further.

Generalization to periodic models

We could extend (3.123) to periodic models in several ways:

(a) Instead of using the second set of equations in (3.123)) we could assume that the third set of equations corresponding to the noise is a PARMA type model.

(b) We could eliminate the second set of equations and assume that the first set of equations is a PARMA model.

State-space model and smoothing: review and generalizations

Several extensions of the models in (3.121) and (3.121) to (3.123) have been proposed. One interesting extension consists on the use of the smoothing priors methodology proposed by Kitagawa and Gersch (1984) Schiller (1984), and Akaike (1980), and in a more general context the nonparametric function fitting of Wahba and her colleagues (1977, 1975, 1983), and Wecker and Ansley (1985). Basically, the idea is to estimate the trend component and the seasonal component by using stochastically perturbed difference equations. As a goodness of fit criterion use functions related to the topological properties of the different trend components such as the low curvature of the global trend, the stability of the seasonal part, and the noise plus small variance of the stochastic component. The nonparametric approach that these investigators chose led to the smoothing spline. Moreover it is known (Kilmeldorf and

Wahba, 1970) that smoothing splines can be justified from a Bayesian approach. This approach has been used by Akaike (1980) and Kitagawa and Gersch (1984, 1980) tq smooth time series.

The fitting methodology of Kitawaga ad Gersch (1980, 1984) can be reinterpreted as follows. The observations are the result of the cumulative effect of the stochastic processes that act at different levels. For example, many times the global level of the process $\{X_t\}$ could be assumed to follow a random walk. Apart from the variation due to a stochastically changing global level, the process could exhibit seasonal variation. This seasonal variation could be driven by another stochastic process, independent of the global level process, whose expectation, conditional on the past, is periodic. By adding stochastic processes in this form, one could take into account the behaviour of the processes at different frequencies.

One of the apparent shortcomings of the models obtained using Box-Jenkins methodology is that they seem to forecast better than most other approaches when the forecast horizon is short. However, for long term forecasting they seem not to be so accurate. This inability (Makridakis and Hibon, 1979; Makridakis et al. 1982) of forecasting long term trend by Box-Jenkins models perhaps could be overcome by the use, after Kitawaga and Gersch (1980, 1984), of the following class of models. Given N observations of the process X_t, a model similar to (3.121) is fitted to the data. Hence

$$X_t = C_t(1) + C_t(2) + C_t(3) + \ldots + C_t(p), \qquad (3.127)$$

where the $C_t(i)$'s satisfy autoregressive equations, and the autoregressive filters could have zeros on the complex unit circle. The input noise of these autoregressive processes, $\omega_t(i)$, is assumed to be white noise, generally Gaussian white noise, uncorrelated among different components, $C_t(i)$'s, of the model in (3.125) (that is, the different components are stochastically independent). The autoregressive filters are chosen to ensure that the "noisiness" of the components $C_t(i)$ increases as the index 'i' increases, and also the lags of the autoregressive filters are chosen to take into account certain specific behaviour such as seasonality or trading day effect of the time series X_t. A most important characteristic of the model defined by (3.127) is that the $C_t(i)$'s are unobserved variables and they play the same role as state variables for the state space model in (3.126) and in the factor analysis model. Indeed, the model in (3.127) can be represented readily in a Markovian state space form.

The feature that differentiates the model in (3.127) from the type of models attainable using a state space representation of the process $\{X_t\}$ is the role that the variances of the components, $C_t(k)$, play. They control the "smoothness of the stochastic component and hence they play a similar role to the smoothing constant of nonparametric regression. They are similar quantities to the measures of signal to noise ratio used mainly by engineers. However, the choice of these smoothing constants is the most crucial aspect of the model. Several forms have been proposed. For example, Akaike (1980) assumes a Bayesian model for these parameters; Kitawaga and Gersch (1984) proposed to use the estimate of the Akaike information Criteria (AIC) over a grid of values of the smoothing constants λ_k in (3.124), and choose the model that gives the smallest AIC.

As an example of the use of the model in (3.127), Kitawaga and Gersch (1984) analyzed economic time series data. Because of the particular nature of economic time series they chose four as the value of p, where p is the number of stochastic components. For their model, $C_t(1)$ corresponded to a local polynomial trend of low curvature, that satisfied,

$$\nabla^k(B)C_t(1) = \omega_t(1), \tag{3.128}$$

where $\nabla = (1 - B)$ is the first order differencing operator. Their second component corresponded to the seasonal variation, given by

$$C_t(2) = \sum_{j=1}^{S} C_{t-j}(2) + \omega_t(2). \tag{3.129}$$

Their third component $C_t(3)$ corresponded to trading day effect. $C_t(3)$ satisfied the following relationship, where i is the day of the week corresponding to day t,

$$C_i(3) = C_{i-1}(3); \text{for} \quad i = 1, \ldots, 6; \tag{3.130}$$

and

$$\sum_{i}^{7} C_i = 0. \tag{3.131}$$

Finally, the last component corresponded to the 'noise' of the system, and was modeled assuming that it satisfied an autoregressive model. The most important parameters to estimate, those which control the 'wigginess' of the trends are the variances of the white noise sequences $\omega_t(i)$. The noise processes are assumed to satisfy the relationship in (3.123). The parameters that have to be estimated are the relative size λ_K of these variances. Under the Bayesian methodology that these authors adopt, these parameters are hyperparameters (Lindley and Smith, 1972). The Gaussian assumption about the residuals implies the use of the sum of the squares of the residuals corresponding to the different input sequences. One can use

$$l = \sum_{k=1}^{p} \sum_{t=1}^{N} \hat{\omega}_t^2(k)/\lambda_k, \tag{3.132}$$

as a criterion to be minimized in order to estimate the parameters. To compute the residuals $\omega_t(i), i = 1, \ldots, N_r$ the state space representation of the model in (3.127 is used. This is easily obtained from the state space representation of the autoregressive processes $C_t(i)$'s. Using this Markovian representation, recursive estimation of the state and of the parameters can be achieved by means of the the Kalman filter. To use the Kalman filter, the recursion has to be initialized by giving the values of the state variables at stage zero, and by specifying the covariance matrix of the state vector at time zero. These parameters need to be estimated in most cases. The estimation could be done by backforecasting the initial state vector and its covariance matrix. A computational efficient way of doing this is by using the backward form of the Kalman filter.

In the models that Kitawaga and Gersch (1980, 1984) employ, the exact form of the filters is known. However, the noise the system is assumed to be unknown, either by lack of knowledge of the variances of the white noise for each component, or by assuming an autoregressive model for the noisiest part of the model.

The recursive equations used by Kitawaga and Gersch (1980, 1984) are the following. The overall state space variable at time t is denoted by $\mathbf{S}_t$, where

$$\mathbf{S}_t = (\mathbf{S}_t(1), \ldots, \mathbf{S}_t(p)). \tag{3.133}$$

The evolution of the state variables for the components of the process is governed by equations similar to (3.126). Then, the overall state variable satisfies a equation similar to (3.126) but with the form $(\mathbf{F}, \mathbf{G}, \mathbf{H})$ given by

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{F}_2 & \ldots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & - & \ldots & \mathbf{F}_p \end{pmatrix} \tag{3.134}$$

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & 0 & \ldots & 0 \\ 0 & \ddots & & \vdots \\ 0 & \ldots & 0 & \mathbf{G}_p \end{pmatrix} \tag{3.135}$$

and,

$$\mathbf{H} = (\mathbf{H}_1, \ldots, \mathbf{H}_p). \tag{3.136}$$

If we denote the predicted value of the state variable at time t, given the observations up to time s, by $S(t|s)$, then, the recursive equations used by

Kitagawa and Gersch (1980, 1984) (Kalman Filter equations) are given by,

$$S(t|t-1) = \mathbf{F}S(t-1|t-1), \tag{3.137}$$

$$S(t|t) = S(t|t-1) - \mathbf{K}_t e_t, \tag{3.138}$$

$$e_t = X_t - \mathbf{H}S(t|t-1), \tag{3.139}$$

$$\mathbf{K}_t = \mathbf{V}(t|t-1)\mathbf{H}_t'\mathbf{V}(t|t-1)^{-1}, \tag{3.140}$$

$$\mathbf{V}(t|t-1) = \mathbf{F}\mathbf{V}(t-1|t-1)\mathbf{F}' - \mathbf{G}\mathbf{Q}\mathbf{G}', \tag{3.141}$$

$$\mathbf{V}(t|t) = \mathbf{I} - \mathbf{K}_t\mathbf{H}_t\mathbf{V}(t|t-1). \tag{3.142}$$

REMARK 3.11 The extension of this model to the periodic case is [illegible]. We assume that the appropriate component is a PARMA type model.

3.13 Periodic EAR(1) models

Lawrence and Lewis (1980) proposed a family of stationary time series with exponential marginals and autoregressive autocorrelation function, the so called EAR models. These models can be generalized to the case of periodic models as follows. We will assume an autoregression of order one process

$$y_t = \alpha_t y_{t-1} + \epsilon_t \tag{3.143}$$

where α_t and the probability distribution of ϵ_t are periodic in an extended sense. We want to find the probability distribution of ϵ_t, such that the probability distribution of y_t is exponential with rate λ_t, λ_t is periodic.

The Laplace transform of the density of y_t will be denoted by $\phi_t(s)$ and that of the density of ϵ_t by $\phi_{\epsilon_t}(s)$. Then, by stationarity

$$\phi_t(s) = \phi_{t-1}(\alpha_t s)\phi_{\epsilon_t}(s). \tag{3.144}$$

Hence

$$\phi_{\epsilon_t}(s) = \frac{\phi_t(s),}{\phi_{t-1}(\alpha_t s)}. \tag{3.145}$$

Then, because

$$\phi_t(s) = \frac{\lambda_t}{s + \lambda_t}, \tag{3.146}$$

we have that

$$\begin{aligned} \phi_{\epsilon_t}(s) &= \frac{\lambda_t}{\lambda_{t-1}}\alpha_t + \left(1 - \alpha_t \frac{\lambda_t}{\lambda_{t-1}}\right) \frac{\lambda_t}{s + \lambda_t} \\ &= p_t + q_t \frac{\lambda_t}{s + \lambda_t}. \end{aligned} \tag{3.147}$$

We have then the following result

THEOREM* 3.12 If the probability density of the noise ϵ_t is a mixture of an exponential density with rate λ_t and the probability distribution that has all its mass at zero, the mixture proportions given by $p_t = \lambda_t/\lambda_{t-1}\alpha_t$ and $q_t = 1 - p_t$, then the probability density of y_s in (3.127) is exponential with parameter λ_t.

REMARK 3.13 We are also including in this theorem what could be called left exponential densities or densities defined on the negative real axis. Then the rate λ_t is negative.

There is a a constraint in the rates λ_t and the parameters λ_t, namely

$$0 < p_t = \frac{\lambda_t}{\lambda_{t-1}}\alpha_t < 1. \tag{3.148}$$

Hence λ_t and α_t have to have the same sign if $\lambda_{t-1} > 0$ and opposite signs if $\lambda_{t-1} < 0$. Also, $\lambda_t\alpha_t < \lambda_{t-1}$ if λ_{t-1} positive and $\lambda_t\alpha_t > \lambda_{t-1}$ if λ_{t-1} negative.

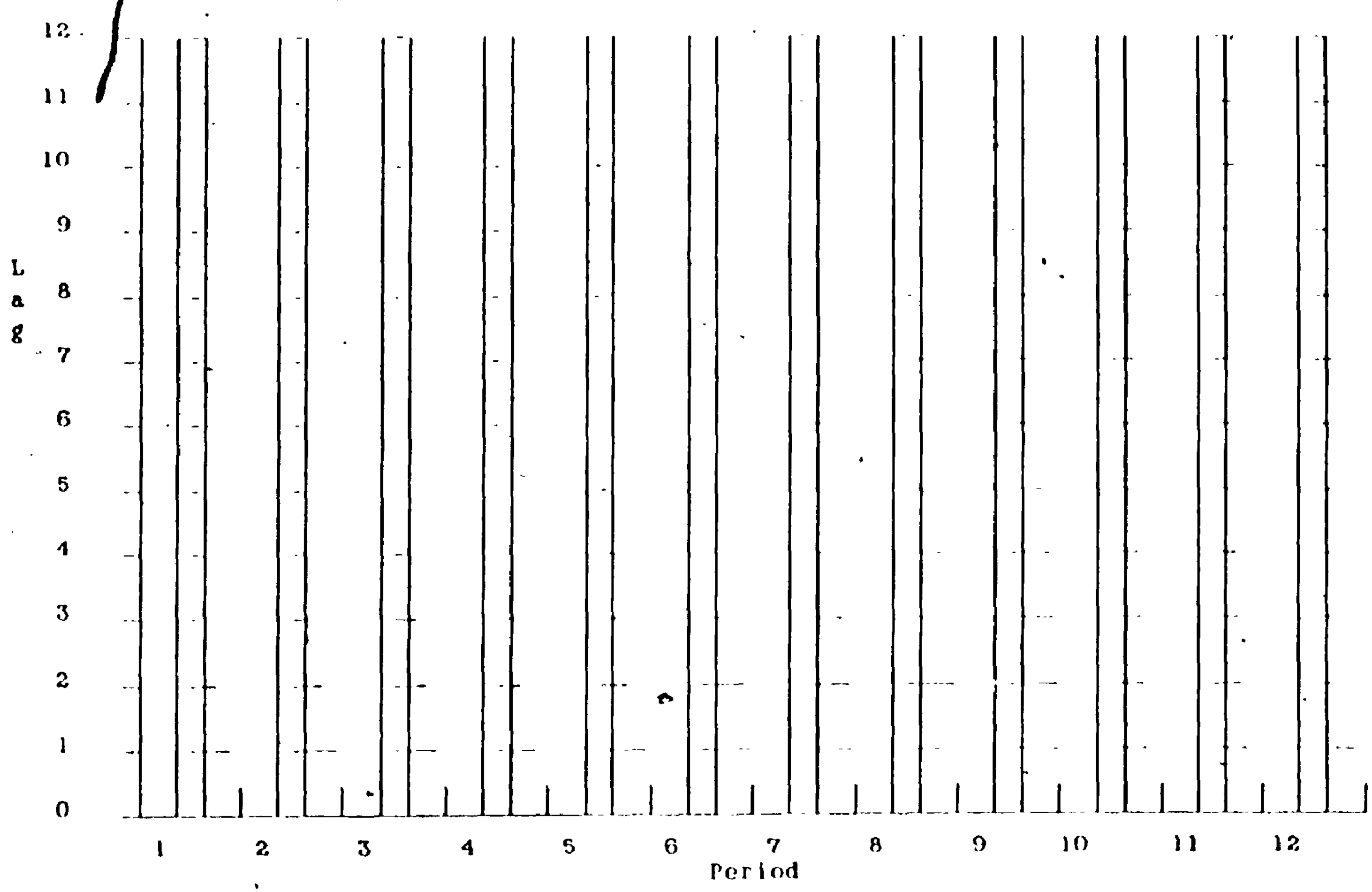


Figure 3.1 Periodic ACF for the Namakan River.
Vertical denote a 95% confidence band.

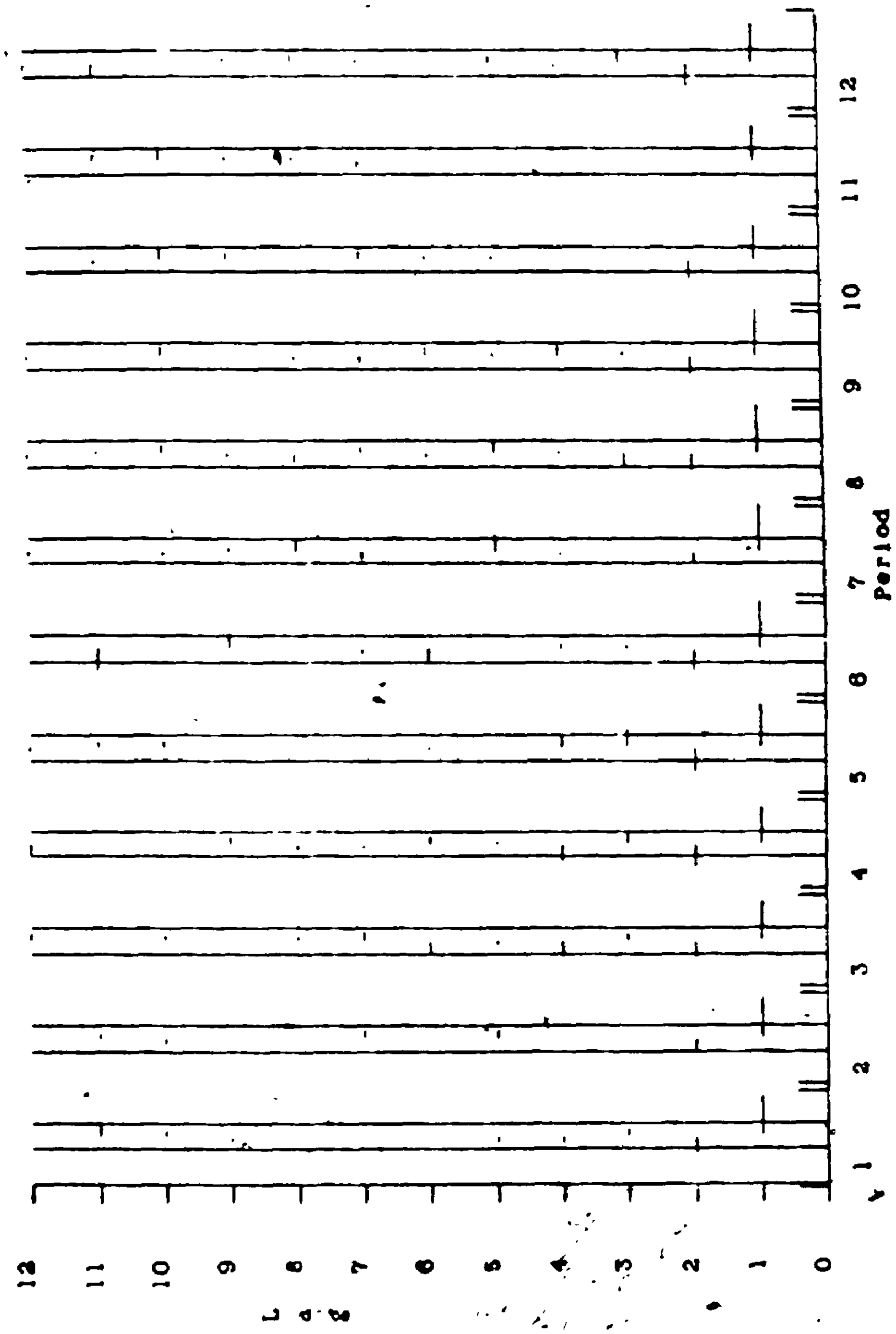


Figure 3.2 Periodic sample PACF for the Namakan River. Vertical lines denote a 95% confidence interval.

	AIC	
Period	Model 1	Model 2
1	-232	-251
2	-266	-288
3	-244	-249
4	-108	-112
5	-84	-79
6	-120	-132
7	-129	-140
8	-145	-150
9	-143	-151
10	-90	-105
11	-116	-135
12	-181	-186

Table 3.1
Namakan's River AIC's for each season for model 1 and 2
See text for explanation of Models 1 and 2.

