Autocorrelation-Robust Inference*

P. M. Robinson and C. Velasco

1. Introduction

Time series data occur commonly in the natural and engineering sciences, economics and many other fields of enquiry. A typical feature of such data is their apparent dependence across time, for example sometimes records close together in time are strongly correlated. Accounting for serial dependence can considerably complicate statistical inference. The attempt to model the dependence parametrically, or even nonparametrically, can be difficult and computationally expensive.

In some circumstances, the serial dependence is merely a nuisance feature, interest focusing on "static" aspects such as a location parameter, or a probability density function. Here, we can frequently base inference on point or function estimates that are natural ones to use in case of independence, and may well be optimal in that case. Such estimates will often be less efficient, at least asymptotically, than ones based on a comprehensive model that incorporates the serial dependence. But apart from their relative computational simplicity, they often remain consistent even in the presence of certain forms of dependence, and can be more reliable than the "efficient" ones, which can sometimes become inconsistent when the dependence is inappropriately dealt with, leading to statistical inferences that are invalid, even asymptotically. In this sense the estimates can be more "robust" than "efficient" ones, and the main focus in this paper is their use in statistical inference, though we also discuss attempts at more efficient inference.

The stress in this paper is on large sample inference, due to the relative intractability of finite-sample properties in all but a few situations. First order asymptotic theory is relatively simple to use, and also often relies on milder assumptions than finite-sample, or even higher order, asymptotic theory, and can thus be more widely applied, at least given data sets of reasonable size. Even a comprehensive discussion of first order asymptotics is impossible, however, because the form of the asymptotic distribution can depend on the nature of the dependence structure, which can vary immensely, across many forms of

stationarity and a vast range of nonstationary processes. We mainly stress inferences based on asymptotically normal statistics. Here, a very important role is played by stationarity or asymptotic stationarity, and indeed on further restrictions on the dependence of a (possibly unobservable) process associated with the data and the statistical model.

Valid inference based on an asymptotically normal statistic requires only a consistent estimate of the variance matrix in the limiting distribution. Then the statistic can be studentized, consistent confidence regions set, and asymptotically valid hypothesis tests carried out, where one-sided tests are possible in single-parameter cases but two-sided ones based on the χ_p^2 distribution would be used in p-parameter cases with $p > 1$. Usually the variance is affected by the dependence, and requires a different, and more complicated, type of estimate than that under independence. This can be based on a parametric model for the autocorrelations. However nonparametric types of estimate are more popular, being consistent under broader conditions.

The next section discusses inference on the mean of a covariance stationary series, based on the sample mean. The methods used to robustify many of the more complicated statistics discussed subsequently, and especially methods used recently in econometrics, can be seen as extensions of those proposed much earlier for the mean by such authors as Jowett (1955), Hannan (1957) and Brillinger (1979). The limiting variance is proportional to the spectral density at frequency zero of the process, and so we briefly describe some methods of spectrum estimation. Section 3 considers inference on slope parameters in linear regression, using least squares. While the methods of Section 2 can be used here, an alternative approach due to Eicker (1967) is described. Section 4 discusses extensions to M-estimates of location and regression models, and other robust estimates. Section 5 considers nonlinear and more general econometric models. In each case, nonparametric spectrum estimation is involved in one way or another, and Section 6 discusses the important problem of selecting bandwidth numbers. Section 7 departs from the first-order asymptotics stressed previously to consider higher order theory, and the bootstrap. In each of Sections 2–7, the type of dependence assumed is weak, or short-term, and the methods of inference considered differ from those under independence. The final two sections each depart from one of these features. Section 8 concerns long range dependence, while Section 9 concerns inference on smoothed probability density and regression estimates.

2. Inference based on the sample mean

Let $\{X_t, t = 1, 2, ...\}$ be a covariance stationary sequence of random variables with mean and lag-j autocovariance given by

$$\mu = \mathrm{E}(X_1), \quad \gamma_j = \mathrm{E}(X_1 - \mu)(X_{1+j} - \mu) \ , \tag{2.1}$$

and spectrum $f(\lambda)$ given by

$$\gamma_j = \int_{\pi}^{\pi} f(\lambda) \cos j\lambda \, d\lambda \ . \tag{2.2}$$

To estimate μ given observations $X_1, ..., X_N$, consider the sample mean

$$\overline{X} = \frac{1}{N} \sum_{t=1}^{N} X_t \ , \tag{2.3}$$

which is the ordinary least squares estimate (OLSE) of μ.

In case the X_t are also uncorrelated, that is,

$$\gamma_j = 0, \quad |j| \geq 1 \ , \tag{2.4}$$

$\overline{X}$ is the best linear unbiased estimate (BLUE), and if in addition X_t is Gaussian it is the maximum likelihood estimate (MLE), having minimum variance, γ_0/N, within the class of all regular estimates. γ_0 can be estimated unbiasedly by $Nc_0/(N-1)$ where for future use we define the sequence

$$c_j = \frac{1}{N} \sum_{1 \leq t, t+j \leq N} (X_t - \overline{X})(X_{t+j} - \overline{X}), \ 0 \leq j < N \ . \tag{2.5}$$

Then $t = (N-1)^{1/2}(\overline{X} - \mu)/c_0^{1/2}$ has the t_{N-1} distribution. If the X_t are not Gaussian, but independent and identically distributed, then $t \to_d N(0,1)$, as $N \to \infty$.

When (2.4) is not true, $\overline{X}$ is no longer a BLUE or MLE. However, it is still unbiased, and Grenander (1954) showed that if:

$$f(\lambda) \text{ is continuous and positive at } \lambda = 0 \ , \tag{2.6}$$

$\overline{X}$ is asymptotically efficient within the class of linear unbiased estimates, so no differential weighting need be considered. However, t is no longer asymptotically $N(0,1)$, indeed

$$V(\overline{X}) = \frac{\gamma_0}{N} + \frac{2}{N} \sum_{j=1}^{N} \left(1 - \frac{j}{N}\right) \gamma_j \ , \tag{2.7}$$

and the second component is only zero for all N under (2.4). We have nevertheless the central limit theorem (CLT)

$$\frac{\overline{X} - \mu}{V(\overline{X})^{1/2}} \to_d N(0,1) \tag{2.8}$$

under various conditions permitting dependence in X_t, such as (see Ibragimov and Linnik, 1971, Hannan, 1979):

Condition A X_t is stationary, and for some $\delta > 0, \mathrm{E}|X_1|^{2+\delta} < \infty$ and X_t is α-mixing with $\sum_{j=1}^{\infty} \alpha_j^{2/(2+\delta)} < \infty$, where α_j is the j^{th} α-mixing number of X_t, or:

Condition B $X_t = \mu + \sum_{j=-\infty}^{\infty} \beta_j \varepsilon_{t-j}$, where the ε_t^2 are uniformly integrable, $E(\varepsilon_t | F_{t-1}) = E(\varepsilon_t^2 - E\varepsilon_t^2 | F_{t-1}) = 0$ a.s., where F_t is the σ-field generated by $\{\varepsilon_t, \varepsilon_{t-1}, \ldots\}$, and $\sum_{j=-\infty}^{\infty} |\beta_j| < \infty$.

