

Introduction to Whittle (1953) "The Analysis of Multiple Stationary Time Series"

Matthew Calder, Colorado State University
Richard A. Davis, Colorado State University

1 Introduction

During the 1940's and 1950's the subject of statistical time series analysis matured from a scattered collection of methods into a formal branch of statistics. Much of the mathematical groundwork was laid down and many of the practical methods were developed during this period. The foundations of the field were developed by individuals whose names are familiar to most statisticians: Yule, Wold, Wiener, and Kolmogorov. Interestingly, many of the methodological techniques that originated 40 to 50 years ago still play a principal role in modern time series analysis. A key participant in this golden-age of time series analysis was Peter Whittle.

Statisticians have looked at temporally sequential data, or time series, since the recording of numbers began. However, it was not until Yule developed the concept of the autoregressive sequence that the modelling of time series was considered to be of a substantially different character than other statistical modelling problems. Prior to that, time series were modelled as the sum of a deterministic function and a random noise component, in much the same way as any other regression problem. Yule's autoregressive model on the other hand, relates the present value of a series directly to it's past. This seemingly obvious step raises a host of fundamental questions concerning the existence, uniqueness, and characterization of such processes.

Many of these questions were answered by the spectral theory of Wiener and Kolmogorov. The spectral theory connects the autocorrelation function of a series to a spectral distribution function, via the Fourier transform. This leads to the spectral representation of a stationary process which underpins much of modern time series analysis. Complimenting the spectral theory was a time domain theory that culminates in the decomposition theorem of Wold. This alternative representation decomposes a stationary time series into a superposition of an infinite order moving average and a stationary deterministic process. Both of these theories put Yule's autoregressive models on a firm mathematical footing and set the stage for the flurry of activity that was to come.

2 Historical Background

One of the main results in "The Analysis of Multivariate Stationary Time Series" is an approximation to the likelihood of a stationary multivariate Gaussian process. This approximation is a generalization of the univariate case previously derived in Whittle (1951). The development of the result in the univariate case parallels that of the multivariate case, and the form of the results are similar. A quick synopsis of the univariate approximation will shed some light on the more general results.

Consider a set of observations X_1, \dots, X_N to be taken from a univariate stationary Gaussian process $\{X_t\}$ with mean zero. Assume that the correlation function of the time series can be parameterized by the vector $\beta = (\beta_1, \dots, \beta_m)'$ and that the spectral density $\frac{\sigma^2}{2\pi}g(\cdot; \beta)$ exists, i.e.,

$$\text{Cov}(X_s, X_t) = \frac{\sigma^2}{2\pi} \int_0^{2\pi} e^{i(s-t)\omega} g(\omega; \beta) d\omega.$$

Then, ignoring constants, $-2\log$ of the likelihood of the parameter vector $\theta = (\beta, \sigma^2)$ based on the observation vector $\mathbf{X} = (X_1, \dots, X_N)'$ is given by:

$$L(\theta) = \sigma^{-2} \mathbf{X}' G^{-1} \mathbf{X} + \log |\sigma^2 G|, \quad (1)$$

where $\sigma^2 G = \sigma^2 G(\beta)$ is the covariance matrix of \mathbf{X} . The maximum likelihood estimate, $\hat{\theta}$, is found by maximizing $L(\theta)$ over the allowable parameter space. Explicit solutions to this maximization problem rarely exist, and hence maximization of the likelihood is performed using numerical optimization techniques, requiring multiple evaluations of $L(\theta)$. These evaluations can be daunting due to the complicated matrix operations involved, and must have been especially so in 1953 before the advent of high speed digital computers and sophisticated optimization techniques.

Whittle's idea was to approximate the quadratic form in such a way as to avoid computing the explicit inverse and determinant of G . The approximation is based on the observation that G can be approximated by a circulant matrix, \tilde{G} , which in turn can be easily diagonalized by the matrix U consisting of rows,

$$\mathbf{u}_j = N^{-\frac{1}{2}} (1, e^{i\omega_j}, e^{2i\omega_j}, \dots, e^{(N-1)i\omega_j}),$$

where $\omega_j = 2\pi j/N$ represent the Fourier frequencies (see Brockwell and Davis (1991) p. 134-136). In particular,

$$UGU' \approx U\tilde{G}U' = D = \text{diag}(g(\omega_0; \beta), \dots, g(\omega_{N-1}; \beta))$$

and

$$U\mathbf{X} = \mathbf{Z} = (Z_0, \dots, Z_{N-1})',$$

where $Z_j = N^{-\frac{1}{2}} \sum_{t=0}^{N-1} e^{itw_j} X_{t+1}$ is the j^{th} ordinate of the discrete Fourier transform of \mathbf{X} . Using these properties, it follows that,

$$\mathbf{X}'G^{-1}\mathbf{X} \approx \mathbf{X}'\tilde{G}^{-1}\mathbf{X} = \mathbf{X}'U'D^{-1}U\mathbf{X} = \sum_{j=0}^{N-1} \frac{|Z_j|^2}{g(w_j; \beta)}$$

and

$$\log |\sigma^2 G| \approx \log |\sigma^2 U D^{-1} U'| = \sum_{j=0}^{N-1} \log(\sigma^2 g(w_j; \beta)).$$