3.14 Applications

As an application of the results in this chapter we will fit a PARMA model to a monthly riverflow record. The river that we consider for fitting is the Namakan at Lac de la Croix in Ontario. The monthly time series record in cubic meters per second (cms) goes from 1923 to 1977. Hence, altogether we have 648 observations. Its mean annual flow is 108 cms. We will transform the data by taking natural logarithms. We will use the first 600 observations to fit a model and the remaining 48 observations to obtain an estimate of the one-step ahead forecast error.

There are several methods for deciding which PARMA model we should fit to the time series. One is to use the cut-off properties of the periodic ACF and the PACF of the time series record. Figure 3.1 gives the sample ACF and Figure 3.2 gives the sample PACF. These graphs give, for each period, the 95% asymptotic confidence interval that is obtained assuming that the observations are independent. These 95% confidence limits are drawn as vertical lines. If the autocorrelation function, the horizontal lines, crosses the confidence interval we consider that there is correlation at the particular lag where the crossing happened. If the pattern of crossings for the autocorrelations damps out while the partial autocorrelations cut-off we will consider fitting an autoregressive model to that particular period. If we have the same pattern of crossings but with the roles of the ACF and PACF interchanged, we will consider the use of a moving average model. If both, the ACF and the PACF cut-off we will consider an ARMA model for that particular period.

A PAR(1) model (Model 1) is popular among hydrologists for fitting monthly riverflows. Indeed, by the graphs of the sample ACF and the sample PACF, given in Figure 3.1 and 3.2 respectively, this model seems be a possibility. The overall AIC of this model is −1880. The first column of Table 3.1 give the AIC for each season. The estimated one-step ahead forecast error, based on the last forty-eight observations is 0.082. A simple alternative to the PAR(1) model is a PARMA(1;1) model (Model 2). The overall AIC for this model is −1955. The second column of table 3.2 give the AIC for this model. Comparing these AIC's we see that except for period number 5, the AIC's of Model 2 are smaller, and in some cases substantially smaller, than the AIC's of Model 1. The one-step ahead forecast for Model 2 is 0.057. Then, there is as 17% improvement in the forecasting ability of Model 2 as compared with the forecasting ability of Model 1.

Based on these comparisons we may conclude that the introduction of moving average terms is a substantial improvement, for the Namakan River, over the popular PAR(1) model used by some hydrologists. We may also investigate also the use of higher-order lags but it was found that the improvement, if there was some, was only marginal.

3.15 Summary

Although PARMA models are equivalent to multiple ARMA models, they are interesting to investigate for two reasons. The parametrization of PARMA models permits us to use our insight about the time series process naturally (without imposing restrictions on the parameter space). That is, many times it is more natural to work with the PARMA parametrization than with the multivariate ARMA family. This implies that if we find that the PARMA family is not adequate for the particular seasonal time series we are working with, we might find an adequate family, extending, generalizing the PARMA family. Some of these generalizations are equivalent to generalizations of multiple ARMA models, however, most are not equivalent (nonlinear time series, bilinear extensions, threshold models, state-dependent models, random seasons). The other reason for using PARMA models instead of multivariate ARMA is related to the first one. Many times,in order to obtain comparable fits, the number of parameters that have to be specified when we use the PARMA family is much smaller that the number of parameters needed by the multivariate ARMA parametrization.

The PARMA family has not been studied from the point of view of state-space. However, when we are just interested in the second order properties of the time series process, a natural way of representing the process is through the state-space representation. We have done this in this chapter. The state-space representation is not unique and different representations have different advantages. Hence it is interesting to obtain different state-space representation of PARMA models. These different representations give different conditions that imply the seasonal stationarity of PARMA models (Theorem 3.2). The state-space representation is also useful to give faster exact Gaussian maximum likelihood algorithms, as the one given in Section 3.8. This algorithm is faster than previous algorithms (Vecchia, 1985) and it is also easily amenable for using with missing observations, aggregated data, nonstandard seasons and more nonstandard situations.

Next, we study the utocovariance structure of PARMA processes. We show that the autocovariance function may be obtained solving a system of linear equations similar to Mcleod's (1975). We rederived Sakai's algorithm and definition of partial correlation coefficients to show that contrary to his conclusion that the partial correlation coefficient are not useful for identification, they may be used to identify the process. This is due to a cut-off property that they posses. The inverse autocorrelations have been found useful for identification (Cleveland, 1972; Mcleod et al. (1977); Chatfield, 1979). We defined the inverse autocorrelation of PARMA process and gave an algorithm for their estimation. They may be used as an additional means for identifying the time seties process.

In order to make tests about the fitted parameters of a PARMA model we need to compute the information matrix. This is simple if the model does not have moving average parameters. In this case the information matrix is b.ock triangular, with a block for each season and each block may be computed separately from the other seasons. In the general case, the moving average parameters act as a kind of feedback that precludes the information matrix to be block triangular for each season. We found an expression for the information matrix, and we studied several cases, where we may approximate the information matrix by a block triangular matrix.

We give two algorithms for obtaining the information matrix. One is similar to the algorithm given in Box and Jenkins (1976). This algorithm may also be used to approximate the gradient and the Hessian of the sum of the squared residuals. This is useful for minimization of the sum of squares. The other algorithm is based on the multivariate representation of PARMA processes. Following McLeod (1984), we defined duality for PARMA processes and indicated how it may be used to obtain an approximate information matrix.

Many times we are interesting in fitting a PARMA process for forecasting. In this case it is of interest to have an expression of the MSE of the prediction. We have derived an expression for the MSE.

We have assumed the process stationary. In reality many processes exhibit some kind of trend. One approach with such as process is to eliminate the trend by smoothing. We indicated how this may be done based on the smoothing priors methodology of Kitagawa and Gersch (1984). Finally, we gave a model similar to the EAR(1) model in Lawrence and Lewis (1978).

References

1. Akaike H. (1980). Likelihood and the Bayes Procedure in Bayesian Statistics, in *Bayesian Statistics*, Eds. J.N. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, Valencia, Spain University Press, 141–166.

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

3. Ansley, C. F., Kohn, R. (1983). Exact likelihood of a vector autoregressive process with missing or aggregated data. *Biometrika,* **70**, 275–278.

4. Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis, Forecasting, and Control*, San Francisco Holden–Day.

5. Chatfield, C. (1979). Inverse autocorrelations. *J. Royal Soc., Ser A*, **142**, 363–377.

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

7. Cipra T. and Tlustý, P. (1987). Estimation of multiple autoregressive-moving average models using periodicity. *J. Time Series,* **8**, 293–301.

8. Cipra, T. (1985a) Periodic moving average processes. *Aplikace Matematily* **30**, 218–29.

9. Cipra, T. (1985b) Statistical analysis of multiple moving average processes using periodicity. *Kybernetica* **21**, 335–45.

10. Cleveland W.P. and Tiao G. C. (1979). Modeling Seasonal Time Series. *Revue Ecomonic Appliéd*, **32**, 107–129.

11. Croley and Rao,(1978). A manual for hydrological time series deseasonalizing and serial dependence reduction. *Institute of Hydraulic Research*, report no. **199**, University of Iowa.

12. Craven, P. and Wahba G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31, 370–493.

13. Dagum, Estela B., (1978). Modelling, Forecasting, and Seasonally Adjusting Economic Time Series with the X-11 Arima Method, *The Statistician*, **27**, 203–216.

14. Delleur, J. W., Tao P. C., and Kavvas M. L. (1976). An Evaluation of the Practicality and Complexity of Some Rainfall and Runoff Time Series Models. *Water Resources Research*, **12**, 953–970.

15. Dunsmuir W. (1981). Estimation of Periodically Varying Means and Standard Deviations in Time Series Data. *J. Time Series,* **2**, 3., 129–153.

16. Dunsmuir W. (1983). Time Series Regression with Periodically Correlated error and Missing Data. In *Time Series Analysis of Irregularly Observed Data*, Lecture Notes in Statistics, No. 25

17. Gardner, G., Harvey, A. C., Phillips, G. D. A. (1980). An algorithm for the exact maximum likelihood estimation of autoregressive moving average models by means of Kalman filtering. *Appl. Statistics*, **29**, 311–322.

18. Gladyshev, E. G. (1961). Periodically correlated Random Sequences. *Soviet Mathematics*, **2**, 385–388.

19. Gladyshev, E. G. (1963). Periodically and almost-periodically correlated random processes with a continuous time parameter. *Theory of Probab. and its Appl.*, **8**, 173–177.

20. Harvie, A. C. (1981). *The Econometric Analysis of Time Series.* Oxford, Philip Allen.

21. Harvie, A. C. (1981). *Time Series Models.* Oxford, Philip Allen.

22. Harrison, P. J. and Stevens, C. F. (1976). Bayesian Forecasting. *J. Royal Soc., Ser B,* **38**, 205–247.

23. Hillmer, S. C. and Tiao, G. C. (1982). An ARIMA-Model-Based Approach to Seasonal Adjustment, *Journal of the American Statistical Association*, **77**, 63-70.

24. Hillmer, S. C., and Tiao, G. C. (1979). Likelihood Function Of Stationary Multiple Autoregressive Moving Average Models.*Journal of the American Statistical Association*, **74**, 652–660.

25. Jazwinski, A. H. (1970). *Stochastic processes and filtering theory.* Academic Press, New York.

26. Jones, R. H. and Brelsford, W. M. (1967). Time Series with Periodic Structure. *Biometrika*, **54**, 403–408.

27. Kalman, R. E. (1960). A new basic approach to linear filtering and prediction problems. *Trans. ASME, Journal of Basic Engineering, Ser. D*, **80**, 35–45.

28. Kitagawa G. and Gersch, W. (1984), A Smoothness Priors-State Space Modeling of time Series with Trend and Seasonality, *Journal of the American Statistical Association*, **79**, 378–389.

29. Kimeldorf, G. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, **41**, 495–502.

30. Kohn, R., Ansley, C. F. (1982). A note on obtaining the theoretical

autocovariances of an ARMA process. *J. Statist. Comput. Simul.*, **15**, 273–283.

31. Lawrence, A. J. and Lewis, P. A. W. (1980). The exponential autoregressive moving average EARMA(p, q) process. *J. Royal Soc., Ser. B,* **42**, 150–161.

32. Lawrence, A. J. and Lewis, P. A. W. (1981). A new autoregressive time series model in exponential variables (NEAR(1)). *Advances in Probability*, **13**, 826–845.

33. Lawrence, A. J. and Lewis, P. A. W. (1985). Modelling and residual analysis of nonlinear autoregressive time series in exponential variables. *J. Royal Soc., Ser. B,* **47**, 165–202.

34. Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with Discussion). *J. Royal Soc., Ser. B,* **34**, 1–18.

35. Makridakis, S. and Hibon, M. (1979). Accuracy of forecasting: an empirical investigation (with Discussion). *J. Royal Soc., Ser. A,* **142**, 97-145.

36. Makridakis, S., Andersen, A., Carbone R., Fildes R., Hibon M., Lewandowski, Newton J., Parzen E., and Winkler R. (1982). Accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, **1**, 111-154.

37. Makridakis, S. and Hibon, M. (1979). Accuracy of forecasting: an empirical investigation (with Discussion). *J. Royal Soc., Ser. A*, **142**, 97-145.

38. Mckenzie, (1986) Discrete autoregressive models. In Modelling with Application to Water Resources. *Special Issue on Time Series Analysis in Water Resources*. K. W. Hipel, editor.

39. McLeod A. I. (1975). The derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Appl. Statist.*, **24**, 255–256.

40. McLeod A. I. (1977). Improved Box-Jenkins estimators. *Biometrika*, **64**, 531-534.

41. McLeod A. I. (1984). Duality and other properties of multiplicative autoregressive-moving average models. *Biometrika*, **71**, 207–11

42. McLeod, A. I. and Hipel, K. W. (1978). Developments in monthly autoregressive modelling. Technical report No. 45–XM-011178, Dept. of Systems Design Engineering, Univ. of Waterloo, Ontario, Canada.

43. McLeod, A. I. and Hipel, K. W., and Lennox, W. C. (1977). Advances in Box-Jenkins modelling. 2. Applications. *Water Resourc. Res.*, **13**, 577–586.

44. Menard, and Roy, (1986). Robust estimation of covariances. *J. Statist. Comput. Simul.*, **19**.

45. Morf, M., Sidhu, G. S., and Kailath T. (1974). Some new algorithms for recursive estimation in constant, linear, discrete-time systems. *IEEE*, **AC-19**, 315–323.

46. Morgan, J. A. and Tartar, J. F. (1972). Calculation of the residuals sum of squares for all possible autoregressions. *Technometrics*, **14**, 673–640.

47. Newbold, P. (1974). The exact likelihood function for a mixed autoregressive-moving average models. *Biometrika*, **66**, 265–270.

48. Newton, J. (1982). Using periodic autoregressions for multiple spectral estimation. *Technometrics*, **24**, 109–116.

49. Noakes, D. J., McLeod, A. I., Hipel, K. W. (1979) Forecasting monthly riverflows time series. *International Journal of Forecasting*, **1**, 179–190.

50. Pagano, M. (1978). On Periodic and Multiple Autoregressions. *Annals of Mathematical Statistics*, **6**, 1310–1317.

51. Pearlman, J. G. (1980) An algorithm for the exact likelihood of a high-order autoregressive-moving average process. *Biometrika*, **67**, 232–233.

52. Priestley, M. B. (1980). State-dependent models: A general approach to nonlinear time series analysis. *J. Time Series*, **1**, 47–71.

53. Rao, A. R. and Kashrap, P. L., (1974) Stochastic modeling of riverflows. *IEEE Transactions on Automatic Control*, **AC-19**, 874-888.

54. Salas, J. D., D. C. Boes and Smith R. A. (1982). Estimation of ARMA models with Seasonal Parameters. *Water Resources Research*, **18**, 1006–1010.

55. Salas, J. D., Tabios III, G.Q. and Bartolini, P. (1985). Approaches to Multivariate Modelling of Water Resources Time Series. *Water Resources Bulletin*, **21**, 4.

56. Sakai, H. (1982). Circular Lattice Filtering Using Pagano's Method. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-30**, 279–286.

57. Shea B.L. (1987). Estimation of multivariate time series. *J. of time series*, **8**, 95–111.

58. Shiller, R. (1973). A distributed lag estimator derived from smoothness priors. *Econometrica*, **41**, 775–778.

59. Shiller, R. (1984). Smoothness priors and nonlinear regression. *J. Amer. Statist. Assoc.*, **79**, 609–615.

60. Shiskin, J., Young, A. H. and Musgrave, J. C. (1967). The X-11 Variant of the Census Method II Seasonal Adjustment Program. Technical Paper, Washington, D.C. U.S. Department of Commerce, Bureau of the Census.

61. Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1985a). Grouping of Periodic Autoregressive Models. Time Series Analysis: Theory and Practice 6, edited by O. D. Anderson, J. K. Ord and E. A. Robinson, North Holland, Amsterdam, 35–49.

62. Thompstone, R. M., Hipel, K. W. and McLeod, A. I. (1985b). Forecasting Quaterly-Monthy Riverflow. *Water Resources Bulletin.*, **21**, 5.

63. Tao, P. C., and Delleur, J. W. (1976). Seasonal and nonseasonal ARMA models. *ASCE, Journal of the Hydrology Division*, **102**, 1541–1559.

64. Tiao G. C and Gruppe M. R. (1981). Hidden Periodic Autoregressive-Moving Average Models. *Biometrika*, **67**, 365–373.

65. Tiao, G. C. and Box, G. E. P. (1981). Modelling Multiple Time Series with Applications. *Journal of the American Statistical Association*, **76**, 802–816

66. Tiao, G. C. and Box, G. E. P., Grupe, M. R., Hudak, G. B., Bell, W. R., and Chang, I. (1979). The Wisconsing Multiple Time series (WTMS-1) Program: A Preliminary Guide. Department of Statistics, University of Wisconsin, Madison, Wisconsin.

67. Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *J. Royal Soc., Ser. B*, **42**, 245–292.

68. Tsay, R. S. and Tiao, G. C. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. *J. Amer. Statist. Assoc.*, **79**, 84–96.

69. Vecchia, A. V. J. T. Obeysekera, J. D. Salas, and Boes, D. C. (1983). Aggregation and Estimation for Low-Order Periodic Arma Models. *Water Resourc. Res.*, **19**, 1297–1306.

70. Vecchia, A. V. (1983). Aggregation and Estimation for Periodic Autoregressive-Moving Average Models. Ph. D. dissertation, Dept. of Statistics, Colorado State University, Fort Collins.

71. Vecchia, A. V. (1985). Maximum Likelihood Estimation for Periodic Autoregressive-Moving Average Models. *Technometrics*, **27**, 375-384.

72. Vecchia, A. V. (1986). Periodic Autoregressive-Moving Average (PARMA) Modelling with Applications to Water Resources. In Special publication on Time Series Analysis in Water Resources. K. W. Hipel, editor.

73. Yevjevich (1972) *Stochastic Processes in Hydrology*. Water Resources Publ., Littleton, Colorado.

74. Wahba, G. (1975). Smoothing noisy data with spline functions. *Numerische Matematik*, **24**, 383–393.

75. Wahba G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *J. Royal Soc., Ser. B,* **40**, 364–372.

76. Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *J. Royal Soc., Ser. B,* **45**, 133–150.

77. Wecker, W.P. and Ansley, C. F. (1983) The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, **78**, 81–89.

Chapter 4

ON BOOTSTRAPPING

4.1 Introduction

In order to investigate the medium to large sample behaviour of statistical estimates that are functions of the empirical distribution function F_N of the observations, Efron's (1979) bootstrap has been used with some success. It is possible to give several statistical procedures similar to Efron's (1979) bootstrap and in this chapter we present several of these definitions. These procedures are generally intended for use with a regression type model, and more generally for an optimization functional type of statistical procedure.

4.2 First definition

Suppose that we have N observations, y_i, $i = 1, \ldots, N$ of a where, ϵ_i are independent and identically distributed random variables. The vectors x_i are random variables in $\mathbb{R}^p$, but we allow the possibility that they are degenerate (constant), $(\cdot , \cdot)$ is the inner product in $\mathbb{R}^p$. In order that consistency and asymptotic normality of the estimates hold, we assume certain conditions on x_i and ϵ_i. We will assume that

$$\frac{\mathbf{X}'\mathbf{X}}{N} \rightarrow V, \tag{4.2}$$

where $\mathbf{X}$ is the $N \times p$ matrix with rows equal to x_i; the matrix V is a positive definite matrix. Also, we will assume that $\mathbf{X}$ is of full rank. To estimate θ_o, we form the residuals $e_i, i = 1, \ldots, N$, where

$$e_i = e_i(\theta) = y_i - (x_i, \theta). \tag{4.3}$$

The parameter θ belongs to Θ, a subset of $\mathbb{R}^k$, assumed to be compact. We assume that the true value of θ lies in Θ. We obtain an estimate of θ by minimizing a loss function such as

$$L_n(\theta) = \frac{1}{N}\sum_{1}^{N} \psi(e_i(\theta)). \tag{4.4}$$

The value that minimizes (4.4) will be called $\hat{\theta}$. Initially we will assume that $\psi(e) = e^2$. Hence, we will be working with the usual regression model.