Both conditions imply that the γ_j decay fast enough to be absolutely summable, and that $f(\lambda)$ is continuous for all λ, and so can be referred to as 'weak dependence' conditions. Thus we also have from Fejér's theorem (Zygmund, 1977, p.89)

$$NV(\overline{X}) \to 2\pi f(0), \quad \text{as } N \to \infty \ , \tag{2.9}$$

indeed this follows also from (2.6), which also implies that the right side is positive. We deduce that $\overline{X}$ is $N^{1/2}$-consistent and

$$N^{1/2}(\overline{X} - \mu) \to_d N(0, 2\pi f(0)), \quad \text{as } N \to \infty \ . \tag{2.10}$$

To do large-sample inference it remains to find an estimate $\hat{f}(0)$ of $f(0)$ such that

$$\hat{f}(0) \to_p f(0), \quad \text{as } N \to \infty \ , \tag{2.11}$$

for then

$$\frac{N^{1/2}(\overline{X} - \mu)}{\{2\pi \hat{f}(0)\}^{1/2}} \tag{2.12}$$

can be approximated by a $N(0,1)$ variate. Given a parametric model for γ_j, equivalently for $f(\lambda)$, for example a stationary autoregression (AR), invertible moving average (MA), or stationary and invertible autoregressive moving average (ARMA), we can estimate the unknown parameters, insert them in the formula for $f(0)$, and obtain an $N^{1/2}$-consistent estimate. However, (2.11) requires no rate of convergence, and there has been greater stress on using nonparametric spectrum estimates which are consistent in the absence of assumptions on functional form, though as even the early papers on studentizing the sample mean of a time series by Jowett (1955), Hannan (1957) illustrate, the same estimates of $\hat{f}(0)$ can be interpreted as either parametric or nonparametric. These authors both chose quadratic functions of X_t for $\hat{f}(0)$, mirroring estimates stressed in the early nonparametric spectrum estimation literature. These are still widely used today and are described as follows.

Introduce the mean-corrected periodogram

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{t=1}^{N} (X_t - \overline{X}) e^{it\lambda} \right|^2 \tag{2.13}$$

and, a real, even, bounded and integrable function $K(\lambda)$, satisfying

$$\int_{-\infty}^{\infty} K(\lambda)\, d\lambda = 1 \ . \tag{2.14}$$

Define

$$K_M(\lambda) = M \sum_{j=-\infty}^{\infty} K(M(\lambda + 2\pi j)) \ , \tag{2.15}$$

where M is called a "lag number", or "bandwidth" number. Weighted autocovariance estimates of $f(\lambda)$ are given as

$$\begin{aligned} \hat{f}_c(\lambda) &= \int_{-\pi}^{\pi} K_M(\lambda - \theta) I(\theta)\, d\theta \\ &= \frac{1}{2\pi} \sum_{1-N}^{N-1} k\left[\frac{j}{M}\right] c_j \cos j\lambda \ , \end{aligned} \tag{2.16}$$

where

$$k(x) = \int_{-\infty}^{\infty} K(\lambda) e^{ix\lambda}\, d\lambda \ . \tag{2.17}$$

For the same K and M, $\hat{f}_c(\lambda)$ is typically closely approximated by the averaged periodogram form

$$\hat{f}_p(\lambda) = \frac{2\pi}{N} \sum_{j=1}^{N-1} K_M(\lambda - \lambda_j) I(\lambda_j) \ , \tag{2.18}$$

where $\lambda_j = 2\pi j/N$, and because these $I(\lambda_j)$ are invariant to location shift in X_t no mean correction of X_t is necessary in (2.18).
One set of sufficient conditions for (2.11) with $\hat{f}(0)$ given by (2.16) and (2.18) is as follows (see e.g. Hannan, 1970, Chapter 5): X_t is fourth-order stationary with

$$\sum_{j=0}^{\infty} j|\gamma_j| < \infty, \quad \sum_{h,i,j=-\infty}^{\infty} \sum \left|\mathrm{cum}(X_1, X_{1+h}, X_{1+i}, X_{1+j})\right| < \infty \ , \tag{2.19}$$

$k(x)$ is continuous, $\lim_{x \to 0} |k(x) - 1|/|x| \to 0$, and $1/M + M/N \to 0$ as $N \to \infty$.

The choice of M and K has been extensively discussed, mainly from the standpoint of bias and variability. For given K, variances tend to increase with M. On the other hand, a small M can lead to $K_M(\lambda)$ not being heavily concentrated around the origin, and bias from the influence of frequencies near zero, most likely negative bias if there is a spectral peak there. For given M, a similar dilemma is faced in the choice of K. Frequencies distant from zero can also cause bias. Many K produce side-lobes in K_M, so coincidence of a side-lobe with a large spectral peak in $f(\lambda)$ can give an inflated estimate of $f(0)$. The precise relevance of moment behaviour of $\hat{f}(0)$ to (2.12) is unclear, but obviously too large an $\hat{f}$ will produce an overly narrow interval estimate for $\overline{X}$, and vice versa. On computational grounds k having compact support in $\hat{f}_c(0)$, $k(x) = 0$, $|x| > 1$, is desirable because then only about M c_j need be computed, whereas the

(contradictory) practice of choosing K to have compact support in $\hat{f}_p(0)$, so $K(\lambda) = 0$, $|\lambda| > \pi$, is desirable because then only about $N/2M$ of the $I(\lambda_j)$ need be computed. On the other hand, all $N-1$ c_j and $N/2$ $I(\lambda_j)$ can be rapidly computed via the fast Fourier transform, so these considerations are of minor importance. The desirable requirement of non-negativity of a variance estimate is implied if

$$K(\lambda) \geq 0, \text{ all } \lambda \ . \tag{2.20}$$

The estimates of $NV(\overline{X})$ proposed by Jowett (1955) and Hannan (1957) approximate the so-called truncated version of $\hat{f}_c(0)$, where

$$k(x) = 1, \quad |x| \leq 1; = 0, \quad |x| > 1, \tag{2.21}$$

but the corresponding K does not satisfy (2.20) and has relatively large side-lobes. Jowett derived his estimate without reference to spectrum estimation, and his discussion of asymptotic theory is essentially appropriate to the parametric finite MA case, with M remaining fixed as N increases. Hannan pointed out the connection with spectrum estimation, and also indicated that the more general class $\hat{f}_c(0)$ could be used. In particular, K due to Parzen and Bartlett are still commonly used, satisfying (2.20), and have more acceptable side-lobes than those inherent in (2.21). For example, the modified Bartlett k is given by

$$k(x) = 1 - |x|, \quad |x| \leq 1; = 0, \quad |x| > 1. \tag{2.22}$$

Hannan also referred to the desirability of $\hat{f}(0)$ being almost uninfluenced by μ, but noted that this conflicts with the need to reduce the bias due to smoothing.