These approximations suggest replacing the log-likelihood in (1) by

$$L_W(\theta) = \sum_{j=0}^{N-1} \left(\log(\sigma^2 g(w_j; \beta)) + \frac{|Z_j|^2}{\sigma^2 g(w_j; \beta)} \right). \quad (2)$$

In the univariate case the likelihood can be further simplified by first maximizing over the parameter σ^2 . This can be done independently of the other parameters, giving the estimate,

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{j=0}^{N-1} \frac{|Z_j|^2}{g(w_j; \beta)}$$

Substituting this estimate back into (2) gives the concentrated likelihood,

$$L_W(\beta, \hat{\sigma}^2) = N \log \left(\frac{1}{N} \sum_{j=0}^{N-1} \frac{|Z_j|^2}{g(w_j; \beta)} \right) + \sum_{j=0}^{N-1} \log(g(w_j; \beta)), \quad (3)$$

which is a function of β only. Equation (2) is known as the Whittle approximation to the likelihood. Calculation of $L_W(\theta)$, and hence numerical optimization of $L_W(\theta)$, is substantially simpler than that based on $L(\theta)$ since the quadratic form, $\mathbf{X}'G^{-1}\mathbf{X}$, and the determinant, $|G|$, are not explicitly calculated. Also the parameters come in only through the spectral density, which for many types of models has an explicit formula. The periodogram must be calculated, but it too can be calculated quite easily via the fast Fourier transform.

The estimation procedure for multivariate processes derived in "The Analysis of Multiple Stationary Processes" minimizes a quantity analogous to (2) above. Although there are some fundamental differences between the univariate and multivariate cases, it is comforting to see that Whittle's approximation to the likelihood takes essentially the same form for both.

3 Summary of the Paper

In Section 1 of the paper, Whittle stresses the need to develop methods for the analysis of multivariate time series. Much of the theory of time series

is concerned with a single sequence of measurements, however, data obtained in practice are often multidimensional. Economic time series might consist of price indices, interest rates, and cost indicators all measured simultaneously. Radar data are often generated by antennae arrays, with each antennae producing its own time series. And as any audiophile will tell, stereo sound is much more interesting than mono.

For a standard linear model, the least squares estimate is the same as that obtained by maximizing the likelihood assuming a Gaussian model. Analogously, Whittle describes his estimator as "least squares" because it is derived by maximizing the Gaussian likelihood even though the underlying model is not necessarily Gaussian. In the vernacular of modern time series, such estimates are called "maximum (Gaussian) likelihood" while "least squares" refers to estimates that are obtained by minimizing the quadratic form occurring in the Gaussian likelihood.

A first step in Whittle's approximation to the Gaussian likelihood is to express the process $\{\mathbf{X}_t\}$ as an infinite order moving average,

$$\mathbf{X}_t = \sum_{k=0}^{\infty} C_k \boldsymbol{\eta}_{t-k}, \quad (4)$$

where $\{C_k\}$ is a sequence of matrices with C_0 equal to the identity, and $\{\boldsymbol{\eta}_t\}$ is a sequence of uncorrelated random vectors with covariance matrix \mathbb{X} . Assuming (4) can be inverted, this yields the AR representation,

$$\mathbf{X}_t - \sum_{k=0}^{\infty} D_k \mathbf{X}_{t-k} = \boldsymbol{\eta}_t,$$

Apart from end effects, $(\mathbf{X}_1, \dots, \mathbf{X}_N)$ can be expressed in terms of $(\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_N)$ and vice versa so that $-2\log$ of the Gaussian likelihood of $(\mathbf{X}_1, \dots, \mathbf{X}_N)$ is approximately

$$\sum_{t=1}^N \boldsymbol{\eta}_t' \mathbb{X}^{-1} \boldsymbol{\eta}_t + N \log |\mathbb{X}|.$$

In the remainder of the introduction, Whittle points out some of the difficulties that arise in the multivariate case that are not present in the univariate case. One major hurdle is the lack of separation of the parameters describing the "essential features" of the process, from those describing the "variances of the residuals". For example, in the MA representation in (4), the estimation of \mathbb{X} cannot be done independently of the estimation of the $\{C_k\}$ as it was in the univariate case (see (3) above). In addition, if the matrices C_k are functions of a parameter vector $\boldsymbol{\beta}$, then the asymptotic variance of the maximum likelihood estimate of $\boldsymbol{\beta}$ will typically depend on \mathbb{X} .