Definition 4.1 **The idea is to randomly perturb the observations y_i, generate residuals with these new observations and again minimize (4.4). This will be accomplished as follows: generate residuals $e_i(w_i, \theta)$ defined by**

$$e_i(w_i, \theta) = y_i^* - (x_i, \theta), \tag{4.5}$$

where,

$$y_i^* = y_i + s_N\chi_i = y_i\omega_i, \tag{4.6}$$

$i = 1, \ldots, N$, the $\chi_i, i = 1, \ldots, N$, are identically and independently distributed random variables with mean 0 and variance 1; and

$$s_N^2 = \frac{1}{N}\sum_1^N e_i^2(\hat{\theta}). \tag{4.7}$$

▽

The explicit expression of $\hat{\theta}_\omega$ is

$$\hat{\theta}_\omega = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t WY, \tag{4.8}$$

where $W = \text{diag}(w_1, \ldots, w_N)$.

THEOREM 4.2 The expectation of $\hat{\theta}_\omega$ is equal to $\hat{\theta}$; and its variance is equal to $s_N^2(X^tX)^{-1}$. ▽

From now on we will consider the centered and standardized estimate $\sqrt{N}(\hat{\theta}_\omega - \hat{\theta})$. Notice that

$$\sqrt{N}(\hat{\theta}_\omega - \hat{\theta}) = \sqrt{N}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\chi. \tag{4.9}$$

Now, because $(\mathbf{X}^t\mathbf{X})^{-1}N$ converges almost surely (a.s.) to a positive definite matrix V as $N \to \infty$, and because $\hat{\theta}$ converges asymptotically to a normal random variable, then $\sqrt{N}(\hat{\theta}_\omega - \hat{\theta})$ also converges asymptotically to a normal random variable with the same limiting multivariate mean (zero), and the same variance matrix V^{-1} as $\sqrt{N}(\hat{\theta} - \theta)$.

THEOREM 4.3 The random variable $\sqrt{N}(\hat{\theta}_\omega - \hat{\theta})$ is asymptotically normal with mean zero and variance matrix V^{-1}. ▽

This result holds irrespective of the distribution of the χ_i's, perhaps we could choose the distribution so that the medium to large sample size distribution properties of both $\sqrt{N}(\hat{\theta}_\omega - \hat{\theta})$ and $\sqrt{N}(\hat{\theta} - \theta)$ are similar. We could envision the χ_i's as residuals and hence a reasonable choice of the distribution of χ_i's is to set it equal to the empirical distribution of the estimated residuals $e_i(\hat{\theta})$, mean centered, or a smoother version of this empirical distribution function.

REMARK 4.4 Freedman's (1981) definition of bootstrapping for a regression model, i.e. when **X** it is not random, is as follows. Generate residuals $e_i^*, i = 1, \ldots, N$ with distribution function equal to the empirical distribution function, (e.d.f.), of the estimated residuals $e_i(\hat{\theta})$. Generate new observations $y_i^*, i = 1, \ldots, N$

$$y_i^* = (x_i, \hat{\theta}) + e_i^*. \tag{4.10}$$

Notice that

$$\begin{aligned} y_i^* &= y_i + (x_i, (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\epsilon) + e_i^*, \\ y_i^* &= y_i - \hat{e}_i + e_i^* \end{aligned} \tag{4.11}$$

whereas (Definition 4.1)

$$y_i w_i = y_i + \chi_i. \tag{4.12}$$

Hence, the two bootstrapping procedures are similar in spirit but not identical. However, if the model assumed is correct both definitions lead to

Hence, the two bootstrapping procedures are similar in spirit but not identical. However, if the model assumed is correct both definitions lead to the same bootstrapped estimate. The advantage of the present definition is clear when we want to generalize bootstrapping to time series models. To keep with the regression setting that we have been using consider the generalization of bootstrap to autoregressive processes. The definition of the bootstrap presented in this chapter carries over exactly as stated. Using Freedman's (1981) definition we generate the new observations by feeding into the estimated model residuals that have the same distribution as the centered-estimated residuals. Then, for example, if the model that we have assumed is not correct, the generalization of Freedman's definition of the regression bootstrap will not give any indication of this fact as we are generating observations from the assumed model. Even if the assumed family of models is correct, errors in the estimated residuals, in the sense that the distribution of the estimated residuals $e_i(\hat{\theta})$ may be a poor estimation of the true distribution function of the innovations ϵ_i, will enter the model in two different forms. With the definition of bootstrapping presented in this chapter, the resampled residuals enter the model in an additive form hence unusual residuals with have only a local effect in the generation of new observations. However, with the generalization of Freedman's bootstrap the resampled residuals enter the model in an innovation form. Hence, the effect on the new observations with be present in all future computations. Moreover, there is a computational advantage in using (4.12) versus (4.11)

4.2.1 Asymptotic analysis

We are assuming that the random variable $\sqrt{N}(\hat{\theta} - \theta)$ is asymptotically normal, with mean zero and variance matrix equal to $\sigma^2 V^{-1}$

4.3 Alternative definition

In this section we will present yet another definition of the bootstrap. It is based on an idea of Cover and Unny (1986). Suppose that one has N observations $y_1, \ldots, y_N$, and that one wants to estimate a parameter θ using the loss function $L(\theta, F_N)$, where F_N is the nonparametric estimate of the cumulative distribution function of the observations $y_1, \ldots, y_N$. Assume that $L(\theta, F_N)$ can be written as

$$L(\theta, F_N) = \sum_{i=1}^{N} T(y_i, \theta). \tag{4.13}$$

Now the idea of the new definition of bootstrap is that instead of minimizing the function $L(\theta, F_N)$ (4.13) we minimize the following random loss function

$$L(\theta, F_N, \omega) = \frac{1}{N} \sum_{i=1}^{N} \omega_i T(y_i, \theta) \tag{4.14}$$

where ω is a vector of random weights that satisfy some conditions that will be specified later.

We will write $\hat{\theta}_\omega$ for the value of θ that minimizes $L(\theta, F_N, \omega)$ in Θ, and we will write $\hat{\theta}$ for the corresponding value with respect to $L(\theta, F_N)$.

Note that both values $\hat{\theta}_\omega$ and $\hat{\theta}$ depend on N, but we will note state it explicitly.

First, we will give some examples.

EXAMPLE 4.5 Suppose that that observations y_i are independent and identically distributed. Suppose that the random weights are such that their sum is equal to M, and their are chosen, subject to this condition, uniformly over the integers $0, \ldots, M$. Then, the result is the classical bootstrap.

Efron (1979) and others have also introduced smoother versions of the bootstrap. The idea is to use a smoothed version of the cumulative distribution function F_N. To obtain a smooth version of F_N they generally convolute F_N with a smooth density, such as the normal density, to obtain $\tilde{F}_N = \phi_\lambda * F_N$, where ϕ_λ is a smooth density, and λ is a parameter that tends to zero such that asymptotically $\tilde{F}_N$ tends to the true F, together with its derivatives. If in (4.14), the version of the bootstrap that we have just given, we use weights that are distributed randomly under some sum restriction, we obtain the estimates that are similar to the classical smoothed bootstrap estimates. Many statistical procedures are defined naturally in terms of minimizing a loss function. Even the estimation of the sample mean of independent and identically distributed random variables can be seen as the minimization of a sum of squares of residuals. Hence, it is of interest to bootstrap the loss function.

EXAMPLE 4.6 Let $y_1, \ldots, y_N$, be N independent observations of a random variable with mean μ_0. Then the bootstrap estimate of μ_0 is obtained by minimizing

$$\sum_1^N (y_i - \mu)^2 w_i = \min!, \tag{4.15}$$

with respect to μ, where w_i are random weights. The estimate $\hat{\mu}_w$ is given by

$$\hat{\mu}_w = \frac{\sum_1^N w_i y_i}{\sum w_i} \tag{4.16}$$

EXAMPLE 4.7 Let (y_i, x_i), be N observations of a bivariate random variable with correlation coefficient ρ_0. The usual estimate of ρ_0 may be obtained using a regression model approach. Hence, we assume that

$$y_i = \rho_0 x_i + e_i. \tag{4.17}$$

Assuming that the errors e_i are independent and identically distributed, the bootstrap estimate of ρ_0 is obtained by minimizing

$$\sum_1^N (y_i - \rho x_i)^2 w_i, \tag{4.18}$$

where we assume that the observations (y_i, x_i) are centered around the sample mean $(\bar{y}, \bar{x})$, and the w_i are random weights. The explicit estimate is given by

$$\hat{\rho}_w = \frac{\sum_1^N w_i y_i x_i}{\sum_1^N w_i x_i^2}. \tag{4.19}$$

EXAMPLE 4.8 Let $y_1, \ldots, y_N$, be N independent observations of the following regression model

$$y_i = x_i' \beta_0 + e_i, \tag{4.20}$$

where x_i' is a vector in $\boldsymbol{R}^k$ then the minimization bootstrap is the solution of

$$\sum_1^N (y_i - x_i'\beta)^2 w_i = (Y - X\beta)'W(Y - X\beta) = \min!, \qquad (4.21)$$

with respect to β, where the w_i are random weights, X is the matrix formed with its rows equal to x_i', Y is the column vector of the observations, and W is a diagonal matrix with entries equal to w_i. The explicit form of the estimate is given by

$$\hat{\beta}_w = (X'WX)^{-1}X'WY. \qquad (4.22)$$

4.3.1 Conditions on the random weights

We want to minimize the random loss function (in 4.14). However it is obvious that the minimizer will not change if we divide each weight ω_i by the same constant α. Hence, either by taking $\alpha = \sum \omega_i$ or $\alpha = \max \omega_i$ we see that we could assume that $\omega_i \leq \rho < \infty$, where ρ is some constant. Because the random weights are bounded, we can use powerful inequalities such as the Hoeffding's inequality (Hoeffding, 1956). We will assume that the random weights are bounded form now on.

Also, we want the expectation of $L(\theta, F_N, \omega)$ to be equal to $L(\hat{\theta}, F_N)$. Hence, it is reasonable to set $\boldsymbol{E}\{w_i\} = 1$, for $i = 1, \ldots, N$.

4.3.2 Analysis of the Bootstrap

First we want to prove asymptotic consistency. We give some notation. Let $V_{N,\omega}(\theta)$ be equal to $L(\theta, F_N, \omega) - L(\theta, F_N)$. Then define

$$\|f\|_\Theta = \sup_{\theta\in\Theta} |f(\theta)|. \tag{4.23}$$

We will write $T_i(\theta)$ for $T(x_i, \theta)$ and

$$s_N(\theta) = \frac{1}{N} \sum T_i^2(\theta). \tag{4.24}$$

A generic constant will be always denoted by c. To prove consistency we want to prove a Glivenko-Cantelli type of theorem, i.e. we want to prove that $\|V_{N,\omega}(\theta)\|_\Theta \to 0$ as $N \to \infty$.

THEOREM 4.9 Assume that Θ is a compact. Assume that $T_i(\theta)$ is locally Lipschitz and such that $s_N(\theta)$ converges a.s. . Then

$$\|V_{N,\omega}(\theta)\|_\Theta \to 0 \qquad \text{as} \qquad N \to \infty. \tag{4.25}$$

$\bigtriangledown$

Proof. Because Θ is compact for any $\delta > 0$ there is a finite cover of Θ with balls of radius less or equal to δ and centers θ_α, denote by B_α above the balls. Then

$$\boldsymbol{P}\{\|V_N(\theta)\|_\Theta > \epsilon\} = \sum_\alpha \boldsymbol{P}\{\|V_N(\theta)\|_\alpha > \epsilon\} \tag{4.26}$$

but

$$\boldsymbol{P}\{\|V_N(\theta)\|_\alpha > \epsilon\} \leq \boldsymbol{P}\{V_{n,\omega}(\theta_\alpha) > \epsilon\} + \boldsymbol{P}\{\|V_{n,\omega} - V_{n,\omega}(\theta_\alpha)\|_{B_\alpha} > \epsilon\}. \tag{4.27}$$

Consider the first term of the R.H.S. of (4.27). Then by Hoeffding's inequality (Hoeffding, 1956)

$$\mathbb{P}\{V_{n,\omega}(\theta_\alpha) > \epsilon\} \leq 2\exp\{\frac{-c\epsilon^2 N}{s_N(\theta)}\}, \tag{4.28}$$

where c is constant that depends on the bounds of ω_i. Now, consider the second term of the right hand side of (4.27)

$$V_{N,\omega}(\theta) - V_{n,\omega}(\theta_\alpha) = \frac{1}{N}\sum \omega_i|T_i(\theta) - T_i(\theta_\alpha)|. \tag{4.29}$$

But because T_i is locally Lipschitz, if the radius of the balls is small enough

$$|T_i(\theta) - T_i(\theta_\alpha)| \leq d\rho^\psi, \tag{4.30}$$

for some constant ψ. Then again using Hoeffding inequality

$$\mathbb{P}\{\|V_{n,\omega}() - V_{N,\omega}(\theta)\theta_\alpha\|_{B_\alpha} > \epsilon\} \leq 2\exp\{-d\epsilon^2 N\} \tag{4.31}$$

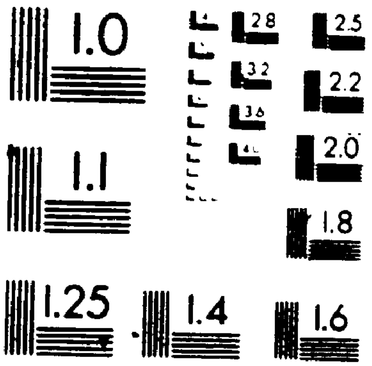
Then by (4.27), (4.28) and (4.31), we obtain

$$\mathbb{P}\{\|V_N(\theta)\|_\Theta > \epsilon\} \leq c(\epsilon)\exp\{-c\epsilon^2 N\}. \tag{4.32}$$

Hence, the Borel-Cantelli theorem implies that

$$\|V_{N,\omega}(\theta)\|_\Theta \to 0 \quad \text{as} \quad N \to \infty \quad \text{a.s.} \tag{4.33}$$

REMARK 4.10 The conditions of the theorem are fulfilled in many common situations, as in multiple regression with the usual conditions in the matrix of independent variables, and also in the case of ARMA time series models.


1.0
2.8
2.5
3.2
2.2
3.6
2.0
1.1
1.8
1.25
1.4
1.6

Now we will consider the asymptotic normality of $\hat{\theta}_\omega$.

THEOREM 4.11 We assume the same conditions on ω_i, T_i and Θ as in theorem (4.9). Also, we assume conditions on T_i sufficient to ensure the asymptotic normality of $\hat{\theta}$ with covariance matrix

$$I^{-1} = \boldsymbol{E}\{\varphi\} \tag{4.34}$$

Also, assume that T_i can be expanded such that

$$\hat{\theta}_n - \hat{\theta}_0 = (\frac{1}{n}\sum \phi_i'(\theta_0))^{-1}(\frac{1}{n}\sum \phi_i(\theta_0)) + \ldots.$$

such that $\boldsymbol{E}|\phi'(\theta_0; X)|$ exists, and $\phi(\theta_0; X)$ satisfies conditions that assure the Central Limit Theorem. Then $\sqrt{N}(\hat{\theta}_n - \theta_0)$ and $\sqrt{N}(\hat{\theta}_w - \hat{\theta}_N)$ have asymptotically the same distribution. Moreover, if $\boldsymbol{E}\phi^3(\theta_0; x)$ is defined, then the distance between their probability laws is of order $1/\sqrt{n}$. ▽

4.3.3 The mean

We have observations $y_i, i = 1, \ldots, N$, with mean μ, and finite variance σ_i^2, with out loss of generality we will assume that the unknown true mean $\mu = 0$. The estimate of μ is obtained by minimizing

$$\sum_{i=1}^{N} \omega_i (y_i - \mu)^2 \tag{4.35}$$

where the weights ω_i are positive independent and identically distributed random variables, with mean one and finite variance σ_ω^2. The estimate corresponding to the set of weights $\omega = (\omega_i, \ldots, \omega_N)$ is given by

$$\begin{aligned} \hat{\mu}_w &= \frac{\sum_{i=1}^{N} \omega_i y_i}{\sum \omega_i} \\ &= \bar{y} + \frac{\sum_{i=1}^{N} \omega_i (y_i - \bar{y})}{\sum_{i=1}^{N} \omega_i}. \end{aligned} \tag{4.36}$$

Notice that

$$\sum_{i=1}^{N} \lambda_i = 1,$$

and that λ_i are exchangeable random variables. Then we have the following result.

THEOREM **4.12** The mean of $\hat{\mu}_\omega$ is equal to

$$\mathbb{E}\{\hat{\mu}_\omega\} = \bar{y}. \tag{4.38}$$

The variance of $\hat{\mu}_\omega$ is equal to

$$\text{var}\{\hat{\mu}_\omega\} = (1 - N^2 \mathbb{E}\{\lambda_1 \lambda_2\}) \frac{1}{N} \sum_{i=1}^{N} e_i^2, \tag{4.39}$$

where, $e_i = y_i - \bar{y}_i$. Hence,

$$\text{var}\{\hat{\mu}_\omega\} \leq \frac{1}{N} \sum_{i=1}^{N} e_i^2, \tag{4.40}$$

and $\hat{\mu}_\omega \to 0$ as $N \to \infty$. ▽

Next, we are interested in the asymptotic behaviour of $\mu_\omega - \bar{y}$. The denominator of μ_ω is not N, however by the Strong Law of the Large Numbers, (SLLN)

$$\frac{\sum \omega_i}{N} \to 1, \qquad \text{a.s.}. \tag{4.41}$$

Also, by Bernstein's inequality the probability that this sum deviates from one by more than λ, converges exponentially to zero, as the sample size increases. Therefore, we will assume that the denominator of μ_ω is equal to N. Then,

$$\hat{\mu}_\omega - \bar{y} = \frac{\sum \tilde{\omega} y_i}{N}, \tag{4.42}$$

Also, by Bernstein's inequality the probability that this sum deviates from one by more than λ, converges exponentially to zero, as the sample size increases. Therefore, we will assume that the denominator of μ_ω is equal to N. Then,

$$\hat{\mu}_\omega - \bar{y} = \frac{\sum \tilde{\omega} y_i}{N}, \tag{4.42}$$

and $\tilde{\omega} = \omega - 1$. We will write ω for $\tilde{\omega}$ from now on.

Next, use the Berry-Esséen Theorem to investigate the asymptotic behaviour of $\hat{\mu}_\omega - \bar{y}$. Thus, the following it is true

THEOREM **4.13** Let y_i be independent observations of random variables with finite variance and common mean μ. Suppose that the random weights are positive, independent and identically distributed random variables with mean and variance one. Then the following holds true

$$\sup_z |\boldsymbol{P}(\frac{\hat{\mu}_\omega - \hat{\mu}}{\sigma_w} \leq z) - \Phi(z)| \leq \alpha \frac{\sum e_i^3}{(\sum e_i^2)^{3/2}} \tag{4.43}$$

where $\sigma_w^2 = \frac{1}{N}\sum_{i=1}^N \epsilon_i^2$, and α is an universal constant.