Elaborations on $\hat{f}_c$ and $\hat{f}_p$, involving such techniques as tapering and prewhitening, have been proposed in the spectrum estimation literature. Tapering (Tukey, 1967) multiplies X_t by a sequence which decays smoothly to zero at $t = 1$ and $t = N$ in order to reduce the effect of contamination of $I(\lambda)$ from other frequencies. Prewhitening (Press and Tukey, 1956) entails fitting a preliminary AR to X_t, forming $\hat{f}_c$ or $\hat{f}_p$ from the residuals, and then multiplying by the AR transfer function. This recognizes that a quadratic spectrum estimate may not be very good at fitting a sharp peak, such as indeed appears at zero frequency in many empirical series. In fact pure AR spectrum estimation, without the kernel smoothing involved in $\hat{f}_c$ or $\hat{f}_p$, became popular, see Burg (1975), Parzen (1969); here the AR order replaces M as the bandwidth and is regarded as increasing slowly to infinity with N in the asymptotics. Mixed ARMA models have also been used. Other spectrum estimates with similar 'high-resolution' potential have been proposed, see e.g. Pisarenko (1972). These and other ideas are discussed further by Robinson (1983a) in the *Handbook of Statistics* volume on Time Series in the Frequency Domain.

Brillinger (1975, 1979) developed the approach of Jowett (1955) and Hannan (1957) to an important class of multivariate problems. Let X_t temporarily be vector-valued with mean vector μ and autocovariance matrix $\Gamma_j = \mathrm{E}(X_1 - \mu)(X_{1+j} - \mu)'$ and spectral density matrix $f(\lambda)$ given by $\Gamma_j = \int_{-\pi}^{\pi} f(\lambda) e^{ij\lambda} \, d\lambda$. Brillinger (1975) gave a CLT and covariance matrix estimate for the discrete Fourier transform

$(2\pi N)^{-1/2}\sum_{t=1}^{N} X_t e^{it\lambda}$, which reduces to one for the sample mean at $\lambda = 0$. Brillinger (1979) noted that apart from the mean correction the elements of the sample autocovariance matrix $N^{-1}\sum_{1\le t,t+j\le N}(X_t - \overline{X})(X_{t+j} - \overline{X})'$, are essentially sample means themselves, so that the preceding methods of inference extend readily to these quadratic statistics, while the delta method can then be applied to determine inference rules for such statistics as sample autocorrelations. The approach developed collectively by Jowett, Hannan and Brillinger will be seen to extend readily to other situations described in the paper such as ones considered in econometrics, and we shall refer to it as the **JHB** approach.

Brillinger (1973) discussed the problem of inference on μ in case it is the mean of a continuous time process, which is either observed continuously or at possibly unequally-spaced times that are either finite or random. He suggested splitting the observation interval $(0,N)$ into m disjoint subseries of lengths $\ell = N/m$, showed that for fixed m the sample means of each stretch are asymptotically independent and identically distributed as $N \to \infty$, and then used the sample variance of the m sub-series means to estimate the variance of the overall sample mean. This gives a studentized statistic that is asymptotically t_{m-1}, as $N \to \infty$. Of course for this approach to be useful it must be possible to choose m fairly large, at the same time as ℓ is large. Thus, Carlstein (1986) (in the context of a more general class of statistics than $\overline{X}$) developed Brillinger's approach by allowing m, as well as ℓ, to increase with N, in discrete time series. He showed that Brillinger's estimate can then be consistent for $2\pi f(0)$, and discussed the choice of ℓ and m when X_t is a first-order AR. The estimates considered by Brillinger and Carlstein are, like $\hat{f}_c(0)$ and $\hat{f}_p(0)$, quadratic functions of the data, and so the asymptotic theory of all these estimates can undoubtedly be treated in a unified way. Indeed Künsch (1989) considered a jackknifed estimate of $NV(\overline{X})$ based on overlapping subseries, and showed it to be a special case of $2\pi\hat{f}_c(0)$.

In a somewhat similar vein, Politis and Romano (1995), Bertail et al. (1995) have developed a procedure of Wu (1990), proposed for iid data and statistics that are asymptotically normal, to data that can be dependent and statistics T_N, say, that can have an arbitrary limit distribution. Politis and Romano propose to construct blocks of size ℓ with a given (possibly zero) degree of overlap, and to construct the statistic of interest (say the sample mean) from each. Then the empirical distribution across blocks of the latter, centred at T_N and suitably normalised, converges, at continuity points, to the distribution function of T_N centred at its probability limit and suitably normalised, in case of stationary α-mixing observations. Clearly this approach to constructing asymptotically valid confidence intervals has extremely wide applicability.

The problem of inference on the mean of a stationary time series has also been discussed in the operations research literature. Many estimates of $V(\overline{X})$ considered there are also quadratic forms but tend to have been motivated by considerations other than asymptotic theory. For a recent review see Song and Schmeiser (1993).

3. Inference on linear regression

The simple location model of the previous section extends to the multiple regression

$$Y_t = \beta' Z_t + X_t, \quad t = 1, 2, \ldots , \tag{3.1}$$

where X_t is, as before, covariance stationary, but it now has zero mean and is unobservable, the scalar Y_t and q-dimensional column vector Z_t being observed. Throughout the present section we presume for convenience that the spectrum $f(\lambda)$ of X_t satisfies at least the condition:

$$f(\lambda) \text{ is continuous and positive for all } \lambda \in (-\pi, \pi] \ . \tag{3.2}$$

Again the properties under (2.4) of the OLSE

$$\hat{\beta} = \left[\sum_{t=1}^{N} Z_t Z_t'\right]^{-1} \sum_{t=1}^{N} Z_t Y_t \tag{3.3}$$

of the vector β partially extend to dependent situations. In particular, Grenander (1954) gave conditions under which $\hat{\beta}$ is asymptotically as efficient as the BLUE of β, in case of nonstochastic Z_t. These conditions are satisfied under (3.2) if, for example, Z_t is a vector of polynomials or trigonometric functions in t. Even when these conditions are not satisfied $\hat{\beta}$ is often consistent and asymptotically normal, whereas a generalized least squares estimate (GLSE) of β, obtained under a misspecified parametric model for $f(\lambda)$, might be inconsistent. However, whether or not it is asymptotically efficient, the asymptotic variance of $\hat{\beta}$ is not the same as it would be under (2.4).