In Section 2 Whittle discusses the Wold decomposition, similar to that given in (4). It is noted that, in general, the MA coefficient matrices $\{C_k\}$

are not uniquely determined by the spectral matrix function $F(\omega)$. That is, it may not be possible to discriminate between different parameterizations of the model from a given F . Whittle also gives conditions necessary to invert the MA representation to obtain an autoregressive (AR) representation. He further connects the spectral matrix function to the transfer functions corresponding to the respective AR and MA filters.

In Section 3 Whittle derives an expression for the prediction variance of a multivariate stationary process. In the multivariate case there is no scalar prediction error, but rather a vector of prediction errors (the η_t in (4)) with an associated covariance matrix. The determinant of this matrix is taken as the measure of prediction error, and is found to be,

$$V = \exp \left\{ \frac{1}{2\pi} \int_0^{2\pi} \log |F(\omega)| d\omega \right\}, \quad (5)$$

This is analogous to Kolmogorov's formula for the prediction variance of a univariate stationary process. Also, $V^{\frac{N}{2}}$ is approximately the Jacobian of the transformation from the likelihood of the observations $\{\mathbf{X}_t\}$ to the likelihood of the standardized moving average noise terms $\{\Phi^{-\frac{1}{2}}\eta_t\}$. This fact is used to derive the estimation equations in Section 5.

In Section 4 Whittle derives asymptotic formulas for the cumulants of arbitrary linear functions of the sample cross-covariances of a multivariate stationary Gaussian process. Corresponding to any such linear function ξ Whittle shows there exists a Hermitian matrix function $Q(z)$ such that,

$$\xi = N \int_0^{2\pi} \text{tr}(Q(e^{i\omega})f(\omega)) d\omega,$$

where $f(\omega)$ is the periodogram matrix function,

$$f(\omega) = \frac{1}{N} \left(\sum_{t=1}^N \mathbf{X}_t e^{it\omega} \right) \left(\sum_{s=1}^N \mathbf{X}_s e^{is\omega} \right)'$$

The cumulants $k^{(r)}(\xi)$ are then shown to satisfy

$$k^{(r)}(\xi) \sim \frac{2^{r-1}(r-1)!N}{2\pi} N \int_0^{2\pi} \text{tr}(Q(e^{i\omega})F(\omega))^r d\omega.$$

which are subsequently used to derive the limiting behavior of the likelihood function and its derivatives.

In Section 5 the approximation to the likelihood is derived in a surprisingly direct manner, using the infinite moving average representation of the process. The sum of squares of the noise terms is expressed as,

$$\sum_{t=1}^N \eta_t' \Phi^{-1} \eta_t \approx \frac{N}{2\pi} \int_0^{2\pi} \text{tr}(f(\omega)F^{-1}(\omega)) d\omega,$$

which, when combined with (4), implies that $-2\log$ of the likelihood can be approximated by,

$$L_W(\theta) = \frac{N}{2\pi} \int_0^{2\pi} \log(|F(\omega; \theta)|) + \text{tr}(f(\omega)F^{-1}(\omega; \theta)) d\omega. \quad (6)$$

Compare this expression (6) for $L_W(\theta)$ to that obtained in the univariate case (3).

In Section 6 Whittle argues that when the process with true parameter $\theta = \theta_0$ is Gaussian, $\frac{\partial L_W(\theta_0)}{\partial \theta}$ is asymptotically normal with mean 0, and covariance matrix

$$2M = \left[\frac{N}{4\pi} \int_0^{2\pi} \text{tr} \left[\frac{\partial F}{\partial \theta_j} F^{-1} \frac{\partial F}{\partial \theta_k} F^{-1} \right] d\omega \right]_{j,k},$$

and

$$\frac{\partial^2 L_W(\theta)}{\partial \theta^2} = M + o_p(N).$$

It now follows from standard Taylor series expansions of $L_W(\theta)$ and assuming consistency of the estimators that $\hat{\theta}$ is asymptotically normal with mean θ_0 and covariance matrix $2M^{-1}$.

Section 7 uses expansions similar to those in Section 6 to develop testing procedures. A generalized likelihood ratio test based upon the prediction variance V of Section 3 is derived and an example is given demonstrating its use.

In Section 8 a more complete example involving a bivariate series of sunspot measurements is given. The model used in the example is of a simplified symmetrical type, but both fitting and testing procedures are demonstrated.