REMARK **4.14** We could of course use random variables ω_i, that are independent and normally distributed with mean and variance one. In this case the distribution of (4.42) is Gaussian.

Data set Identification	Mean of $\hat{d}$	St. deviation	$\hat{d}$	St deviation
Saugeen	0.110	0.212	0.108	0.100
Dal	0.028	0.177	0.024	0.093
Danube	0.069	0.157	0.059	0.072
French	0.149	0.168	0.134	0.093
Gota	0.365	0.245	0.388	0.064
Mckenzie	0.284	0.142	0.274	0.105
Neumunas	0.105	0.137	0.103	0.068
Rain Phil.	0.210	0.110	0.229	0.078
St. Lawrence	0.495	0.055	0.499	0.080
Thames	0.139	C.149	0.120	0.093
Temperature	0.153	0.079	0.151	0.050

Table 3.1. Estimation of the parameter d using Bootstrapping

Using the bootstrapping technique described in the chapter, the value of the persistence parameter d in the model $\nabla^d(B)x_t = a_t$ was estimated by the mean value of the estimates obtained using the bootstrap, this value is given in the second column. The third column gives the standard deviation of the estimte obtained by bootstrapping. The fourth column lists $\hat{d}$ which is the estimate of d obtained by the maximun likelihood method described in the chapter and the last column gives the asymptotic standard deviation of $\hat{d}$ (this is equal to $\sqrt{6/\pi^2 N}$).

4.4 Applications

We will give some applications of the bootstrapping procedure discussed in this chapter. We consider the fourteen data sets given in Table 1. of chapter 1. It is of interest to consider how the bootstrap behaves if we fit to these data sets a fractionally differenced noise model, FARMA(0, d, 0). Generally, we are interested in the behaviour of the bootstrapped values when the model that we are fitting is correct. But also, we are interested in the behaviour of the bootstrap when the model is not correct. We bootstrapped the residual sum of squares for each data set and we obtained the mean of the bootstrapped fractional differencing parameter, this mean should be close to the estimated value. We obtained the standard deviation of the bootstrapped fractionally differencing parameter. These quantities are given in Table 4.1. We may observe that for the data-sets where the fractionally differencing noise model is not appropriate we standard deviation is much larger than the standard deviation obtained using the asymptotic normal theory. This is particular important for data sets such as the Gota river where using the standard deviation alone is not sufficient for rejecting a FARMA(0, d, 0) model. However in this case it is clear from a residual analysis that a FARMA(0, d, 0) model is not appropriate for the Gota river. A similar comment is true for Philadelphia's annual rainfall. However using the

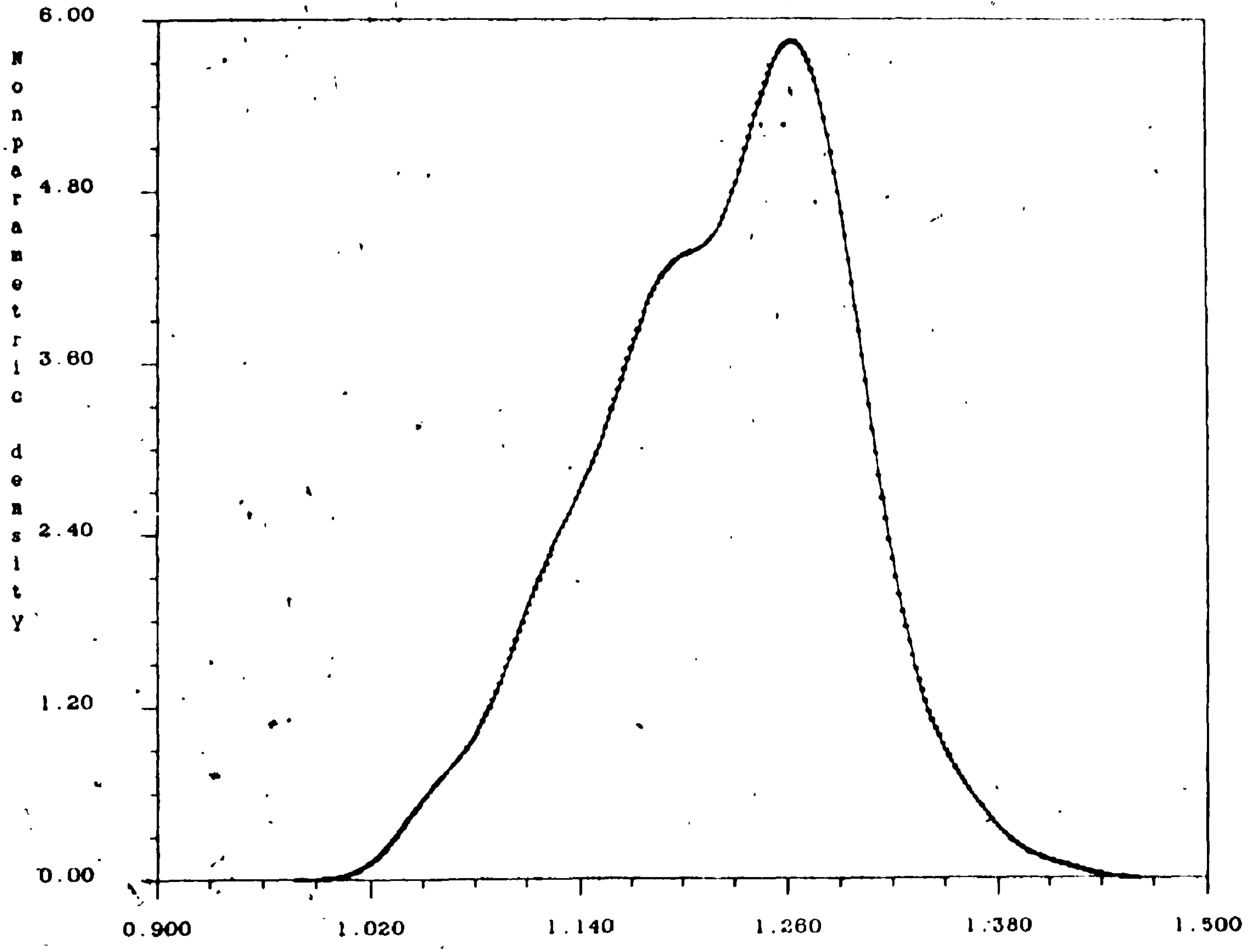


Figure 4.1 Density of the first coefficient of an AR(11) model of the Sunspots by bootstrapping.

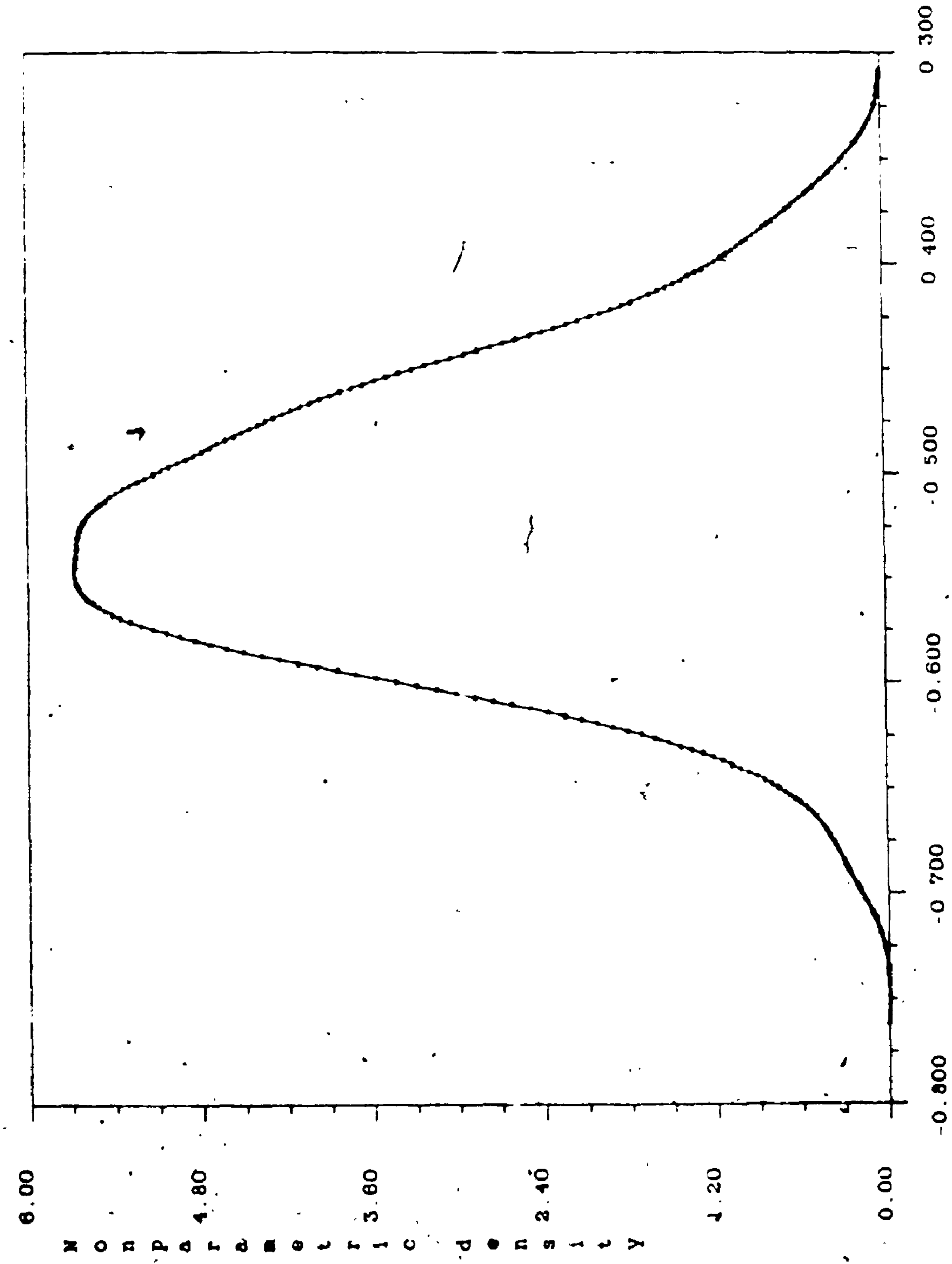


Figure 4.2 Density of thesecond coefficient of an AR(11) model for the Sunspots by bootstrapping.

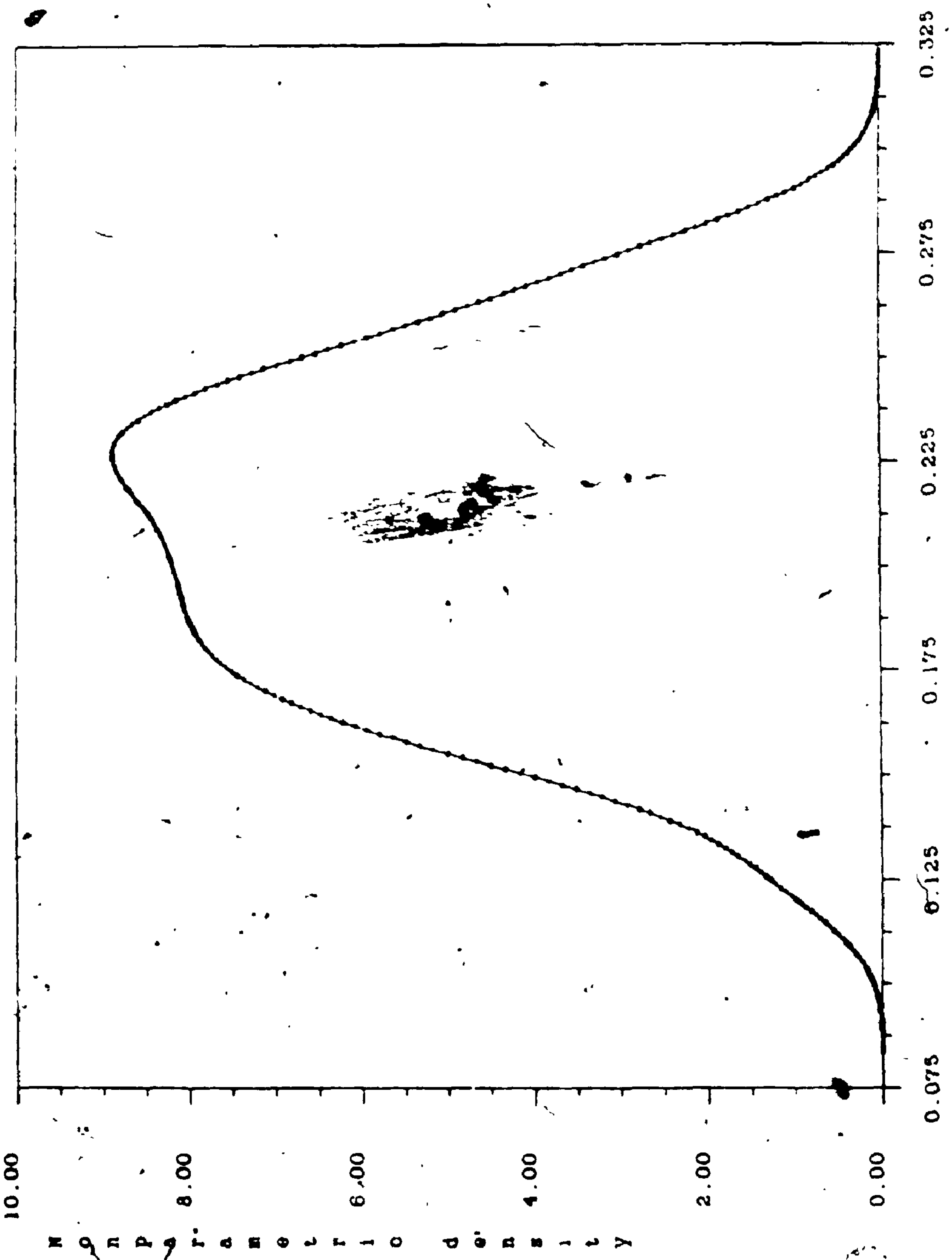


Figure 4.3 Density of the last coefficient of an AR(11) model for the Sunspots by bootstrapping.

bootstrap we are able to reject the null hypothesis of a FARMA$(0, d, 0)$ process quite conclusively. For the case of the St. Lawrence river the standard deviation of the bootstrapped values is more than 20% smaller than the standard deviation obtained using the asymptotic normal theory. The Central England temperature record is one set where we can not give a conclusive result. While the asymptotic normal theory does not reject a FARMA$(0, d, 0)$ model the bootstrapping procedure gives as a borderline case for rejection at the 5% level.

As another application of the bootstrapping technique that we have proposed in this chapter we will consider the estimation of the sunspots series. This time series record is well known among time series researches. A possible model for fitting to the sunspot numbers is a AR(11), with all the coefficients set equal to zero, except the coefficients corresponding to the first, second, and eleventh lags. The bootstrap technique described above was employed to find an approximate expression for the density of each coefficient in the AR(11) model that we fitting. The graphs of these bootstrapped densities is given in Figures 4.1 to 4.3. It is evident, just by looking at these graphs, that for the given sample size the distribution of the estimated AR coefficients is not close to a normal distribution. These plots also suggest that the densities of the estimated coefficients may arise by superposition of two separate probability densities. Note, also, that symmetric confidence intervals do not seem appropriate because the densities are skew.

4.5 Summary

The extension of the bootstrap to a regression like context is not straightforward. We present two alternatives to the definitions in the literature. One is based on the idea of perturbing the observations. The other definition is based on bootstrapping the loss function when the parameters are estimated by minimizing a loss function.

We prove a general theorem to show that the asymptotic distribution of the bootstrap coincides with the asymptotic distribution of the estimated parameter. Also, we give some results in the spirit of Singh (1981) to prove that the bootstrap coincides with a random Edgeworth expansion of the distribution of the estimated parameter. For the mean of a stochastic process, we present more accurate results. Finally we present some applications of bootstrapping time series models.

References

1. Abramovitch, L. and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap. *Annals of Mathematical Statistics,* **13**, 116-132.

2. Babu, G, J. and Singh, K. (1982). Edgeworth expansions for sampling without replacement from finite populations. *J. Multivariate Anal.*,

3. Babu, G, J. and Singh, K. (1983). Inference on means using the bootstrap. *Annals of Mathematical Statistics*, **11**, 999–1003.

4. Cover, K. A. and Unny, T. E. (1986). Application of computer intensive statistics to·parameter uncertainty in streamflow synthesis. *Water Resources Bulletin*, **22**, 495–499.

5. Efron, R. (1979). Bootstrap methods: another look at the Jacknife. *Annals of Mathematical Statistics*, **7**, 1–26.

6. Efron, B. (1981). Nonparametric estimates of the standard error: the jacknife, the bootstrap and other methods, *Biometrika,* **68**, 589-599.

7. Efron, B. (1982). *The Jacknife, the Bootstrap and other Resampling Plans*. Philadelphia: SIAM.

8. Freedman, D. A. (1981). Bootstrapping Regression Models. *Annals of Mathematical Statistics*, **9**, 1218–1228.

9. Hall, P. (1983). Inverting an Edgeworth expansion. *Annals of Mathematical Statistics*, **11**, 569–576.

10. Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. of the American Statistical Association*, **58**, 13–30

11. Singh, K. (1981). On the accuracy of Efron's bootstrap. *Annals of Mathematical Statistics*, **9**, 1187–1195.

Chapter 5

NONPARAMETRIC FUNCTION FITTING

5.1 Introduction

The use of parametric models has been until relatively recently the only feasible way of estimating functions. Most of the statistical estimation of functions has been done within the framework of parametric families and within these parametric families the Gaussian family of probability estimates has been extensively used. Although in many situations there are many valid reasons for using a particular parametric family, it is also true that in many other situations it is not known to which parametric family the probability distribution of the data belongs. Hence, it seems that there is room in the statistical literature for methods of estimation that do not rely upon a parametric family of probability distribution functions to fit the data. All the techniques used to obtain either a density or a regression type function without using a parametric family are given the generic name of nonparametric function fitting techniques. In this chapter, we are concerned with these techniques, especially in the case of stochastic processes.

There are two important situations for using nonparametric function fitting: (a) when we are interested in obtaining a probability density; (b) when we are interested in obtaining a regression type function, i.e., a regression function, hazard function, discriminant function, etc. We will study both of the above cases.

The first technique to obtain nonparametric densities may have been the histogram or something close to it. But in reality, the first paper that was concerned about the problem of nonparametric function estimation was Rosenblatt's (1956), where he showed the nonexistence of an unbiased estimator of a density function, under general conditions.

Nonparametric families

Nonparametric function fitting does not assume a family of models that may be parametrized with a finite (and small) number of coordinates. Instead, it assumes that the unknown function belongs to a particular set of functions. For example, we may desire that the estimate we are looking for to belong to $\mathcal{C}[a,b]$, or $\mathcal{C}^{(k)}[a,b]$, or convex functions, or monotone functions, or piece-wise functions, etc.. As the examples just given show, the emphasis is on some general topological properties that the unknown function should possess.