Consider two general specifications for Z_t, taking Z_{jt} to be Z_t's j^{th} element and $D_j = \left(\sum_{t=1}^{N} Z_{jt}^2\right)^{1/2}$.

Condition C Z_t is nonstochastic and, as $N \to \infty$,

$$D_i \to \infty, \quad i = 1, \ldots, q \ , \tag{3.4}$$

$$\lim \max_{1 \le t \le N} \frac{|Z_{it}|}{D_i} \to 0, \quad i = 1, \ldots, q \ , \tag{3.5}$$

$$\sum_{t=1}^{N-|j|} \frac{Z_{ht} Z_{i,t+j}}{D_h D_i} \to \rho_{hij}, \ h, i = 1, \ldots, q, \ j = 0, \pm 1, \ldots \ . \tag{3.6}$$

Condition D Z_t is strictly and covariance stationary and ergodic and independent of X_u for all t, u, and we put $\rho_{hij} = \mathrm{E}(Z_{ht} Z_{i,t+j}) / \{\mathrm{E}(Z_{ht}^2)\mathrm{E}(Z_{it}^2)\}^{1/2}$.

In either case we assume that $R_j = (\rho_{hij})_{q \times q}$ is positive definite for $j = 0$, and we have

$$R_j = \int_{-\pi}^{\pi} e^{ij\lambda} dF(\lambda), \ j = 0, \pm 1, \ldots \ , \tag{3.7}$$

for some F with Hermitian nonnegative definite increments, which is uniquely defined by the requirement that it be continuous from the right and $F(-\pi) = 0$. Then under similar conditions on X_t to those described in Conditions A or B,

$$D(\hat{\beta} - \beta) \to_d N(0, R_0^{-1} S R_0^{-1}), \text{ as } N \to \infty \ , \tag{3.8}$$

where $D = \text{diag}\{D_1, \ldots, D_q\}$ and

$$S = 2\pi \int_{-\pi}^{\pi} f(\lambda)\, dF(\lambda) \ . \tag{3.9}$$

Because R_0 is consistently estimated by $D^{-1} \sum_{t=1}^{N} Z_t Z_t' D^{-1}$, it remains to estimate S. Equivalently, we have the approximation

$$\hat{\beta} \sim_d N\left[\beta, \left[\sum_{t=1}^{N} Z_t Z_t'\right]^{-1} V\left[\sum_{t=1}^{N} Z_t X_t\right]\left[\sum_{t=1}^{N} Z_t Z_t'\right]^{-1}\right] \ , \tag{3.10}$$

where $V(.)$ now refers to the covariance matrix of its argument, and comparison with (3.8) indicates that

$$V\left[\sum_{t=1}^{N} Z_t X_t\right] \sim D^{1/2} S D^{1/2} = T \text{ as } N \to \infty \ . \tag{3.11}$$

Thus we require an estimate $\hat{T}$ such that

$$D^{-1/2}(\hat{T} - T) D^{-1/2} \to_p 0 \ . \tag{3.12}$$

Eicker (1967) proposed (see his equation (4.124))

$$\hat{T}_1 = N^2 \sum_{j=1-M}^{M-1} \frac{C_j \hat{c}_j}{N - |j|} \ , \tag{3.13}$$

where

$$C_j = \frac{1}{N} \sum_{1 \le t, t+j \le N} Z_t Z_{t+j}', \ \hat{c}_j = \frac{1}{N} \sum_{1 \le t, t+j \le N} \hat{X}_t \hat{X}_{t+j} \ , \tag{3.14}$$

$$\hat{X}_t = Y_t - \hat{\beta}' Z_t \ , \tag{3.15}$$

and M is analogous to the M of Section 2, increasing slowly with N in the asymptotics. He indicated that (3.12) holds for $\hat{T} = \hat{T}_1$ under reasonable conditions. $\hat{T}_1$ is suggested by applying Parseval's equality,

$$S = \sum_{j=-\infty}^{\infty} R_j \gamma_j \ , \tag{3.16}$$

and then inserting sample estimates and truncating the sum. However $\hat{T}_1$ has the disadvantage of not necessarily being positive semi-definite (psd). It is possible to obtain a guaranteed psd $\hat{T}$ by introducing suitable weights in (3.13), to correspond to weights guaranteeing a non-negative estimate of $f(0)$, see Section 2. This is best seen in the frequency-domain version

$$\hat{T}_2 = (2\pi)^2 \sum_{j=1}^{N} I_Z(\lambda_j) \hat{f}(\lambda_j) \ , \tag{3.17}$$

considered by Hannan and Robinson (1973), Robinson (1976), where $I_Z(\lambda) = (2\pi N)^{-1} \left(\sum_{t=1}^{Z} Z_t e^{it\lambda} \right) \left(\sum_{t=1}^{N} Z_t' e^{-it\lambda} \right)$, and $\hat{f}(\lambda)$ is a smoothed estimate of $f(\lambda)$. If $\hat{f} = \hat{f}_c$ or $\hat{f}_p$, with K satisfying (2.20), then $\hat{T}_2$ is psd. Unlike Eicker (1967) and Hannan (1979), who stressed Condition C, Hannan and Robinson (1973), Robinson (1976) imposed Condition D, but (3.17) is equally appropriate in either case. In the asymptotic theory of these authors, and of Eicker (1967), the bandwidth M in (3.13) and in $\hat{f}(\lambda_j)$ in (3.17), is treated as nonstochastic with $M^a/N + N/M^b \to 0$ for suitable a, b such that $1 \leq a < b$. In practice, and as described in Section 6, M may well depend on the same data $\{Y_t, Z_t, t = 1, ..., N\}$, and this was permitted by Robinson (1991a), who showed that $(C_0^{-1} \hat{T}_2 C_0^{-1})^{-1/2} (\hat{\beta} - \beta)$ has the same limiting multivariate standard normal distribution under a condition on the stochastic M of the type $M^a/N + N/M^b \to_p 0$ as it does under $M^a/N + N/M^b \to 0$ for deterministic M.