4 Subsequent Developments

The approximation given in (6) is often referred to as the Whittle likelihood in the literature. It has taken a progression of researchers to nail down all of the details and to formalize the *theorems* of Whittle in both the univariate case (Walker (1964) and Hannan (1973)) and in the multivariate case (Dunsmuir and Hannan (1976); Deistler, Dunsmuir, and Hannan (1978); Pötscher (1987); and Dahlhaus and Pötscher (1987)).

Much of this work centered on questions of identifiability, that are intimately connected with establishing consistency of the estimators. Whittle pays little attention to this important problem which for the case of multivariate ARMA models is not a simple matter. Consider for example the following bivariate AR(1) model

$$\begin{aligned} X_{1t} &= \phi X_{2,t-1} + \eta_{1t} \\ X_{2t} &= \eta_{2t}, \end{aligned}$$

where $\boldsymbol{\eta}_t = (\eta_{1t}, \eta_{2t})'$ are IID random vectors. This model also has the MA(1) representation,

$$\begin{aligned} X_{1t} &= \eta_{1t} + \phi\eta_{2,t-1} \\ X_{2t} &= \eta_{2t}, \end{aligned}$$

and hence the likelihood cannot discriminate between either of these two models.

A more subtle but important point not addressed by Whittle is the quality of the approximation to the exact Gaussian likelihood. For example, it is not readily apparent that the estimate based on Whittle's approximation, and that based on the exact Gaussian likelihood are asymptotically equivalent. Subsequent results have shown this asymptotic equivalence for a wide class of time series models and extended the results of Sections 6 and 7 to cases where the underlying distribution is non-Gaussian.

Although the Whittle likelihood provides a simple approximation to the Gaussian likelihood, it does not appear to have been used much for ARMA modelling. This may be in part due to the work being "done too early ... to reap its due recognition" (Whittle (1986)) and hence simply overlooked, or perhaps because of the development of alternative methods, such as the Box-Jenkins backcasting algorithm. Today there is no reason to use either of these methods for this purpose. Recursive techniques based upon the Kalman filter and state-space representations can be used to compute exact Gaussian likelihoods efficiently.

There has been a resurgence of interest in the Whittle approximation to the likelihood. For fractionally integrated ARMA models and certain classes of spatial models, finite state space representations do not exist which preclude the use of recursive methods to calculate the likelihood efficiently. Nevertheless, these models do have explicit expressions for the spectral density and therefore Whittle's approximation becomes an indispensable tool.

5 Biography

Peter Whittle was born in Wellington, New Zealand in 1927. He became interested in stochastic processes and time series while working for the New Zealand Department of Scientific and Industrial Research. After completing B.Sc. degrees in mathematics and physics, and a M.Sc. degree in mathematics in 1948 he began post graduate work with Dr. Herman Wold at the University of Uppsala, Sweden. During this time he, in his own words, "constructed the asymptotic likelihood theory for the stationary Gaussian process, essentially if unrigorously."

After earning a doctorate of philosophy in 1951, he returned to New Zealand and the Department of Scientific Research. He continued work in

time series and also worked on problems involving nonlinear models, spatial models, random partial differential equations, and reversibility. After spending a year in Canberra, he began working on problems of prediction and optimization. It was at this time, 1959, that he moved to Britain to join the Statistical Laboratory at Cambridge. In 1961, he became the head of the Statistical Laboratory at Manchester University, replacing Maurice Bartlett for whom he has a deep reverence. In 1966 he moved on to the Churchill Chair in the Mathematics of Operational Research at Cambridge.

REFERENCES

- Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd edition, Springer, New York.
- Dahlhaus, R. and Pötscher, B.M. (1989). Convergence results for maximum likelihood type estimators in multivariable ARMA models II, *J. Multivariate Anal.* **30**, 241–244.
- Dunsmuir, W.J.M. and Hannan, E.J. (1976). Vector linear time series models, *Adv. in Appl. Probab.* **8**, 339–364.
- Hannan, E.J. (1973). The asymptotic theory of linear time series models, *J. Appl. Prob.* **10**, 130–145.
- Pötscher, B.M. (1987). Convergence results for maximum likelihood type estimators in multivariable ARMA models, *J. Multivariate Anal.* **21**, 29–52.
- Walker, A.M. (1964). Asymptotic properties of least squares estimates of parameters of the spectrum of a stationary non-deterministic time series. *J. Austral. Math. Soc.* **4**, 363–384.
- Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis*. Thesis, Uppsala University, Almqvist and Wiksell, Uppsala.
- Whittle, P. (1986). In the late afternoon, *The Craft of Probabilistic Modelling*, edited by J. Gani, Springer, New York.