The main problem with the specification of families of functions based on topological properties is that these families contain too many members and, moreover, most of the functions belonging to these families are close to being non smooth functions. Hence, one must choose with great care the family of functions that may contain the unknown function. This problem is similar to the problem of finding solutions of operator equations, where we must choose the range and the domain of the operator with care in order for a solution to exist and/or be unique. Therefore, to decide to which family f may belong we are interested in specifying the smallest possible family and at the same time a priori knowledge of the unknown function must be satisfied. Small in this context means two things: (a) it means that we will work within the smallest family under inclusion, i.e., if the function is assumed to be differentiable, we will work within the family of differentiable functions and not,.for example within the family of continuous functions; note, that this refers to the type of norm with which we want to measure the distance between two functions; (b) another way of measuring how small is a set of functions is by the use of the Kolmogorov's metric entropy, or equivalently by the capacity of of the set.

DEFINITION 5.1 *Metric entropy of a set of functions $\mathcal{F}$ with metric d.* Let $N(\epsilon)$ be the minimum integer such that there is a subset $\mathcal{A}$ of $\mathcal{F}$, of cardinality N such that any function of $\mathcal{F}$ is within ϵ of some function in $\mathcal{A}$. The logarithm of N is the metric entropy function of the set $\mathcal{F}$ with respect to the metric d. The capacity $C(\epsilon)$ of $\mathcal{F}$ is defined as the maximum number of functions in $\mathcal{F}$ that are at a distance at least ϵ between them. Note that this N and C are equivalent.

REMARK **5.2** Generally it is difficult to compute the metric entropy of a particular family of functions, $\mathcal{A}$. However, if its upper bounds are known, many times these are very useful to prove functional limit theorems relating to the family $\mathcal{A}$, as done, for example by Vapnik and Cervonenkis (1971,1981).

Two other considerations that are generally taken into account in order to decide within which set of function we should estimate the function are : (a) the visual "smoothness" of the function. This is because many times we are interested in plotting the function and it seems easier to comprehend slowly changing functions versus functions with many valleys and crests. Also, it seems to be more visually appealing to have functions without many corners and discontinuities versus functions that do not present these characteristics. (b) Another aspect that has to be taken into account is the speed of computation. However this is largely incompatible with the computation of a "smooth" function. For example, it is much faster to compute piece-wise linear functions than quadratic polynomials.

5.2 Nonparametric function fitting and approximation of functions

The topological properties of the unknown function f are included in the set $\mathcal{H}$ of functions f it is assumed to be in. However, many times we do not work with $\mathcal{H}$ directly but with a subset S_n of $\mathcal{H}$ such that as $n \rightarrow \infty, S_n \rightarrow S$ where S is a dense subset of $\mathcal{H}$. Thus, the idea is similar to the part of mathematics known as "approximation of functions" and shares many of its ideas with it. The main difference is that in the approximation of functions the values of the function, or quantities related to it, are observed without random error, whereas in nonparametric function fitting, the values of the function are either observed with random error or not observed directly but through a related quantity. Another important difference is that the data available in nonparametric function fitting at a given instant is limited and it is either impossible or impractical to obtain more values. This implies that the "size" of S_n, $(size)$, cannot be as large as we would like because the variability of the estimates will increase for a fixed number of data points with $size(S_n)$. This means that we have to extract more information using the same amount of observations. However, as $size(S_n)$ increases the probability that the unknown function lies within S_n increases. Then the $size(S_n)$ has to be chosen in such a way as to counterbalance these two tendencies. This is formalized using the variance and the bias of the estimates.

REMARK 5.3 Many times in statistics we want to optimize a criterion function. Generally, this criterion is a single valued function. However, there are cases where we want to use multivalued criterion functions, i.e multiple criteria. This is the case above where the two criteria are the variance and bias of the estimator. The optimization and statistics literature generally deals with the single valued criterion case. This is so because there is not a 'best' value solution of multivalued criterion problems but rather we have to work with nondominated solutions, efficient solutions, etc. The usual strategy when dealing with multicriterion problems has been to optimize a weighted sum of the individual criteria. For the case of nonparametric function estimation, we could use the MSE of the estimates. In statistics, it has also been common to work with unbiased estimators, i.e., with a subset of the set of efficient solutions. In nonparametric function fitting, there are not unbiased estimators for the whole unknown function f (Rosenblatt, 1956; Prakasa Rao, 1983). Hence, we cannot use this approach. We will encounter other cases of this situation later with splines. Also, in view of the recent developments in the field of multiple criteria optimization, it may be interesting to relate them to statistical procedures.

5.3 Some basic strategies for approximation of functions

5.3.1 Kernel estimates

From the point of view of approximation of functions the basic theory of kernel estimates is given by Shapiro (1966). Also, interesting results about the differentiation of functions can be found in Devroye and Györfi (1984). From the point of view of statistical estimation, the initial results are in Rosenblatt (1956) and Parzen (1962). The basic idea is to convolve a function with another function called the kernel. There is also a parameter λ called the smoothing parameter. The relationship between the kernel and the smoothing parameter is such that as $\lambda \to 0$, the kernel function concentrates all its mass at a single point, i.e., the kernel function "peaks up" to a Dirac delta function. Thus, as $\lambda \to 0$ the convolution approaches the function. This is familiar in Fourier series where we work with the Fejer kernel, or the Poisson kernel. It has an old connection with statistics through the estimation of spectral frequencies by nonparametric methods.

Let f and k be functions. Define the following function by convolution

$$\hat{f}_\lambda(x) = \int f(y) K_\lambda(x - y)\, dy, \tag{5.1}$$

where $K_\lambda(u) = \lambda^{-1} K(u/\lambda)$. Then, if K satisfies some reasonable conditions $\hat{f}(x) \to f(x)$. Moreover, if f is uniformly continuous then under similar conditions on K as in the case of pointwise convergence, the convergence of $\hat{f}(x)$ to $f(x)$ is uniform. Some examples of typical conditions where uniform convergence occurs are

$$\begin{array}{lll} (C.1) & |k(x)|x^{i} \to 0 & \text{as } x \to \infty; k(x) \in \mathbf{L}^1;\ \int k(x)\, dx = 1, \\ (C.2) & k(x), & \text{a density function.} \end{array}$$

We will say that a function K is a kernel if satisfies any of these two conditions.

An important property of the convolved function $\hat{f}_\lambda$ is that it inherits some of the topological properties of the convolution kernel. This implies that a "rough" function can be smoothed as much as we like by convolution with a smooth kernel. Moreover, as $\lambda \to 0$, the smoothed version of f approaches f at almost every point, and in many case uniformly.

THEOREM 5.4 (Shapiro, 1966) Let K be an integrable function on $\mathbb{R}$. Let k be a $\mathcal{C}^{(k)}(\mathbb{R})$ kernel, then $\hat{f}_\lambda$ is a $\mathcal{C}^{(k)}(\mathbb{R})$ function. ▽

THEOREM 5.5 Let f be an integrable function on $\mathbb{R}$. Let K be a kernel, then

(a) If f is convex (concave) and K is a non-negative kernel then $\hat{f}_\lambda$ is convex (concave).

(b) If K is convex (concave) and f is nonnegative then $\hat{f}_\lambda$ is convex (concave).

(c) If f is a monotone function in $\mathbb{R}$, and K is a non-negative kernel, then $\hat{f}_\lambda$ is also a monotone function.

(d) If f is an even (odd) function and $\hat{f}_\lambda$ is an even kernel then $\hat{f}_\lambda$ is an even (odd) function.

▽

Proof

(a) If f is convex, then for $0 \le \alpha \le 1$

$$\begin{aligned} \hat{f}_\lambda(\alpha x + (1-\alpha)y) &= \int f(\alpha x + (1-\alpha)y - u)K_\lambda(u)\,du \\ &= \int f(\alpha(x-u) + (1-\alpha)(y-u)K_\lambda(u)\,du \\ &\le \alpha \int f(x-u)K_\lambda(u)\,du \\ &\quad +(1-\alpha)\int f(y-u)K_\lambda(u)\,du \\ &= \alpha\hat{f}_\lambda(x) + (1-\alpha)\hat{f}_\lambda(y) \end{aligned} \tag{5.2}$$

(b) If K is convex the proof is similar to (a).

(c) Assume that f is increasing. Then if $x > y$

$$\hat{f}_\lambda(x) - \hat{f}_\lambda(y) = \int \left(f(x-u) - f(y-u)\right) K_\lambda(u)du \ge 0. \tag{5.3}$$

(d) If f is even then

$$\begin{aligned} \hat{f}_\lambda(x) &= \int f(y)K_\lambda(x-y)\,dy \\ &= \int f(-y)K_\lambda(x-y)\,dy \\ &= \int f(-y)K_\lambda(-x+y)\,dy \\ &= \int f(y)K_\lambda(-x-y)\,dy \\ &= \hat{f}_\lambda(-x) \end{aligned} \tag{5.4}$$

□

COROLLARY 5.6 **If f is a linear combination of functions $f_1, \ldots, f_n$, that fall into the categories (a) to (d) and K_λ is an appropriate kernel function, then $\hat{f}_\lambda$ will also be a linear combination, with the same coefficients as f, of functions $\hat{f}_1, \ldots, \hat{f}_n$, that fall in the same category as $f_1, \ldots, f_n$.** ▽

THEOREM 5.7 Let f be a $\mathcal{C}^{(k)}[a,b]$ function. Let K be a kernel with compact support. Suppose that sign of $f^{(k)}$ is constant on $I = [c,d]$ then the sign of $\hat{f}_{\lambda}^{(k)}$ will be constant on $I(\lambda) = [c(\lambda), d(\lambda)]$, with $I(\lambda)$ included on I and $c(\lambda) \to c, d(\lambda) \to d$ as $\lambda \to 0$. ∇

Proof. This theorem is an immediate consequence of the fact that $\hat{f}^{(k)} \to f^{(k)}$ uniformly (Shapiro, 1966).

□

REMARK 5.8 The "look" or shape of f depends on properties such as convexity or monotonicity, and on its derivatives. It follows that these properties will be shared by $\hat{f}_\lambda$ for any λ or as soon as λ is small enough.

Theorem 5.7 remains true for functions defined on $\mathbb{R}^n$. Parts (a) and (b) of Theorem 5.5 are true if f is componentwise or globally convex (concave). Part (c) of Theorem 5.5 remains true for functions that are componentwise monotone. And finally, if f is a radial function, i.e. $f(\underline{x}) = f(\|x\|)$, then $\hat{f}_\lambda$ is also a radial function.

There are also a collection of results about the speed of convergence of $\hat{f}_\lambda$ to f. The most interesting result is that there are saturation families for each kernel

5.3.2 Kernel Density Estimates

Suppose we observe n values $x_1, \ldots, x_n$, of a stochastic process with a stationary marginal density function f. Based on $x_1, \ldots, x_n$, we form the empirical cumulative distribution function, e.c.d.f., $F_n(x)$. Suppose we want to obtain $f(x) = F'(x)$. Then a crude estimate of f will be $F_n'(x)$. However, $F_n'(x)$ is not defined in the usual way but as a sum of generalized functions, Dirac delta functions. Now suppose that we know that f is in $\mathcal{C}^{(k)}$, or that it is even and convex. Then based on the properties of convolutions, a reasonable approach to estimate f could be to convolve $F_n'(x)$ with a kernel K that shares with f the same topological properties (even, convex, in $\mathcal{C}^{(k)}$). Call the convolution $\hat{f}_\lambda$

$$\begin{aligned}\hat{f}_\lambda(x) &= \textstyle\int K_\lambda(x-y)dF_n(y) \\ &= n^{-1}\textstyle\sum_{i=1}^n K_\lambda(x-x_i);\end{aligned} \tag{5.5}$$

$\hat{f}_\lambda(x) = \hat{f}_n$ is called the kernel density estimate of f based on F_n.

Moments of the np $\hat{f}_n$ density

Note that

$$\mathbb{E}\{x|\hat{f}_n\} = \int x\hat{f}_n(x)\,dx = \frac{1}{n\lambda}\sum_i \int xK_\lambda(x-x_i)\,dx$$

Hence,

$$\begin{aligned}\mathbb{E}\{x|\hat{f}_n\} &= \frac{1}{n}\sum\int(x_i+\lambda u)K(u)\,du \\ &= \bar{X} + \lambda\textstyle\int uK(u)\,du \\ &= 0,.\end{aligned}$$

since we are assuming that K is even. So, the nonparametric estimate of the density preserves the sample mean. Next, we look at higher order moments. First, because K is even, $\hat{f}_n$ preserves the moments of odd degree. Consider now the second order moment,

$$\begin{aligned}
\mathbb{E}\{x^2|\hat{f}_n\} &= \int x^2 \hat{f}_n(x)dx \\
&= \frac{1}{n\lambda}\sum\int x^2 K_\lambda(x - x_i)\,dx \\
&= \frac{1}{n}\sum\int (X_i + \lambda u)^2 K(u)du \\
&= \bar{X}^2 + \lambda^2 \int u^2 K(u).\,du
\end{aligned}$$

We see then that in order to preserve the sample variance, and, more important, in order to preserve its $\sqrt{n}$ order of convergence to the true variance, we must assume that

$$\int u^2 K(u)\,du = 0.$$

In general, we have the following theorem. A kernel of order p is a kernel with its first p moments equal to zero. It is not nonnegative if p is greater than two.

THEOREM 5.9 If K is a kernel of order p, then:

(a)

$$\mathbb{E}\{x^i|\hat{f}_n\} = \bar{X}^i; \quad 0 \le i \le p.$$

(b)

$$\mathbb{E}\{x^{(p+1)}|\hat{f}_n\} = \bar{X}^{(p+1)} + O(\lambda^{(p+1)}\int u^{(p+1)}K(u)du).$$

▽

REMARK 5.10 Note that we could have used α-trimmed versions of the estimated moments. Keep in mind also that the above is true only in the case of infinite range. If the range of definition of the density is finite, we will have asymptotically negligible end effects. Finally, note that p order kernels are being used to increase the order of convergence to zero of the bias of the nonparametric kernel.

Quantile function

It is also interesting to assess how close is the nonparametric quantile function associated with the estimate $\hat{f}_n$ to the empirical quantile function.

Let p be a value between 0 and 1. Then, $Q(p)$ is defined as the root of the equation

$$\int_{-\infty}^{Q(p)} \hat{f}_n(u)d\,u = p.$$

We assume that the support of k is included in a finite interval that for simplicity we assume to be equal to $[-1,1]$. But then

$$\int_{-\infty}^{Q(p)} K_\lambda(u - x_{(i)})d\,u = \begin{cases} 0, & \text{if } Q_{(p)} - x_{(i)} \geq \lambda, \\ 1, & \text{if } Q_{(p)} - x_{(i)} \leq -\lambda, \\ \int_{-\infty}^{\tau} K(u)d\,u & \text{otherwise,} \end{cases}$$

where $\tau = (Q(p) - x_{(i)})/\lambda$. Hence, if λ is sufficiently small,

$$Q(\frac{i}{n}) \approx x_{(i)} - \lambda.$$

This could be used as an ad hoc way of choosing the smoothing parameter: choose λ such that some prescribed set of the nonparametric density quantiles agree with the sample quantiles.

5.3.3 MSE and IMSE

If we want to assess the quality of the estimator $\hat{f}_\lambda$ of f we can do this pointwise or globally. First we will consider how good the estimator $\hat{f}_\lambda$ is at each point x. A reasonable measure is the MSE at a point x. As it is well known we can decompose the MSE as

$$\mathbb{E}\{\hat{f}_\lambda(x) - f(x)\}^2 = (bias(x))^2 + var(\hat{f}_\lambda(x)), \tag{5.6}$$

where

$$bias(x) = \hat{f}_\lambda(x) - \mathbb{E}\{\hat{f}_\lambda(x)\} \tag{5.7}$$

and

$$var(\hat{f}_\lambda(x)) = \mathbb{E}\{\hat{f}_\lambda - \mathbb{E}\{\hat{f}_\lambda(x)\}\}^2. \tag{5.8}$$

First, consider the *bias* term. By stationarity,

$$\mathbb{E}\hat{f}_\lambda(x) = \int K_\lambda(x-y)f(y)\,dy, \tag{5.9}$$

hence the *bias* is the pointwise error committed by the smoothing operation viewed from the point of view of deterministic approximation of functions.

The following results have been obtained by some researchers (see Prakasa Rao (1983) and references therein). However, most of these results have been obtained for identically distributed and independent observations. Hence, they have some interest for stochastic processes.

THEOREM **5.11** Suppose that f is a $\mathcal{C}^{(k)}$ density and K is a kernel. Then

$$\begin{aligned} bias(x) &= \sum_{j=1}^{k-1} f^{(j)}(x) \int_{\mathbb{R}} \frac{(y-x)^j}{j!} K_\lambda(y-x)\,dy \\ &\quad + \int_{\mathbb{R}} \int_x^y \frac{(z-x)^k}{k!} K_\lambda(x-y) f^{(k)}(z)\,dz\,dy. \end{aligned} \tag{5.10}$$

▽

Proof. Simply just use a k-term Taylor's expansion of f.

□

COROLLARY **5.12** Assume that f is a $\mathcal{C}^{(k)}(\mathbb{R})$ density and K a kernel such that first k moments except the first one are zero, i.e.

$$\int y^k K(y)dy = 0, \qquad j = 1, \dots, k, \tag{5.11}$$

then

$$bias(x) = \int_{\mathbb{R}} \int_x^y \frac{(z-x)^k}{k!} K(x-y) f^{(k)}(z)\,dz\,dy = O(\lambda^k). \tag{5.12}$$

▽

DEFINITION **5.13** *Kernels of order p.*

We say that a function K that satisfies conditions $C1$, and $C2$ is a kernel of order p if it satisfies (5.11) for $j = 1, \dots, p$, and

$$\int y^{(p+1)} K(y)dy \neq 0.$$

▽

REMARK **5.14** Kernel functions that satisfy the conditions of the corollary of $k \geq 2$ are necessarily negative in some regions. This might produce estimates of the unknown density that are negative at some points. This may seem a disadvantage if, for example, we want to display the estimated density. But if we use the density estimate within some procedure where the nonnegativeness of the density is not crucial, the additional reduction of the bias may be seen as an improvement.

Another result similar in spirit to (5.11) involves the modulus of continuity $\omega(h)$ of f. More precise results can be obtained by using elaborate definitions of the modulus of continuity involving the derivatives of f, but we just want to present an example of the kind of results available using the modulus of continuity.