The OLSE is a special case (with $\phi(\lambda) \equiv 1$) of

$$\hat{\beta}(\phi) = \left\{ \sum_{j=1}^{N} I_Z(\lambda_j) \phi(\lambda_j) \right\}^{-1} \sum_{j=1}^{N} I_{ZY}(\lambda_j) \phi(\lambda_j) \ , \tag{3.18}$$

where $I_{ZY}(\lambda) = (2\pi N)^{-1} \left(\sum_{t=1}^{N} Z_t e^{it\lambda} \right) \left(\sum_{t=1}^{N} Y_t e^{-it\lambda} \right)$ and $\phi(\lambda)$ is a real, non-negative function. For general ϕ,

$$D\{\hat{\beta}(\phi) - \beta\} \to_d$$

$$N\left[0, \left\{ \int_{-\pi}^{\pi} \phi(\lambda) \, dF(\lambda) \right\}^{-1} 2\pi \int_{-\pi}^{\pi} f(\lambda) \phi^2(\lambda) \, dF(\lambda) \left\{ \int_{-\pi}^{\pi} \phi(\lambda) \, dF(\lambda) \right\}^{-1} \right] , \tag{3.19}$$

under Conditions A or B, Conditions C or D, and other conditions (e.g. Hannan, 1973), and also

$$\left[\left\{\frac{1}{N}\sum_{j=1}^{N}\phi(\lambda_j)I_Z(\lambda_j)\right\}^{-1}\frac{1}{N}\sum_{j=1}^{N}\hat{f}(\lambda_j)\phi^2(\lambda_j)I_Z(\lambda_j)\left\{\frac{1}{N}\sum_{j=1}^{N}\phi(\lambda_j)I_Z(\lambda_j)\right\}^{-1}\right]^{-1/2}$$
$$(\hat{\beta}(\phi)-\beta) \tag{3.20}$$

is approximately multivariate standard normally distributed, where again $\hat{f}$ is a smoothed nonparametric estimate.

The user-chosen function ϕ can achieve two distinct goals. One is to reduce bias due to misspecification of (3.1) from errors in the observation of regressors. In particular the actual model may be $Y_t = \beta' Z_t^* + X_t$ where $Z_t = Z_t^* + U_t$ is observed, and the signal-to-noise ratio may vary across frequency in a manner that can be approximately predicted. For example the spectral mass of Z_t^* might be assumed to be relatively large at low frequencies, and so if U_t is assumed to be white noise the signal-to-noise ratio may be expected to be relatively high at low frequencies, but possibly low at higher frequencies. On the other hand if seasonal noise U_t is anticipated the signal-to-noise ratio is likely to be low in the neighbourhood of the seasonal frequencies. In either case we take $\phi(\lambda) = 0$ where signal-to-noise ratio is feared to be low, thereby approximately achieving another form of 'robustness'. This idea of omission of frequencies was proposed by Hannan (1963a), and justified theoretically by Hannan and Robinson (1973), Robinson (1972). It was developed in an econometric context, and named 'band spectrum regression', by Engle (1974). Notice that there will nevertheless be 'leakage' from the omitted frequencies, and Robinson (1986) showed that this can be reduced by tapering the data.

Irrespective of whether or not we eliminate frequencies, there remains the choice of ϕ at remaining frequencies. Hannan (1963b) showed, for a modification of (3.18), that the data-dependent choice $\phi(\lambda) = \hat{f}^{-1}(\lambda)$, with $\hat{f}$ again representing a nonparametric spectrum estimate, achieves the same asymptotic efficiency as the GLSE using a parametric model for $f(\lambda)$, and

$$\left\{\sum_{j=1}^{N}\hat{f}(\lambda_j)^{-1}I_{ZZ}(\lambda_j)\right\}^{1/2}\left\{\hat{\beta}(\hat{f}^{-1})-\beta\right\} \tag{3.21}$$

is asymptotically multivariate standard normal. Robinson (1991a) showed that this remains true in case of a general data dependent bandwidth M in $\hat{f}$, and allowed also for omission of frequencies. Samarov (1987) investigated if adapting unknown correlation is necessary when there is only a small unspecified autocorrelation. He described in terms of the relation between regressors and disturbances the loss of efficiency of the OLSE when the disturbances depart from a central model (usually white noise) with additive, autocorrelated contamination. He found that the loss of efficiency is higher when the spectrum $dF(\lambda)$ of the regressors Z_t exhibits large peaks. In a minimax context, he obtained the function ϕ in (3.18) that minimizes the mean squared error of the regression estimates at

the central model subject to an upper bound on the loss of efficiency of the OLSE. That function follows f^{-1} when dF is small, but it is proportional to dF^{-1} at the frequencies where the spectrum of the regressors shows high peaks, down-weighting their contribution to the variance in (3.19).

Because of the ease with which they extend to more general problems, we shall refer to the use of convolution-type estimates such as (3.13) or (3.17) as the E approach, following Eicker's original suggestion. However, the JHB approach also applies to (3.1), because under Condition D we can estimate T by

$$\hat{T}_3 = 2\pi N \hat{f}_u(0) \ , \tag{3.22}$$

where $\hat{f}_u(\lambda)$ is an estimate of the spectral density matrix of $U_t = Z_t X_t$, computed with X_t replaced by $\hat{X}_t$ by methods analogous to those of Section 2. This approach has been stressed in the econometric literature and it has the advantage over $\hat{T}_1$ and $\hat{T}_2$ of not requiring independence, at least up to fourth moments, between Z_t and X_t, but only that X_t and Z_t are uncorrelated. On the other hand, it makes less use of the structure of the model if such independence is reasonable, when it might be expected to possess inferior finite-sample properties.

A different line of investigation was pursued by authors such as Bickel and Herzberg (1979), namely robustness of experimental design with respect to autocorrelated errors. They considered a linear model similar to (3.1), but with observations on Y and Z recorded at possibly unequally spaced points of time, where the choice of time points is the design problem. They developed an asymptotic theory under which equal-spacing retains its known optimality under independence of X_t, in a strong sense for location-estimation and in a weaker sense for slope estimation. Their approach is one of trying to minimize the variance of least squares in case of dependence, subject to a minimum of efficiency in case of independence.

4. Inference based on robust estimates

The essentially linear estimates of location and regression models of the preceding sections are generally not fully efficient asymptotically when X_t is non-Gaussian, and also can be distorted in finite samples by a few 'outliers', seemingly contaminated observations that are of much greater magnitude than are the bulk of the data. Both criticisms can be to some extent overcome by estimates that seem appropriate to heavy-tailed distributions.