THEOREM **5.15** Let f be a function with modulus of continuity ω. Let K be a kernel. Assume $\omega \in \mathbb{L}^1(k)$. Then

$$|bias(x)| = \int \omega(y - x)|K_\lambda(y - x)|\, dy. \tag{5.13}$$

▽

COROLLARY **5.16** Under the same conditions as in the theorem (5.15)

$$\begin{aligned} |bias(x)| &\leq \omega(h) + \textstyle\int_{B(h)} \omega(y - x) K_\lambda(y - x)\, dy \\ &\leq \omega(h) + \epsilon(h, \lambda), \end{aligned} \tag{5.14}$$

where $\epsilon(\lambda) \to 0$ as $\lambda \to 0$. ▽

COROLLARY **5.17** If f is locally a Lipschitz function of type α at x ($|f(y) - f(x)| \leq \beta |x-y|^\alpha$) then

$$|bias(x)| \leq \gamma \lambda^\alpha, \tag{5.15}$$

where

$$\gamma \leq \beta \int_{\mathbb{R}} |u|^\alpha |k(u)|\, du. \tag{5.16}$$

▽

We will consider now the variance of $\hat{f}_\lambda(x)$. The variance is given by

$$\begin{aligned} var(\hat{f}_\lambda) &= (n\lambda)^{-1} \int \bar{K}_\lambda^2(u) f(x-\lambda u)\, du + \\ &\quad +2/n \textstyle\sum_{i=1}^{n-1} (1-i/n) \int \bar{K}_\lambda(u) \bar{K}_\lambda(v) dF_{1,1+i}(x-\lambda u, x-\lambda v) \\ &= I + II. \end{aligned} \tag{5.17}$$

where $\bar{K}$ is the centered kernel

$$\bar{K}_\lambda(u) = K(u) - \mathbb{E} K_\lambda(x - x_i), \tag{5.18}$$

and $F_{1,i+1}$ is the joint probability distribution of x_1 and x_{i+1}.

REMARK **5.18** 1. The second term of the variance does not involve terms in λ^{-1}. This implies that if this cross covariance term II in (5.17), disappears the variance will behave as in the case of independent observations. This will imply that the same rate of decay in both cases.

2. The cross covariance term II in (5.17) will tend to zero as $n \to \infty$ if and only if $\int \bar{K}_\lambda(x-u) \bar{K}_\lambda(x-v) dF_{1,i+1}(u,v) \to 0$.

3. This result was hinted at by Rosenblatt (1970), and again for some special cases by Yakowitz (1985, 1987) and Robinson (1983, 1986).

4. Call $a_i(x, \lambda) = \int \tilde{K}_\lambda(u)\tilde{K}_\lambda(v) dF_{1,i+1}(u, v)$. Because of 1. above we are interested in process such that

$$\sum_{i=1}^{n-1}(1 - i/n)a_i(x, \lambda), \qquad (5.19)$$

is summable. Then the cross covariance term II is of order $O(n^{-1})$.

We have then the following result

THEOREM 5.19 Suppose $\{x_t\}$ is a stochastic process with stationary marginal density $f(x)$. Assume that the process satisfies some asymptotic independence assumption which implies that

$$\int_{R^2} \tilde{K}_\lambda(u)\tilde{K}_\lambda(v)\, dF_{1,i+1}(x - \lambda u, x - \lambda v) \to 0 \quad \text{as} \quad i \to \infty. \qquad (5.20)$$

Also, assume that f is such that $bias(x) = O(\lambda^k)$. Then,

$$MSE(\hat{f}_\lambda(x)) = O(\lambda^k + \frac{1}{n\lambda}), \qquad (5.21)$$

and the optimal smoothing parameter $\hat{\lambda}$ is as in the case of independent observations of order

$$\hat{\lambda} = O(n^{-1/(2k+1)}). \qquad (5.22)$$

Using $\hat{\lambda}$,

$$MSE(\hat{f}_\lambda(x)) = O(n^{-2k/(2k+1)}). \qquad (5.23)$$

▽

REMARK 5.20 1. Note that the variance term I is of order

$$\frac{f(x)}{n\lambda}\int K_\lambda^2(u)\, du. \qquad (5.24)$$

Hence, we may want to use a kernel K such that

$$\begin{aligned} &\int K^2(u)\,du = \text{minimum} \\ &\int K(u)\,du = 1, \quad K \geq 0. \end{aligned} \tag{5.25}$$

These kernels were studied by Epanechnikov (1969), and are given by

$$K_e(u) = \begin{cases} \dfrac{3}{4\sqrt{5}} - \dfrac{3x^2}{20\sqrt{5}} & \text{for } |x| \leq \sqrt{5}, \\ 0 & \text{otherwise.} \end{cases}$$

However, the ratio between $\int K_e^2$ and $\int K^2$ for many other kernels is so close to one that the other factors in the MSE will eliminate the small reduction in the MSE obtained by using K_e.

2. On many occasions, we are interested in the overall performance of $\hat{f}_\lambda$. This is measured by the MISE. Where the MISE is the defined as the integral of MSE(x) over the region of interest. However, the above results about the MSE generalize to the MISE. In particular, the order of the optimal smoothing factor $\hat{\lambda}$, the rate of convergence and the Epanechnikov kernels are the same.

Optimal kernels of order p

We have seen above that the use of kernels of order p is justified as a means of reducing the bias of the estimate in order that the first p moments of the nonparametric density estimate are identical to the sample moments. It is of some interest to obtain analogues of Epanechnikov's kernels for the case of a p order kernel. We use the classical variational method to obtain an expression for these optimal kernels.

The problem is

$$\int K^2(u)\,du = \text{minimum}$$

subject to

$$\int K(u)\,du = 1,$$

$$\int K(u)u^i\,du = 0, \qquad i = 1, \ldots, p$$

$$K \geq 0.$$

Form the Lagrangian of this variational problem:

$$\text{minimize } \mathcal{L}(K) = \int K^2(u)du + \lambda_0(\int K(u)du - 1)$$
$$+\lambda_1(\int K(u)udu) + \cdots + \lambda_p(\int K(u)u^p du)$$

Then, if δK is an increment around the optimal kernel such $K + \delta K$ is a p order kernel,

$$\mathcal{L}(K + \delta K) - \mathcal{L}(K) = \int (2K(u) + \lambda_0 u + \cdots + \lambda_p u^p)\delta K = 0,$$

for any increment δK. Thus,

$$(2K(u) + \lambda_0 u + \cdots + \lambda_p u^p) = 0.$$

We require K to be an even function. This implies that $\lambda_{(2i+1)} = 0$. Hence,

$$2K(u) = \lambda_0 + \lambda_2 u^2 + \cdots + \lambda_u^2 k,$$

where $2\mu = p$. Now, if v is such that $K(v) = 0$ set $K(u)$ equal to zero for $|u| \geq v$, then K must satisfy the constraints. Hence

$$\lambda_0 \frac{u^{(2i+1)}}{(2i+1)} + \lambda_2 \frac{u^{(2i+3)}}{2i+3} + \cdots + \lambda_{(2\mu)} \frac{u^{(2i+2\mu+1)}}{2K+1} = 0, \tag{5.26}$$

for $i = 0, \ldots, k$. Also, it must satisfy the equation

$$K(v) = \lambda_0 + \lambda_2 u^2 + \cdots + \lambda_{(2k)} = 0, \tag{5.27}$$

plus the equation

$$\int K(u) du = 1$$

$$\lambda_0 u + \lambda_2 u^3 + \cdots + \lambda_{(2\mu)} u^{(2\mu+1)}. \tag{5.28}$$

These nonlinear equations (5.26-5.28) may be solved by iteration.

5.3.4 Choosing the smoothing constant

In practice, we need some method to decide which smoothing constant λ we should use. There are several approaches to this question. One method is to compute the optimal smoothing factors for special cases, such as when the density is normal. In this case $\hat{\lambda} = 1.06\sigma$. This value should work reasonably well for densities close to the normal. For case of densities with tails heavier than normal, we could use $\hat{\lambda} = 0.79\sigma$. These values appear in Silverman (1986). Another approach is advocated by Rudemo (1982) and Bowman (1985). It consists of minimizing an estimate of MISE. Also, one could use the test graphs advocated by Silverman (1986) based on the second derivative of the density. Finally, a quick method of obtaining a reasonable smoothing constant is to choose the constant that gives a smooth density but also agrees with a preselected sample set of quantiles such as the median and the quartiles, etc.. In order to be consistent, this set of preselected quantiles must increase as the sample size increases to include the whole line.

We will study the Rudemo-Bowman procedure for nonindependent observations (Rudemo, 1982; Bowman 1984). Consider

$$\int_{\mathbb{R}} (f(x) - \hat{f}(x))^2 dx = \int_{\mathbb{R}} f^2 dx - 2\int_{\mathbb{R}} f(x)\hat{f}(x) dx + \int_{\mathbb{R}} \hat{f}^2(x) dx. \quad (5.29)$$

The expected value of this equation, the IMSE, is

$$\mathbb{E}\int_{\mathbb{R}} (f(x) - \hat{f}(x))^2 dx = \int_{\mathbb{R}} f^2 dx - 2\int_{\mathbb{R}} f(x)\hat{f}(x) dx + \int_{\mathbb{R}} \hat{f}^2(x) dx. \quad (5.30)$$

As the first term on the right hand side of (5.30) does not depend on the smoothing constant we could drop it. The third term does not contain any unknowns. Hence, it may be estimated directly by, say, numerical quadrature. The second summand contains the unknown density. However, note that it is equal to the expected value of $\mathbb{E}\hat{f}_n(X)$, where X is a random variable with density f. The standard way of estimating expectations is by the sample mean of the observations. For this case, these correspond to $\hat{f}_n(x_i)$. However it is better to use $\hat{f}_{(i)}(x_i)$ where $\hat{f}_{(i)}$ is the kernel density estimate using the same parameter λ as $\hat{f}_n$ but without the observation x_i. We compute the expectation

$$\mathbb{E}\frac{1}{n}\sum_{i} \hat{f}_{(i)}(x_i) = \frac{1}{n(n-1)}\sum_{i \neq j} \mathbb{E}K_{\lambda}(x_i - x_j) \quad ,$$

$$\boldsymbol{E}\frac{1}{n}\sum_{i}\hat{f_{(i)}}(x_i) = \frac{1}{(n-1)}\sum_{\nu=0}^{n-1}\int K_{\lambda}(u,v)f_{\nu}(u,v)\,d\,u\,d\,v.$$

Hence, if

$$\lim_{\nu\to\infty}\int K_{\lambda}(u,v)f_{\nu}(u,v)\,d\,u\,d\,v = \lim_{\nu\to\infty}\int K_{\lambda}(u,v)f(u)f(v)\,d\,u\,d\,v, \tag{5.31}$$

we could use

$$-2\frac{1}{n}\sum_{i}\hat{f}_{(i)}(x_i) + \int_{\boldsymbol{R}}\hat{f}^{2}(x)\,d\,x \tag{5.32}$$

as an estimate of

$$-2\int_{\boldsymbol{R}}f(x)\hat{f}(x)\,d\,x + \int_{\boldsymbol{R}}\hat{f}^{2}(x)\,d\,x,$$

and minimize (5.32) to obtain an acceptable value of the smoothing constant λ. In order for (5.31) to be true it is sufficient that

$$|f_{\nu}(x,x) - f^{2}(x)| \to 0 \quad \text{as } \nu\to\infty, \tag{5.33}$$

for all x except in a set of measure zero. We have proven the following theorem

THEOREM **5.21** If X_t is an strictly stationary process, then if (5.33) is satisfied for all x except in a set of measure zero, then

$$-2\frac{1}{n}\sum_{i}\hat{f}_{(i)}(x_i) + \int_{\boldsymbol{R}}\hat{f}^{2}(x)\,d\,x$$

$$-2\int_{\boldsymbol{R}}f(x)\hat{f}(x)\,d\,x + \int_{\boldsymbol{R}}\hat{f}^{2}(x)\,d\,x \to 0,$$

as $n\to\infty$. ▽

5.3.5 A global weak convergence result

As in Bickel and Rosenblatt (1973), we are interested in the process

$$Y_n(x) = \sqrt{n\lambda}\frac{\hat{f}_n(x) - f(x)}{\sqrt{f(x)}}$$

which may be written as

$$Y_n(x) = \frac{1}{\sqrt{\lambda f(x)}}\int_{\mathbb{R}} K(\frac{x-u}{\lambda})\, dE_n(F(u)),$$

where $E_n = \sqrt{n}(F_n(F^{-1}(x) - x)$, the empirical cumulative process for the $F(x_i)$ observations. To study the weak convergence of Y_n, we will follow Bickel and Rosenblatt (1973) closely, even using generally the same notation. The idea is to approximate Y_n by a series of processes, in such a way that the distribution theory of the last approximation may be obtained. Define

$$Y_{0n}(x) = \frac{1}{\sqrt{\lambda_n f(x)}}\int_{\mathbb{R}} K(\frac{x-u}{\lambda_n})\, dB_n(F(u)),$$

where B_n is an standard Brownian bridge. We will make now the crucial assumption. Assume that

$$\|E_n - B_n\|_\infty = o_p(a(n)), \tag{5.34}$$

a.s., where $\| \|_\infty$ is the sup norm, and $a(n) \to 0$ as $n \to \infty$. Then

THEOREM **5.22** (a)

$$\|Y_n - Y_{0n}\|_\infty = O_p(\frac{a(n)}{\lambda_n}). \tag{5.35}$$

(b) If f is continuous positive and bounded then

$$\|Y_{0n} - Y_{1n}\|_\infty = O_p(\sqrt{\lambda_n}), \tag{5.36}$$

where

$$Y_{1n}(x) = \frac{1}{\sqrt{\lambda_n f(x)}} \int_{\boldsymbol{R}} K(\frac{x-u}{\lambda_n})\, dW_n(F(u)),$$

W_n is a standard Wiener process in the interval $[0,1]$.

(c) If f satisfies (b), $2\sqrt{f}$ is absolutely continuous, and if its derivative $f'(x)/f(x)$ is bounded then

$$\|Y_{2n} - Y_{3n}\|_\infty = O_p(\sqrt{\lambda_n}),$$

where

$$Y_{2n}(x) = \frac{1}{\sqrt{\lambda_n f(x)}} \int_{\boldsymbol{R}} K(\frac{x-u}{\lambda_n})\sqrt{f(u)}\, dW_n(u),$$

and

$$Y_{3n}(x) = \frac{1}{\sqrt{\lambda_n}} \int_{\boldsymbol{R}} K(\frac{x-u}{\lambda_n})\, dW_n(u). \tag{5.37}$$

(d) The finite dimensional distributions of Y_{1n} and Y_{2n} are identical. Hence we could identify both processes with each other.

(e) Combining (a) to (d) we conclude that

$$\|Y_n - Y_{3n}\|_\infty = O_p(\sqrt{\lambda_n} + a(n)/\sqrt{\lambda_n}). \tag{5.38}$$

▽

Proof The proof is as in Bickel and Rosenblatt (1973) using (5.34) instead of the rate used by them. Because of (5.38) we should assume that $\lambda_n \to 0$ as $n \to \infty$ and that

$$a(n) = o_p(\lambda_n).$$

□

REMARK 5.23 There is available a wealth of results concerning assumption (5.34). This is an example of the so called strong approximation theorems. For a review of invariance principles for weakly dependent sequences see Philipp, (1986). The best results for strong mixing and absolutely regular sequences are in Dhompongsa (1984). We describe now these results.

The most complete answer about the function $a(n)$ is available when the observations come from independent and identically distributed random variables. For this case strong approximation results began with Brillinger (1969) and Breiman (1968) and ended essentially with Komlos et. al. who obtained the best rate for $a(n)$ as

$$a(n) = \frac{\log(n)}{n}.$$

DEFINITION 5.24 *Strong mixing sequences* Consider a sequence ψ_i of random variables. Call $\mathcal{M}_i^j$ the σ-field generated by $\{\psi_i, \dots, \psi_j\}$, $i \geq j$. The sequence ψ_i is said to be strongly mixing (Rosenblatt, 1956) if the strong mixing coefficient

$$\alpha_n := \sup |\mathbb{P}(A| \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

where the sup is taken over

$$\{A \in \mathcal{M}_1^k, B \in \mathcal{M}_{k+n}^{\infty}, k \geq 1\}$$

goes to zero as $n \to \infty$. ▽

DEFINITION 5.25 *Absolutely regular sequences.*A sequence of $\{\psi_i\}$ of random variables is said to be absolutely regular if

$$\beta_n = \mathbb{E} \sup |\mathbb{P}(B|\mathcal{M}_1^k - \mathbb{P}(B)| \to 0, \quad \text{as } n \to \infty,$$

where $B \in \mathcal{M}_{k+n}^{\infty}$, for $k \geq 1$. . ▽

DEFINITION 5.26 *ϕ-mixing sequences.* A sequence of $\{\psi_i\}$ of random variables is said to be ϕ-mixing if

$$\phi_n = \frac{\sup |\mathbb{P}(B \cap A) - \mathbb{P}(B)\mathbb{P}(A)|}{\mathbb{P}(A)} \to 0,$$

where the $A \in \mathcal{M}_1^k$, and $B \in \mathcal{M}_{k+n}^{\infty}$, for $k \geq 1$. ▽

There are other types of mixing coefficients. For example, the maximum correlation coefficient between the two σ-fields $\mathcal{M}_1^n$ and $\mathcal{M}_{n+1}^{\infty}$, i.e. ρ-mixing, also there is ψ-mixing, and ψ^*-mixing. Finally another type of of mixing coefficient is the one given in (5.33). Important relationships between these mixing coefficients are: (a) α-mixing is implied by the other types of mixing. That is, it is the weakest mixing coefficient between those that we have defined. (b) Also, β-mixing is stronger ψ-mixing. (c) Finally, ρ-mixing and β-mixing are not related in a simple way. When the sequence of random variables is Gaussian, every mixing condition is equivalent to an autocorrelation function that is asymptotically equal to zero. Most of the weak convergence results have been proven for the case of ϕ-mixing coefficients. However, an AR(1) process is not necessarily ϕ-mixing. Hence we prefer to use other types of mixing coefficients. In general it is believed that absolute regularity is the most appropriate mixing coefficient when we are working with time series. Under very weak conditions, AR processes are β-mixing. Moreover there is a much more general result due to Thuam and Tran (1985, 1986) that gives conditions that insure that bilinear time series are absolutely regular with an exponential rate of convergence. The theorems in Dhompongsa (1984) are as follows.

THEOREM **5.27** *Absolutely regular.* Suppose that $\{\psi_i\}$ is a sequence of strictly stationary p-dimensional random vectors with common continuous distribution F such that the absolute mixing coefficient $\beta(n)$ is dominated by $n^{-\tau}$, $\beta_n << n^{-\tau}$; where $\tau > 3 + p$. Then, there is a sequence of independent standard Brownian bridges, B_n such that

$$\sup_{0\leq s\leq 1} |\sum_j(\{F(\psi_j) \leq s\} - s) - \sum_j B_j(s)| << n^{5-\lambda} \qquad (5.39)$$

For some $\lambda > 0$. ▽

THEOREM 5.28 *Strongly mixing* .Suppose that $\{\psi_i\}$ is a sequence of strictly stationary p-dimensional random vectors with a common distribution continuous distribution F such that the strong mixing coefficient $\alpha(n)$ is dominated by $n^{-\tau}$, $\alpha_n << n^{-\tau}$; where $\tau > 3 + p$. Then there is a sequence of independent standard Brownian bridges, B_n,, such that

$$\sup_{0\leq s\leq 1} \sum_{j}(\{F(\psi_j) \leq s\} - s) - \sum_{j} B_j(s)| << n^{5}(\log\log(n))^{-\lambda} \tag{5.40}$$

for some $\lambda > 0$. We have assumed the F is continuous. Continuity may be dropped if we assume: the coefficient $\alpha(n)$ is dominated by $n^{-\tau}$, $\alpha_n << n^{-\tau}$; where $\tau > 4 + 2p$. The conclusion is the same as in (5.40). ▽

We use theorems 5.27, 5.22 and 5.28, 5.22 to obtain the following result.

THEOREM 5.29 (a) If $\{x_i\}$ is a sequence of strictly stationary p-dimensional random vectors with common distribution F such that the absolute mixing coefficient $\beta(n)$ is dominated by $n^{-\tau}$, $\tau > 3 + p$, then

$$\|Y_n - Y_{3n}\|_\infty = O_p(\sqrt{\lambda_n} + n^{-\gamma}/\sqrt{\lambda_n}),$$

some $\gamma > 0$; where Y_{3n} is defined in (5.37).

(b) Suppose that $\{x_i\}$ is a sequence of strictly stationary p-dimensional random vectors with a common continuous distribution F such that the strong mixing coefficient $\alpha(n)$ is dominated by $n^{-\tau}$, $\alpha_n << n^{-\tau}$; where $\tau > 3 + p$. Then

$$\|Y_n - Y_{3n}\|_\infty = O_p(\sqrt{\lambda_n} + (\log\log(n))^{-\gamma}/\sqrt{\lambda_n}),$$

for some $\gamma > 0$.

∇

Proof The theorem follows because under the assumptions stated condition (5.34) is true and this implies the conclusion of Theorem 5.22. Condition (5.34) is true because for part (a) the hypotheses of Theorem 5.27 are valid, and for part (b) the conditions of Theorem 5.28 are satisfied.

□

REMARK **5.30** Several researches have proven a similar theorem. See Silverman (1978) for the case of independent and identically distributed random variables and also Collomb (1984b), Doukham and Ghides (1983) for the case of ϕ-mixing random variables. However their results are not as sharp as the theorem just stated.

5.4 Estimation of state-dependent models

Priestly (1980), introduced the following class of models,

$$y_t = \sum_{k=1}^{p} \alpha_k(S_t) y_{t-k} + \epsilon_t$$

where, ϵ_t is a white noise process, and S_t is the 'state' of the process at time t. These models include a large number of nonlinear models, such as threshold models, bilinear models, nonlinear exponential models, and so on. We will assume that the state at time t just depends on p past observations, and, possibly, some covariates.

To estimate the function $\alpha = (\alpha_1, \ldots, \alpha_k)$ we propose the use of the following nonparametric method. Solve for each value of the state, S, the following minimization problem

$$\min_{\alpha(S)} \sum_t w(S; S_t)(y_t - \sum_{k=1}^{p} \alpha_k(S) y_{t-k})^2,$$

where we assume that the weights are, for each S, a probability function. Also, the weights are such they are zero outside a ball, and that as the smoothing parameter $\lambda \to 0$, this ball shrinks to zero also. Thus, the solution for each S is

$$\alpha(S) = (X^t W X)^{-1} X^t W Y,$$

where $Y = (y_n, \ldots, y_{p+1})$, $X = (y_{t-k})_{t=p+1, \ldots, n}^{k=1, \ldots, p}$ and

$$W = \text{diag}(w(S; S_t)), t = p+1, \ldots, n.$$

Note that these equations are a generalization of the Yule-Walker equations for autoregressive processes. Now, suppose the true function is denoted by α^0, then

$$\begin{aligned} \alpha(S) = & \ (X^t W X)^{-1} X^t W + ((\alpha(S_t))X + \epsilon \\ = & \ (X^t W X)^{-1} X^t W \epsilon + \alpha^0(S) + \\ & \ (X^t W X)^{-1} X^t W((\alpha(S_t) - \alpha(S))S_t)) \end{aligned}$$

Thus,

$$\alpha(S) - \alpha(S) = (X^t W X)^{-1} X^t W \epsilon + (X^t W X)^{-1} X^t W((\alpha^t(S) - \alpha^t(S_t) S_t) \qquad (5.41)$$

Now, $W((\alpha^t(S) - \alpha^t(S_t))S_t) = 0$, outside a ball that shrinks to zero as $\lambda \to 0$. We will assume that α^0 is uniformly continuous over the range of values that we are interested in estimate. This implies in turn that, $W((\alpha^t(S) - \alpha^t(S_t))S_t) \to 0$, uniformly.

We consider now the term (X^tWX). Because W vanishes outside a shrinking ball, as $\lambda \to 0$, we find that

$$(X^tWX) \to (S, S_t).$$

Then, we may ignore asymptotically the second term in (5.41). Now, if we consider the second term we see that is a sum of martingale differences. But, the quadratic variation of this sum is given by

$$(X^tWX)^{-1}X^tW^2X(X^tWX)^{-1}.$$

We have to consider the central term in this expression. Now, w_t is a triangular array of ergodic random variables, hence a strong law of the large numbers applies. Hence the quadratic variation vanishes asymptotically.

Next, note that the term

$$(X^tWX)^{-1}X^tW\epsilon$$

is a sum of martingale differences. We apply a central limit theorem for martingales to conclude that we proposed estimate is consistent and asymptotically Gaussian.

5.5 Combination of nonparametric and parametric estimates

Suppose that the unknown density, f, lies at a distance d from a set of densities that may be given coordinates in $\Theta \subset \mathbb{R}^p$, Θ compact, for simplicity. The 'distance' in this case is given by

$$d(f,\Theta) = \min_{\theta\in\Theta} \int \log(f_\theta(x)d\,x$$

Because, Θ is compact, $d(f,\Theta)$ is finite, and the minimum value is attained by some θ in Θ.

Call, θ_0, to the value of θ that minimizes $d(f,\Theta)$. Suppose that we have a set of n observations, $x_1,\ldots,x_n$. We may try to estimate θ_0 using maximum likelihood, or equivalently we may want to maximize the log-likelihood function, $L(\theta)$. Call $\hat{\theta}$ to the value of θ in Θ that maximizes $L(\theta)$.

Now, if we fix $\theta \in \Theta$, by the Strong Law of the Large Numbers

$$L(\theta)/n \to \mathbb{E}_f\{\log(f_\theta(X))\},$$

as $n \to \infty$, a.s., if the expectation is finite.

Then, if we assume conditions on f and f_θ such that this convergence is uniform, we may conclude that the m.l. estimate of θ_0 is consistent, assuming that the minimum of $d(f,\Theta)$ is unique. Examples of these conditions are entropy bounds in the style of Vapnik and Červonenkis, (1971,1981). Hence, even if f_{θ_0} is not the true density, θ_0 may be estimated consistently.

Now, we consider the asymptotic distribution of $\hat{\theta}$. As with the standard maximum likelihood method we find that

$$\sqrt(n)(\hat{\theta} - \theta_0) = \sqrt(n)(\sum \frac{\partial}{\partial\theta}(\frac{f'_{\theta_0}}{f_{\theta_0}}))^{-1}(\sum(\frac{f'_{\theta_0}}{f_{\theta_0}})) + o_p(\sqrt(n)).$$

Now, by the Central Limit Theorem

$$\frac{1}{\sqrt{n}}\sum(\frac{f'_{\theta_0}}{f_{\theta_0}})$$

is asymptotically normal, assuming that its variance is finite. Assuming that we may interchange the order of differentiation and integration in

$$\boldsymbol{E}_f \frac{f'_{\theta_0}}{f_{\theta_0}}$$

we find that this expectation is zero. For the variance term we find that

$$(\frac{f'_{\theta_0}}{f_{\theta_0}})^2 = (\frac{\partial}{\partial\theta}(\frac{f'_{\theta_0}}{f_{\theta_0}})) + \frac{f''_{\theta_0}}{f_{\theta_0}}$$

Now, the expectation of the second term does not vanish, as is the case for standard maximum likelihood. This is the main difference when we assume that f does not lie within $\{f_\theta\}$. As an example of above suppose that f_θ is a normal density with mean μ and variance σ^2. Then what we just have said, in this case translates into the Strong Law of the Large Numbers and the Central Limit Theorem. Notice that all of the above is valid for a stochastic process.

Now, suppose we want to combine this parametric estimate, $f_{\hat{\theta}}$ with a nonparametric density estimate, $\hat{f}_{np}$. The most obvious way is by a convex combination of $f_{\hat{\theta}}$ and $\hat{f}_{np}$

$$\hat{f}_{sp} = \alpha f_{\hat{\theta}} + (1-\alpha)\hat{f}_{np}$$

In order to obtain a consistent estimate we let $\alpha \to 0$, as $n \to \infty$. However, we want to find how α should be chosen optimally. Write

$$\hat{f}_{sp} - f = \alpha(f_{\hat{\theta}} - f_{\theta_0}) + (1 - \alpha)(\hat{f}_{np} - f) + \alpha(f_{\theta_0} - f)$$

Then, using the $\mathbb{L}_1$ distance between densities we find that

$$\mathbb{E}\|\hat{f}_{sp} - f\| \leq \alpha\mathbb{E}\|f_{\hat{\theta}} - f_{\theta_0}\| + (1 - \alpha)\mathbb{E}\|(\hat{f}_{np} - f)\| + \alpha\mathbb{E}\|f_{\theta_0} - f\|,$$

if we minimize the right hand side of this equation, we obtain an upper bound on $\mathbb{E}\|\hat{f}_{sp} - f\|$. The value of α that minimizes this expression is

$$\begin{array}{ll} \hat{\alpha} = \theta & \text{if } E\|\hat{f} - f_{\theta_0}\| > E\|\hat{f}_{np} - f\| + \mathbb{E}\|f_{\theta_0} - f_{\hat{\theta}}\| \\ \hat{\alpha} = 1 & \hfill \text{otherwise.} \end{array}$$

Hence, based on this upper bound we should switch estimates when the distance of the estimate to the true function is smaller than the combined errors of the other density. Also, this result implies that the convex combination is at least as good as the other two densities.

We can also given lower bounds for the $\mathbb{L}_1$ error committed using the convex estimate. Using the other side of the triangle inequality we have that

$$|\Delta_i - \Delta_j + \Delta_k| < \mathbb{E}\|\hat{f}_{sp} - f\|$$

, where the Δ's are each of the errors in the decomposition of $\mathbb{E}\|\hat{f}_{sp} - f\|$. Based on these results, we see that again, to minimize the $\mathbb{L}_1$ norm we must switch between the parametric estimate and the nonparametric density estimate.

5.6 Bootstrapping a np function estimate

If we want to make inferences concerning the whole unknown function, such as confidence bands, we might use the asymptotic behaviour of the estimate to obtain an approximation of the distributional behaviour of the estimate. However, because we are concerned with a function and not with a point estimate we may view the asymptotics of nonparametric function estimates with some suspicion. The sample size need to obtain reliable estimates may lie outside the sample size at our disposition.

The bootstrap is useful to investigate situations where the sample size is not large. We want to consider the use of bootstrapping techniques to investigate the finite sample probability distribution of nonparametric function estimates. Suppose that we want to obtaining confidence bands for the unknown function. We will assume that the function that we want to estimate is a probability density. Also, we will assume that we are working with a univariate density. We want to obtain confidence bands over the interval $I = [a, b]$. To obtain these bands, using the asymptotic theory we consider

$$I_n(f) \doteq \frac{\sqrt(n)(\hat{f}_n - f)}{\sqrt{f}}.$$

To obtain a confidence band we obtain the sup-norm over the interval I of

$I_n(f)$. The bootstrapped estimate of $I_n(f)$, $I_n^*(f)$, is defined as

$$I_n^*(f) = \frac{\sqrt{n}(\hat{f}_n^* - \hat{f}_n)}{\sqrt{\hat{f}_n}},$$

where

$$\hat{f}_n^* = \int K_\lambda(x-u)dF_n^*(u),$$

and $F_n^*(u)$ is the e.c.d.f. of a random sample of size m from F_n.

Hence,

$$\frac{\sqrt{n}(\hat{f}_n^*(x) - \hat{f}(x))}{\sqrt{\hat{f}_n}} = \frac{\sqrt{n}}{\sqrt{\hat{f}_n}} \int K_\lambda(x-u)d(F_n^* - F_n).$$

We assume that K is a bounded variation kernel, then

$$\frac{\sqrt{n}}{\sqrt{\hat{f}_n}} \int K_\lambda(x-u)d(F_n^* - F_n) = \frac{\sqrt{n}}{\sqrt{\hat{f}_n}} \int (F_n^* - F_n)dK_\lambda(x-u),$$

a similar expression is also true for $I_n(f)$.

Now, we use Komlos et al. (1974,1975) strong approximation for the empirical process corresponding to independent and identically distributed univariate random variables. There is a probability space such that, almost surely,

$$\frac{\sqrt{m}}{\sqrt{\hat{f}_n}}(F_n^* - F_n) = \frac{B \circ F_n}{\sqrt{\hat{f}_n}} + O(\frac{\log m}{m}),$$

and

$$\frac{\sqrt{n}}{\sqrt{\hat{f}}}(F_n - F) = \frac{B \circ F}{\sqrt{\hat{f}}} + O(\frac{\log n}{n}).$$

Hence,

$$\frac{\sqrt{m}}{\sqrt{\hat{f}_n}}(f_n^* - f_n) = \frac{1}{\sqrt{\hat{f}_n}} \int B \circ F_n dK_\lambda(x-u) + O(\frac{\log(m)}{m}),$$

$$\frac{\sqrt{m}}{\hat{f}_n}(f_n^* - f_n) = \frac{1}{\sqrt{\hat{f}}}\int B \circ F dK_\lambda(x-u) + \frac{1}{\sqrt{f}}\int B \circ (F_n - F) dK_\lambda(x-u) +$$
$$+(\frac{1}{\sqrt{\hat{f}_n} - \sqrt{f}})\int B \circ (F_n - F) dK_\lambda(x-u) + O(\frac{\log(m)}{m})$$

Now,

$$\frac{\|B \circ (F_n - F)\|}{\sqrt{f}} = O(\omega(\|F_n - F\|)),$$

where ω is the modulus of continuity of the Brownian motion. Hence, because $\|F_n - F\| \to 0$, we have even an exponential bound for this, $\frac{\|B \circ (F_n - F)}{\sqrt{f}}\| \to 0$, as $m \to \infty$.

Also,

$$\|\frac{1}{\sqrt{f}} - \frac{1}{\sqrt{\hat{f}_n}}\| = O(\|f - \hat{f}_n\|).$$

Hence, we conclude that the distribution of $I_n(f)$, and $I_n^*(f)$, coincide as $n, m \to \infty$. Therefore, we may use this result to approximate the distribution of $I_n(f)$, by the distribution as the bootstrapped quantity.

5.7 Summary

If we do not know if the model belongs to a particular parametric family, an approach for obtaining a density or for fitting a regression type model is by nonparametric function estimation. The literature on nonparametric function estimation is enormous. However, the literature in nonparametric function fitting in time series analysis is quite scarce. In this chapter, we have generalized several nonparametric procedures to a time series context.

The first problems that we were interested in were the estimation of the density function and its derivatives of time dependent observations. To make any progress, we needed to assume that the observations were asymptotically independent as the time span among them increases. There are several definitions on just how this asymptotic independence is achieved. Some of this conditions, such as ϕ-mixing are not satisfied for autoregressive models. Hence they seem of little use in a time series context. However, most of the results about the asymptotic behaviour of the nonparametric estimate of the density use a ϕ-mixing condition. We have given several results using mixing conditions, such as absolutely regular mixing and uniform strong mixing that are satisfied by ARMA models. The best result about the asymptotic behaviour of the density estimate when the observations are independent is by Silverman (1978) who gives the uniform rate of convergence of the density estimate to a particular Gaussian process. We have attempted to generalized this result for time dependent situations using Dhompongsa's (1984) strong approximation result about the e.c.d.f. of a time dependent process.

Apart from the above result we have studied the asymptotic behaviour of the nonparametric density estimate by examining its bias and variance. A most important problem is the choosing of the smoothing constant of the nonparametric estimate. Rudemo's (1981) approach is the most accepted in the literature. We have generalized his results to the case of a time dependent stochastic process. Also, an alternative procedure for choosing the smoothing constant may be to choose the smoothing constant such that it preserves the sample quantiles. Also, it is of interest to generalize Epanechnivok's (1964) optimal kernels for density estimation to optimal kernels of order p for density estimation. We have derived some necessary conditions that these kernels must satisfy.

Turning now to the nonparametric function fitting when the observations are a time dependent process, we have given nonparametric function fitting procedures based on the minimization of a loss function with random weights. We have presented a result similar to the one about the asymptotic behaviour of the nonparametric density for the case of - nonparametric function fitting. Also, we are interested in a time series context in generalizing the autoregressive equations to allow a nonlinear autoregressive model. A simplification may be achieved by thinking of the parameters of the time series process as state-dependent, and not constant. The estimates of these autoregressive coefficients are obtained by procedures similar to ordinary least squares, but with state-dependent weights.

Another interesting situation is when we assume that the governing equation of the time dependent stochastic process is close to a parametric family of models. We have attempted to blend the parametric estimate with the nonparametric one. There are several forms for doing this. We thought the most sensible was a convex combination of both estimates. Although this estimate will inherit the slower convergence of the nonparametric estimate, the combination of both estimates is reasonable from a nonasymptotic point of view. But first for this to be true, it is necessary that the estimated parametric estimate be consistent. We have given some theorems proving consistency for some particular situations.

References

1. Benedetti, J. K. (1977). On the nonparametric estimation of regression functions. *J. Royal Soc., Ser. B* **39**, 248–253.

2. Bickel, P. J., and Rosenblatt, M. On some global measures of the deviations of density function estimates. *Annals of Mathematical Statistics,* , **1**, 1071–1095.

3. Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika,* **71**, 353–361.

4. Breiman, L. (1968). *Probability.* Addison-Wesley, Reading.

5. Brillinger, D. (1969). An asymptotic representation of the sample distribution function. *Bull. Amer. Math. Soc.*, **75**, 545–547.

6. Collomb, G. (1981). Estimation nonparamétrique de la regression: revue bibliographique. *Int. Statist. Rev.*, **49**, 75–93.

7. Collomb, G. (1984a). Nonparametric time series analysis and prediction: Uniform almost sure convergence of the window and the k-NN autoregression estimates. *Math. Oper. Stat., Ser. Statistics.*

8. Collomb, G. (1984b). Properties de convergence presque complete du predicteur a noyan. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **66**, 441-460.

9. Devroye, L. (1978). The uniform convergence of the nearest neighbor regression function estimators and their application in optimization *IEEE Trans. Inform. Theory*, **24**, 142–151.

10. Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Mathematical Statistics*, **9**, 1310-1319.

11. Devyore, L. and Györfi, L. (1985). *Nonparametric Density Estimation: the L_1 View*. John Wiley and Sons, New York.

12. Devroye, L. and Wagner, T. J. (1980a). On the L1 convergence of kernel regression function estimators with applications in discrimination. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **51**, 15-25.

13. Devroye, L. and Wagner, T. J. (1980b). Distribution-free consistency results in nonparametric discrimination and regression function estimators. *Annals of Mathematical Statistics*, **8**, 231–239.

14. Dhompongsa, S. (1984). A note on the almost sure approximation of empirical processes of weakly dependent random vectors. *Yokohama Math. J.*, **32**, 113-121.

15. Doukham, P. and Ghindes, M. (1983). Estimation de la transition de la probabilité d'une Chaine de Markov Doeblin-Recurrence: Etude du cas du process autoregressive general d'order 1. *Stochastic processes and their Applications*, **15**, 271–294.