One important class, which generalizes the OLSE, are M-estimates. In the location problem let μ satisfy

$$\mathrm{E}\psi(X - \mu) = 0 \tag{4.1}$$

for a random variable X and a suitable odd function ψ. Thus $\mu = \mathrm{E}X$ in case $\psi(x) = x$, and $\mu = \text{median } (X)$ in case $\psi(x) = \text{sign}(x)$. Given identically distributed observations $X_1, ..., X_N$ define $\hat{\mu}$ by

$$\sum_{t=1}^{N} \psi(X_t - \hat{\mu}) = 0 \ . \tag{4.2}$$

Huber (e.g. 1964, 1972) discussed such estimates in case of independent X_t, for example establishing that

$$N^{1/2}(\hat{\mu} - \mu) \to_d N(0, \mathrm{E}\psi^2(X_1)/\{\mathrm{E}\psi'(X_1)\}^2) \ , \tag{4.3}$$

where ψ' is the derivative of ψ. An optimal ψ is given by the score function $-\phi'(x)/\phi(x)$, where ϕ now represents the probability density function of X_1 and ϕ' is its derivative, so $\hat{\mu}$ is the MLE of μ. One can form adaptive estimates in which this ψ is replaced by a smoothed nonparametric estimate of the score function, see Stone (1975). On the other hand, Huber and others have suggested alternative ψ which are optimal in another sense or work well in terms of reducing the effects of outliers.

Now suppose X_t are serially dependent, and $\psi(X_t - \mu)$ is covariance stationary, as when X_t is strictly stationary and $\mathrm{E}\psi^2(X_1 - \mu) < \infty$. Then again from Grenander (1954), if we impose (2.4) with f now representing the spectrum of $\psi(X_t - \mu)$, we cannot improve on $\hat{\mu}$ by a weighted estimate so far as asymptotic efficiency, for given ψ, is concerned. Under conditions analogous to Conditions A or B,

$$N^{1/2}(\hat{\mu} - \mu) \to_d N\left(0, 2\pi f(0)/\{\mathrm{E}\psi'(X_1 - \mu)\}^2\right) \ , \tag{4.4}$$

and we can consistently estimate the asymptotic variance using the JHB approach, by

$$2\pi \hat{f}(0) / \left\{ \frac{1}{N} \sum_{t=1}^{N} \psi'(X_t - \hat{\mu}) \right\}^2 \ , \tag{4.5}$$

where $\hat{f}$ is as in Section 2, but based on the $\psi(X_t - \hat{\mu})$ instead of the $X_t - \overline{X}$. Parametric linear modelling of the autocorrelation of X_t (e.g. by an AR) does not in general lead to a simple models for the autocorrelation of $\psi(X_t - \mu)$, or thus to a manageable parametric f, while on the other hand a linear model for the nonlinear transformation $\psi(X_t - \mu)$ might seem contrived, so that nonparametric methods of estimating f seem especially natural here; this remark also has implications for the suitability of Condition B in the asymptotics.

M-estimates of a location parameter have been extended to incorporate a data-dependent scale factor, and to regression models, and analogous methods of inference to allow for serial dependence can be constructed, where the E method is an option in the regression case. M-estimates have also been applied to dynamic time series models, in case of supposed contamination of innovations or observations, see e.g. Beran (1976), Bustos (1981), Martin and Yohai (1986). Use of a ψ that would be optimal in case of independence, need not necessarily lead to optimality in the presence of dependence. This issue was explored by Portnoy (1977, 1978) and Zamar (1990), developing methods of Huber (1964). In case of

independence Huber had chosen ψ by a minimax approach, minimizing the asymptotic variance under the worst possible distribution. Such ψ is at least bounded. Portnoy studied the location problem in a specialized form of MA, deriving estimates which are approximately optimal in Huber's sense, and suggesting the use of piecewise linear ψ with a small amount of redescent. Portnoy derived expressions for the asymptotic variance and showed that no M-estimate can be optimal to the second order without explicitly adapting itself to the dependence. Zamar also employed a minimax approach in the location problem, with known scale. Defining a neighbourhood of contamination for outliers and a measure of serial dependence, he obtained a ψ depending on a given upper bound for the dependency, and as this increases the optimal $\psi(x)$ tends to sign(x), the sample median score function.

In addition to M-estimates, other approaches to inference have been suggested that may be relatively insensitive to outliers. For location and regression models these were developed with independent observations in mind, but we mention examples in which their properties under dependence have also been studied. Gastwirth, Rubin and Wolff (1967) showed that the sign test is no longer distribution-free. Serfling (1969) considered the Wilcoxon two-sample statistic for testing equality of two nonparametric distributions, showing that in case of α-mixing processes (cf Condition A) it is asymptotically normal and consistently estimating the limiting variance by the JHB approach. Gastwirth and Rubin (1971) looked at the effect of dependence on the levels of the t-test, the one-sample Wilcoxon test and the sign test in case of Gaussian processes and AR's with double-exponential marginals. Gastwirth and Rubin (1975) established, under α-mixing, asymptotic normality of linear functions of order statistics which include the median and trimmed mean, as well as the Hodges–Lehmann estimate, and studied relative efficiency in special cases. Koul (1977) gave CLT's for several robust estimates of the slope coefficient in a simple linear regression under α-mixing. In some of the situations studied by the above authors consistent estimation of the limiting variance, under nonparametric dependence, has not been explicitly studied.

5. Inference in econometric models

Due to the occurrence of macroeconomic and financial time series data, many econometric methods are devised with possible serial dependence in mind. In fact relatively early econometric work stressed the efficiency gains due to the GLSE in regression models in the presence of autocorrelated errors, see e.g. Cochrane and Orcutt (1949). This interest has continued, and more recently it has been fashionable to employ point estimates which may well be inefficient, but studentize them to allow for serial dependence. A feature of much econometric work is the relative complexity of modelling, often involving nonlinearity and multivariate data, for example. Some similar models have been used in non-economic applications but we have chosen to categorize as "econometric models" ones which are

more complicated than the simple location and linear regression models treated so far. The purpose of the present section is to discuss some issues that arise when we move to a more complex set-up.

Many important problems are covered by the model

$$Y_t = G_t(\theta) + X_t, \ t = 1, 2, \ldots \ , \tag{5.1}$$

where now Y_t and X_t are $p \times 1$ vectors and $G_t(\theta)$ is a $p \times 1$ vector of possibly nonlinear functions of an unknown $s \times 1$ vector θ, and of observable stochastic or nonstochastic explanatory variables to which explicit reference is suppressed. Again X_t is unobservable. Sometimes $G_t(\theta)$ is linear in observables,

$$G_t(\theta) = \mu(\theta) + \sum_{j=-\infty}^{\infty} A_j(\theta) Z_{t-j} \ , \tag{5.2}$$

where the Z_t are the $q \times 1$ explanatory variables, the $A_j(\theta)$ are $p \times q$ matrix functions of θ, and $\mu(\theta)$ is a vector. The static linear simultaneous equations model of econometrics then arises when $A_j(\theta) \equiv 0$, $j \neq 0$, $A_j(\theta) = \Gamma^{-1}B$, where B and Γ are matrices that are possibly, but not necessarily, linear in θ. "Distributed lag models" are also included, and are of interest even in the scalar case $p = 1$, when, for example, $A_j(\theta) = 0$, $j < 0$; $= \theta_1 \theta_2^j$ for $j \geq 0$ and scalar parameters θ_1, θ_2 with $|\theta_2| < 1$. Also included in (5.1) are possibly static scalar or multivariate nonlinear regression models such as when

$$G_t(\theta) = G(\theta; Z_t) \ , \tag{5.3}$$

for a given function G.