16. Epanechnikov, V. A. Nonparametric Estimation of a Multivariate Probability Density. *Theory of Probability and its Applications*, **14**, 153–158.

17. Gaβer, T. and Mül'r, H. G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.*, **757**, 23–68.

18. Gaβer, T. and Müller, H. G. (1984). Estimating regression functions and their derivatives by the method of kernel.

19. Haggan, V., Heravi, S. M., Priestley, M. B. (1984). A study of the application of state-dependent models in nonlinear time series analysis. *J. Time Series*, **5**, 69–102.

20. Härdle, W. (1984). A LIL for nonparametric regression estimators. *Annals of Mathematical Statistics*, **12**, 624–635.

21. Härdle, W. and Gaβer, T. (1985). On robust kernel estimation of derivatives of regression functions. *Scan. J. Statist.*, **12**, 233–241.

22. Härdle, W. and Marron, J. S. (1986) Optimal bandwidth selection in

nonparametric regression function estimation. *Annals of Mathematical Statistics,* **13**, 1465–1481.

23. Johnston, G. (1979). Probabilities of maximal deviation of nonparametric regression function estimation. *J. Multiv. Anal.*, **12**, 402–414.

24. Komlós, J., Major, P. and Tusnády, G. (1975, 1976). An approximation of the partial sum of partial sums of independent RV's and the sample DF. I and II. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **32**, 111-131 and *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **34** 33-58.

25. Millar, P. (1982). Optimal estimation of a general regression function. *Annals of Mathematical Statistics,* **10**, 717–740.

26. Naradaya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications,* **9**, 141–142.

27. Parzen, M. (1962). On the estimation of the probability function and the mode. *Annals of Mathematical Statistics,* **35**, 1065–1076.

28. Phillip, (1986). Invariance principles. *Dependence in Probability and Statistics.* Birkhausse.

29. Prakasa Rao B. L. S. (1983). *Nonparametric functional estimation.* Academic Press, Orlando.

30. Priestly, M. B. (1980). State-dependent models: A general approach to non-linear time series analysis. *J. Time Series*, **1**, 47–71.

31. Révész, P. (1979). On the nonparametric estimation of the regression function. *Prob. Control Inform. Theory*, **8**, 297–302.

32. Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Mathematical Statistics*, **12**, 1215–1230.

33. Rice, J. and Rosenblatt, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Annals of Mathematical Statistics*, **11**, 141-156.

34. Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Series*, 4, 185-197.

35. Robinson, P. M. (1986). . Robust nonparametric regression. *Robust time series analysis.*

36. Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832–837.

37. Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate Analysis.*, **2** (P. R. Krishnnaiah, ed.), 25-31.

38. Rosenblatt, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference* (ed. M. Puri). Cambridge: Cambridge University Press.

39. Rosenblatt, M. (1971). Curve estimates.*Annals of Mathematical Statistics,* **42**, 1815–1842.

40. Rudemo, M. (1982). Empirical Choice of histogram and kernel density estimators. *Scan. J. Statist.*, 9, 65–128.

41. Schuster, E. and Yacowitz, S. (1979). Contributions to the theory of nonparametric regression with application to system identification. *Annals of Mathematical Statistics,* **7**, 139–149.

42. Shapiro, (1966). *Smoothing and Approximation of Functions.* van Nonstrad, Amsterdam, Holland.

43. Silverman, B. W. (1978). Weak and strong uniform consistency of a density estimate and its derivatives. *Annals of Mathematical Statistics,* **6**, 177–184.

44. Silverman, B. W. (1982a). Kernel estimation using the fast Fourier transform. *Appl. Statist.*, **31**, 93-99.

45. Silverman, B. W. (1982b). On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Mathematical Statistics,* **10**, 785–810.

46. Silverman, B. W. (1986). *Density estimation for Statistics and Data Analysis.* Chapman and Hall, London.

47. Singh, R. S. (1981). On the exact asymptotic behaviour of estimators of a density and its derivatives.*Annals of Mathematical Statistics*, 9, 453-456.

48. Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Mathematical Statistics*, 5, 595–645.

49. Stone, C. J. (1980). Optimal rates of convergence in nonparametric estimators. *Annals of Mathematical Statistics*, 8, 1348–1360.

50. Stone, C. J. (1982). Optimal global rates of convergence in nonparametric regression. *Annals of Mathematical Statistics*, **10**, 1040–1056.

51. Stone, C. J. (1984). An Asymptotically optimal window selection rule for kernel density estimates. *Annals of Mathematical Statistics*, **12**, 1285–1297.

52. Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Mathematical Statistics*, **13**, 689–705.

53. Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Mathematical Statistics*, **14**, 590–606.

54. Thuam, P. D., Tran, (1985) Absolute regularity. *Stochastics processes and Applications*.

55. Titterington, D. M. (1985). Common structure of smoothing techniques in statistics. *Int. Statist. Rev.*, **53**, 141-170.

56. Vapnik, V. N. and Cervonenkis, A. Ya. (1971). On the uniform convergence of the relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, **16**, 264–280.

57. Vapnik, V. N. and Cervonenkis, A. Ya. (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications*, **26**, 532–553.

58. Yakowitz, S. (1985a). Nonparametric density estimation, prediction and regression for Markov sequences. *Journal of the American Statistical Association*80, 215–221.

59. Yakowitz, S. (1985b). Markov flow models and the flood warning problem. *Water Resourc. Res.*, **21**, 81-88.

60. Yakowitz, S., and Szidarovsky, F. (1985). A Comparison of Kriging with nonparametric regression methods. *J. Multiv. Anal.*, **16**, 21-53.

61. Yakowitz, S. (1987). Nearest-neighbour methods for time series analysis. *J. Time Series*, **8**, 235–247.

62. Watson, G. S. (1964). Smooth regression analysis. *Sankhya Ser A*, **26**, 359–372.

Chapter 6

CONCLUSIONS

6.1 Summary

In this thesis we have investigated some aspects of certain nonstandard time series procedures chosen for their relevance in stochastic hydrology. We investigated the behaviour of the mean of FARMA models and found that it is filtered out by the fractionally differencing operator. This has implications on the estimation of the parameters of the model, because the slower rate of convergence of the sample mean to normality of long term memory processes does not affect the rate of convergence to normality of the estimated parameters. Hence, we were able to extend Li and McLeod's (1986) asymptotic normality results for FARMA processes to the case of an unknown mean. After this, the efficiency of the sample mean was investigated and we found that is not efficient when the fractionally differencing parameter is negative. Next, the efficiency of ordinary lest squares estimates of the regression parameters of a process with FARMA noise was considered. This efficiency is linked to the behaviour, as the sample size increases, of the extreme eigenvalues of the autocovariance matrix of the FARMA process.

In the definition of FARMA processes the entire past has to be taken into consideration. This can be done in two manners. One is to use an infinite autoregressive representation of the process. The other is to use the partial correlation coefficients together with the Levinson-Durbin algorithm. An ARMA process with a moving average part may also be envisioned as a process with an infinite autoregressive component. However, the difference with FARMA models is in the rate of decay of the autoregressive coefficients to zero. For ARMA models, they decay to zero exponentially fast, whereas for FARMA models they do so hyperbolically fast. This slow rate of convergence of FARMA models should have an effect on finite sample behaviour of maximum likelihood estimates. Now, it should be noted that even for ARMA models there are simulation results that indicate that certain methods of estimation are superior to maximum likelihood estimates, when the sample size is small. One method that seems to have superior finite sample properties than the maximum likelihood estimate for ARMA process is the average likelihood estimate. We investigated by simulation if this is also true for FARMA models. We found that for small sample sizes, where small is qualified by the above considerations on the hyperbolic rate of decay to zero of the autoregressive parameters, this is true. However, as the differencing parameter approaches the nonstationary boundary the sample size necessary for both methods to have similar finite sample properties decreases. As stated above, we may see FARMA processes as defined by their partial correlation coefficients. This is appealing for two reasons. One is that we do not need to use an infinite autoregressive representations of the process, and, in practice, approximations to this infinite representation. The other is that we may generalize FARMA models by defining process having partial correlation coefficients that are not summable.

Next, we considered multivariate generalizations of FARMA processes. However, full ARMA processes are difficult to identify and physical insights have been used for restricting the family of processes under consideration. A particular useful restriction, for stochastic hydrology, as well for other fields, is the contemporaneous ARMA family, where the time series are just related through the present values of the white noise process. We generalized FARMA models to contemporaneous FARMA models, using both the infinite autoregressive representation and innovation process representation.

Seasonal behaviour is a most important feature of many geophysical time series. The PARMA family is a generalization of an ARMA process that is based on geophysical considerations. A particular type of PAR(1) models have been used extensively in stochastic hydrology for fitting to monthly riverflows. In this thesis, we considered the state-space representation of PARMA models. This is interesting, both for the intrinsic value of the state-space view of the process and because it is ideally suited for the fitting and forecasting of time series. This allows us to give some conditions that are necessary for the process to be stationary. Also, we gave an exact maximum likelihood estimation algorithm based on the state-space representation and the Kalman filter. This algorithm is faster than previous algorithms, and allows us to fit models with a moving average component. For identification purposes, it is useful to use the autocovariance function, the partial autocovariance function and the inverse autocovariance function. Sakai (1982) defined the partial autocovariances and gave a Levinson-Durbin type of algorithm for their computation. However he stated that these coefficients are not useful for identification as they do not have a cut-off property. We derived his algorithm to show that they do have a cut-off property that may be used for identification. We defined the inverse autocovariances and gave an algorithm for their computation. We worked out an algorithm similar to McLeod's (1975) algorithm for the computation of the autocovariance function.

These advances together with the residual analysis presented by McLeod and Hipel (1978) for PAR models, have an straightforward generalizations to PARMA processes, and extend Box-Jenkins' fitting procedures to PARMA models. Hence, it is expected that both for their intrinsic physical relevance and for the flexibility of Box-Jenkins' approach to modelling, PARMA models should have an important role in the modeling of seasonal geophysical time series.

Another commonly used fitting procedure is to think of the time series as the sum of different components, that are largely independent among them. We may consider the case when some of these components are periodic, and for this case PARMA models for these components are a natural consideration. We have considered this extension of PARMA models, together with the smoothness priors approach of Kitawaga and Gersch (1984) that allows us to fit a component type model after deciding upon the amount of the variance that each component explains.

A further generalization should be the consideration of nonlinear and/or nonGaussian periodic processes. We have done this for the case when the marginal univariate density is exponential.

Bootstrapping regression models have been previously done by resampling the fitted residuals and generating new observations by feeding these residuals to the regression model that we were using. This approach has two drawbacks. One is that outliers will affect the performance of the bootstrap. Second, we impose a particular model on the observations, that may not be true. However, after using the fitted residuals to obtain new observations, it becomes the 'true' model. An idea put forward by Cover and Unny (1986) in a stochastic hydrology context is an alternative to this definition of the bootstrap.

We have extended Cover and Unny's bootstrapping procedure to a statistical context, by thinking of the bootstrap as a resampling of the loss function that we are using for estimation of the parameters. This idea seems to us to be interesting in its own right and it could have ramifications to other models in addition to just regression-like models. We have investigated the asymptotic behaviour of the distribution function of a bootstrapped parameter to see if it approaches the distribution of the estimated parameter and how fast it does that. These results are valid for use with stationary processes satisfying some mixing conditions.

Nonparametric function fitting may be considered when we do not know enough about the behaviour of the process under consideration in order to restrict it to a small parametric family of models. The increasing acceptance of this approach for fitting models may be seeing by considering the popularity of the Lowness statistical procedure (Cleveland, 1977). In time series analysis, nonlinear behaviour is apparent by the existence of limit circles, threshold levels and resonance. There are several statistical time series models that model certain nonlinear aspects of a time series process. However, generally we do not know enough about the nature of the nonlinearity to consider the use of these models. In a sense, it is easier to decide if the process is linear or not than to decide the nature of its nonlinearity.

We have attempted to generalize many of the known results for kernel nonparametric function fitting to stochastic processes with a time dependent structure. In order to obtain these results we have to use some asymptotic mixing condition. Some of these results have appeared in the literature, generally with stronger conditions and weaker results. Hence, we have studied the asymptotic behaviour of nonparametric density estimates and their derivatives. We have given an extension of Silverman's (1978) strong approximation result for the nonparametric kernel estimation of the density and its derivatives. We have extended Rudemo's (1981) data based approach for choosing the smoothing constant. However, if we just want to have a feeling about the shape of the density we proposed to chose the smoothing constant by preserving a set of quantiles. We studied similar questions for nonparametric kernel regression fitting.

Sometimes we may have an idea the characteristics that the density function of the process possesses. Moreover, we may be able to give a parametric family of models, such as the sought after unknown function should be close to this parametric family. We have attempted to give an estimate of the unknown function in this situation by using a convex combination of the parametric with the nonparametric estimate. However, the rate of convergence of this estimate is dominated asymptotically by the nonparametric estimate. Nonetheless, we may consider the use of this estimate because it should have better finite sample properties than the other two estimates.

Another situation where we may want to combine the parametric function fitting approach with the nonparametric function fitting approach is when we have some idea about the behaviour of the unknown function over some range of its domain of definition but not over the whole domain. For example, we may want to assume that the tails of the unknown probability density are exponential. We have to try to give a reasonable parametric-nonparametric estimate of this situation by blending the parametric with the

In a time series context it is interesting to consider not full nonparametric families but certain restrictions to these families. For example, the state dependent models of Priestly (1980) seem to constitute a resonable nonparametric family for time series analysis. We have produced a kernel nonparametric fitting estimate that is based on least squares with state-dependent weights.

Finally, we may want to combine ARMA model fitting with nonparametric fitting. For example, if we assume a ARMA noise plus an unknown trend, or a AR(1) model with a nonlinear transfer function. We have approached these situations by using a least squares approach together with state-dependent weights for the nonparametric part together with a least squares approach for the global ARMA coefficients.

6.2 Further research

For FARMA models a hypothesis testing procedure is needed for deciding upon the suitability of FARMA models. This may be done now by using the asymptotic normality of the estimated fractionally differencing parameter. But it does seem that this approach does not have good finite sample properties as is evident by the case of the Gota River in Chapter 1. One problem may be that the standard deviation does not depend on the value of the parameter. Another interesting problem is to obtain statistical results about the existence of long-term memory processes that are Markovian or depend on a few past observations plus on a few past white noise inputs.

In a forecasting study, Noakes et al. (1987) found that, in general, FARMA models are not competitive with ARMA models or with the nonparametric function fitting approach when one studies the behaviour of the one step-ahead forecasts. However, FARMA models may be competitive when we consider forecasts with larger horizons, as FARMA models try to fit an overall property of the time series and not just its behaviour one step ahead.

Another interesting investigation may be the study of the existence of FARMA type models for processes whose range of valid values is discrete.

The averaging likelihood method deserves further investigation in both its theoretical properties and its practical applications. A suitable algorithm should be developed for use with several parameters.

We should consider the extension of PARMA models to nonlinear models. An interesting posibility maybe the class of periodic bilinear models, because they would allow higher order spectra behaviour to enter into the model definition. Also,, we may extend the study of Noakes et al. (1985) regarding the forecasting ability of PAR models to the case of PARMA models.

For the nonparametric procedure we basically need much more knowledge on both its theoretical properties and on its value in practical applications.

For the nonparametric function fitting, we need more knowledge about its practical applications compared with the traditional parametric function fitting methods. Also, it should be interesting to investigate if we can obtain a combination of nonparametric with parametric function estimates that has a better order of convergence than the nonparametric function estimate.

References

1. Cover, K. A. and Unny, T.E. (1986). Application of computer intensive statistics to parameter uncertainty in streamflow synthesis. *Water Resources Bulletin*, **22**, 495–499.

2. Kitagawa G. and W. Gersch (1984), A Smoothness Priors-State Space Modeling of time Series with Trend and Seasonality, *Journal of the American Statistical Association*, **79**, 378–389.

3. Li, W. K. and A. I. McLeod, (1986). Fractional time series modelling. *Biometrika*, **73**, 165–176.

4. McLeod, A. I. and Hipel, K. W. (1978). Developments in monthly autoregressive modelling. Technical report No. 45–XM-011178, Dept. of Systems Design Engineering, Univ. of Waterloo, Ontario, Canada.

5. McLeod A. I. (1975). The derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Appl. Statist.*, **24**, 255–256.

6. Noakes, D. J., McLeod, A. I., Hipel, K. W. (1979). Forecasting monthly riverflows time series. *International Journal of Forecasting*, 1, 179–190.

7. Noakes, D. J., Hipel, K. W., Mcleod, I. A., Jimenez, C., and Yacowizt, (1987). *International Journal of Forecasting*, 1, 179–190.

8. Priestly, M. B. (1980). State-dependent models: A general approach to non-linear time series analysis. *J. Time Series*, 1, 47–71.

9. Rudemo, M., (1982). Empirical Choice of histogram and kernel density estimators. *Scan. J. Statist.*, 9, 65–128.

10. Sakai, H., (1982). Circular Lattice Filtering Using Pagano's Method. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-30, 279–286

11. Silverman, B. W., (1978). Weak and strong uniform convergence of the kernel estimate of a density and its derivatives. *Annals of Mathematical Statistics*, 6, 177–184.

Now we will consider the asymptotic normality of $\hat{\theta}_\omega$.

THEOREM **4.11** We assume the same conditions on ω_i, T_i and Θ as in theorem (4.9). Also, we assume conditions on T_i sufficient to ensure the asymptotic normality of $\hat{\theta}$ with covariance matrix

$$I^{-1} = \boldsymbol{E}\{\varphi\} \tag{4.34}$$

Also, assume that T_i can be expanded such that

$$\hat{\theta}_n - \hat{\theta}_0 = (\frac{1}{n}\sum \phi_i'(\theta_0))^{-1}(\frac{1}{n}\sum \phi_i(\theta_0)) + \dots$$

such that $\boldsymbol{E}|\phi'(\theta_0; X)|$ exists, and $\phi(\theta_0; X)$ satisfies conditions that assure the Central Limit Theorem. Then $\sqrt{N}(\hat{\theta}_n - \theta_0)$ and $\sqrt{N}(\hat{\theta}_\omega - \hat{\theta}_N)$ have asymptotically the same distribution. Moreover, if $\boldsymbol{E}\phi^3(\theta_0; x)$ is defined, then the distance between their probability laws is of order $1/\sqrt{n}$. ▽

4.3.3 The mean

We have observations $y_i, i = 1, \dots, N$, with mean μ, and finite variance σ_i^2, with out loss of generality we will assume that the unknown true mean $\mu = 0$. The estimate of μ is obtained by minimizing

$$\sum_{i=1}^{N} \omega_i (y_i - \mu)^2 \tag{4.35}$$

where the weights ω_i are positive independent and identically distributed random variables, with mean one and finite variance σ_ω^2. The estimate corresponding to the set of weights $\omega = (\omega_i, \dots, \omega_N)$ is given by

$$\begin{aligned} \hat{\mu}_\omega &= \frac{\sum_{i=1}^{N} \omega_i y_i}{\sum \omega_i} \\ &= \bar{y} + \frac{\sum_{i=1}^{N} \omega_i (y_i - \bar{y})}{\sum_{i=1}^{N} \omega_i}. \end{aligned} \tag{4.36}$$