Suppose first that the processes $\{G_t(\theta)\}$ and $\{X_t\}$ are independent of each other and stationary. A minimum-distance or Gaussian estimate of θ is given by

$$\hat{\theta} = \arg\min_{h \in \theta} \sum_{j=1}^{N} \| \Phi(\lambda_j)^{1/2} \{ w_Y(\lambda_j) - w_G(\lambda_j; h) \} \|^2 \ , \tag{5.4}$$

where $\|.\|$ means Euclidean norm, $\Phi(\lambda)$ is a $p \times p$ psd Hermitian matrix, and

$$w_Y(\lambda) = \frac{1}{(2\pi N)^{1/2}} \sum_{t=1}^{N} Y_t e^{it\lambda}, \ w_G(\lambda; h) = \frac{1}{(2\pi N)^{1/2}} \sum_{t=1}^{N} G_t(h) e^{it\lambda} \ , \tag{5.5}$$

and Θ is a compact subset of $\mathbb{R}^S$. In case (5.2) leads to G including some unobservable Z_t (because $Z_1, \ldots, Z_N$ only are observed), we can consider instead

$$\hat{\theta} = \arg\min_{h \in \theta} \sum_{j=1}^{N} \| \Phi(\lambda_j)^{1/2} \{ w_Y(\lambda_j) - A(\lambda_j; h) w_Z(\lambda_j) \} \|^2 \ , \tag{5.6}$$

$$w_Z(\lambda) = \frac{1}{(2\pi N)^{1/2}} \sum_{t=1}^{N} Z_t e^{it\lambda}, \ A(\lambda; h) = \sum_{-\infty}^{\infty} A_j(h) e^{ij\lambda} \ . \tag{5.7}$$

Such estimates were considered by Hannan (1971), Robinson (1972), and cover many cases. Under regularity conditions given in these references,

$$N^{1/2}(\hat{\theta}-\theta) \to_d N(0, \Psi(\Phi)^{-1}\Psi(\Phi f\Phi)\Psi(\Phi)^{-1}) \ , \tag{5.8}$$

where

$$\Psi(\Phi) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\partial}{\partial h_a \partial h_b'} \mathrm{tr}\{\Phi(\lambda)dF(-\lambda; h_a, h_b)\}|_{h_a=h_b=\theta} \ , \tag{5.9}$$

$\Psi(\Phi f\Phi)$ is the same quantity with $\Phi(\lambda)$ replaced by $\Phi(\lambda)f(\lambda)\Phi(\lambda)$, and F satisfies

$$\mathrm{E}\{G_t(h_a)G_{t+j}'(h_b)\} = \int_{-\pi}^{\pi} e^{ij\lambda}dF(\lambda; h_a, h_b) \ . \tag{5.10}$$

We can now apply the E method. Define

$$\hat{\Psi}(\Phi) = \frac{1}{N}\sum_{j=1}^{N} W(-\lambda_j; \hat{\theta})\Phi(\lambda_j)W(\lambda_j; \hat{\theta})' \ , \tag{5.11}$$

$$W(\lambda;\theta) = \frac{\partial}{\partial\theta'}w_G(\lambda;\theta) \text{ or } W(\lambda;\theta) = \frac{\partial}{\partial\theta'}A(\lambda;\theta)w_Z(\lambda) \ . \tag{5.12}$$

Let $\hat{\Psi}(\Phi\hat{f}\Phi)$ be defined analogously, with $\hat{f}(\lambda)$ an estimate of $f(\lambda)$ as in Section 2, based on residuals. Then we treat $\hat{\theta}$ as approximately

$$N\left(\theta, \frac{1}{N}\hat{\Psi}(\Phi)^{-1}\hat{\Psi}(\Phi\hat{f}\Phi)\hat{\Psi}(\Phi)^{-1}\right) \ . \tag{5.13}$$

This approach was suggested by Hannan and Robinson (1973), Robinson (1976) in special cases where G_t is given by (5.2). Robinson (1991a) justified it in case of (5.2), allowing the bandwidth parameter used in a nonparametric $\hat{f}$ to be a quite general function of the same data. The choice $\Phi = I_p$, the $p \times p$ identity matrix, corresponds to nonlinear least squares, while $\Phi = \hat{f}^{-1}$ in (5.4) or (5.6) produces the most efficient estimate within the above class, and thus we have an extension of the approach of Hannan (1967) to a much more general model, and again Robinson (1991a) justified use of a general data-dependent bandwidth in $\hat{f}$.

The JHB method can alternatively be used. For $\Phi \equiv I_p$, $\hat{\theta}$ can be neatly written in time domain form, and when $p = 1$

$$N^{1/2}(\hat{\theta}-\theta) \to_d N(0, A^{-1}BA^{-1}) \ , \tag{5.14}$$

where $A = \mathrm{E}(\partial G_t(\theta)/\partial\theta \partial G_t(\theta)/\partial\theta')$, $B = 2\pi f_u(0)$, where $f_u(\lambda)$ is the spectral density matrix of $u_t = X_t\partial G_t(\theta)/\partial\theta$. Thus the JHB approach can be applied much as it was to linear regression in Section 3. This was advocated by Domowitz and White (1982), though they suggested using the truncated version of $\hat{f}_c(0)$ given by (2.21), whose properties were mentioned in Section 2.

The JHB approach has been stressed in the bulk of subsequent relevant econometric literature, in which special names have been invented for the topic, such as "autocorrelation-consistent variance estimation", "heteroskedasticity and autocorrelation-consistent variance estimation" and "long run variance estimation". Let a vector parameter satisfy the equation

$$\mathrm{E}\{U_t(\theta)\} = 0, \quad t = 1, 2, \ldots , \tag{5.15}$$

where the dimension of U_t is at least as great as that of θ and may also depend on other unknown parametric and/or nonparametric functions. In particular (5.15) embodies the econometric model, or the principal part of it. Given a suitable matrix $\hat{S}$ of full row rank equal to the dimension of θ and converging in probability to a matrix S of the same rank, and proxies $\hat{U}_t(\theta)$ for $U_t(\theta)$, involving estimates of any nuisance parameters/functions, and possibly each depending on all the data, we estimate θ by

$$\hat{\theta} = \arg\min_{h \in \theta} \sum_{t=1}^{N} \hat{U}_t'(h) \hat{S}' \hat{S} \sum_{t=1}^{N} \hat{U}_t(h) \tag{5.16}$$

for a compact set Θ. Under regularity conditions

$$N^{1/2}(\hat{\theta} - \theta) \to_d N(0, C^{-1} D C^{-1'}) , \tag{5.17}$$

where

$$C = SE\left[\frac{\partial U_t(\theta)}{\partial \theta'}\right], \quad D = S 2\pi f_U(0) S' , \tag{5.18}$$

and $f_U(\lambda)$ is the spectral density matrix of $U_t(\theta)$ and can be estimated as in Section 2, from the $U_t(\hat{\theta})$, while C can be estimated by a sample average. The generalized method-of-moments (GMM) estimates of Hansen (1982) fall into this scheme, and he indicated the role of nonparametric spectrum estimation in estimating D, and also that an optimal S is $\mathrm{E}(\partial/\partial\theta U_t(\theta)')\{2\pi f_U(0)\}^{-1}$ so that $C^{-1}DC^{-1'} = [\mathrm{E}(\partial/\partial\theta)U_t(\theta)'\{2\pi f_U(0)\}^{-1}\mathrm{E}(\partial/\partial\theta')U_t(\theta)]^{-1}$, which can be estimated similarly. In Hansen's case the statistic $\hat{\theta}$, though in general only implicitly-defined, entails $U_t(\theta)$ that are known functions of unobservables and θ. Robinson (1989) considered $U_t(\theta)$ which are defined in closed form but are functions of conditional expectations of derivatives of nonparametric regression functions (cf Section 9 below), so that the $\hat{U}_t(\theta)$ involve smoothed nonparametric estimates of these, and gave conditions under which (5.17) holds and D can be consistently estimated by weighted autocovariance estimates. In such circumstances it seems hard to motivate parametric models for the autocorrelation in $U_t(\theta)$, and a nonparametric approach seems natural.

Particular cases of, or modifications of, the various spectral estimation methods of Section 2, have been stressed in the econometric literature. In fact Levine (1983) employed $\hat{f}_c$ with (2.20) but with c_j replaced by $Nc_j/(N - |j|)$, giving even more weight to the longer lags than the truncated estimate, and relating to the undesirable practice of approximating a Fourier series by its partial sum. Newey

and West (1987) advocated the modified Bartlett k (2.22), because of its simplicity and the psd property it bestows on the spectrum estimate. Andrews and Monahan (1992) employed a version of prewhitening, prior to weighted autocovariance estimation, see again Section 2. A number of other contributions focus on alternative estimates and on improvements to rates of convergence and regularity conditions, e.g. Andrews (1991), Gallant and White (1988), Hansen (1992). The dependence conditions stressed in the econometric literature have been mixing conditions, with rate conditions on the mixing numbers. These are attractive because the $U_t(\theta)$ may be complicated, nonlinear functions of underlying variables, but inherit the mixing properties of these; this is not the case if the underlying variables are linear filters of white noise, for example, cf Section 4. The fact that the mixing conditions readily allow a degree of non-trending nonstationarity has also been stressed in the econometric literature, indeed the expectations A, B, C and D can be replaced by limits, indicative of forms of asymptotic stationarity as in Parzen (1962a).

6. Bandwidth selection

The nonparametric methods that have mostly been used in statistics to allow for serial dependence are appealing because they are justifiable in the absence of precise assumptions about the dependence, which can be very hard to motivate. However, the distinction between 'nonparametric' and 'parametric' estimation resides principally in their interpretation in large samples, and in both cases a particular functional form has to be chosen, indeed the same one may be used in either case, and the outcome of inferences will be dependent, possibly greatly so, on the choice of functional form. For example, $\hat{f}_c$ and $\hat{f}_p$ depend on K and M, AR spectrum estimates depend on AR order, and prewhitened quadratic estimates depend on all these quantities. The preceding sections include some discussion of the merits of spectrum estimates, and various K, so in the present section we focus on choice of bandwidth, in particular lag number M.

Consider the choice of M in $\hat{f}_p$ or $\hat{f}_c$ in the simple scalar setting of Section 2. Because bias tends to vary inversely with M, and variance tends to vary directly, minimization of mean squared error (MSE) $\mathrm{E}\{\hat{f}(\lambda) - f(\lambda)\}^2$ was proposed by Grenander and Rosenblatt (1957) as a simple criterion for producing a balancing M. Under regularity conditions

$$\lim_{N\to\infty} N^{2\omega}\mathrm{E}\{\hat{f}(\lambda) - f(\lambda)\}^2 = \frac{f^2(\lambda)}{b}\int_{-\infty}^{\infty} k^2(x)dx\{1 + I(\lambda = 0)\} + b^{2\omega}\{\kappa^{(\omega)} f^{(\omega)}(\lambda)\}^2 \ , \tag{6.1}$$

$$b = \lim_{N\to\infty} \frac{N^{1/(1+2\omega)}}{M} \ , \tag{6.2}$$

$$\kappa^{(\omega)} = \lim_{x \to 0} \frac{1 - k(x)}{x^{\omega}} \quad , \tag{6.3}$$

$$f^{(\omega)}(\lambda) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |j|^{\omega} \gamma_j e^{-ij\omega} \quad , \tag{6.4}$$

and ω is the largest real number for which both (6.3) and (6.4) are assumed finite. Often $\omega = 2$, though in the modified Bartlett case (2.22), $\omega = 1$. In view of (6.2) we should choose M proportional to $N^{1/(1+2\omega)}$, and minimizing (6.1) with respect to b gives the optimal choice of M,

$$M^* = N^{1/(1+2\omega)} / b^* \quad , \tag{6.5}$$

where

$$b^* = \left[\frac{f^2(\lambda) \int k^2(x) dx (1 + I(\lambda = 0))}{2\omega \{\kappa^{(\omega)} f^{(\omega)}(\lambda)\}^2} \right]^{1/(1+2\omega)} \quad . \tag{6.6}$$

In practice, though ω can reasonably be picked as the largest value such that (6.3) is finite (i.e. trusting that (6.4) is finite) there is a strong element of circularity in that b^* depends on $f(\lambda)$ itself, and also on $f^{(\omega)}(\lambda)$. This problem is standard in the smoothed estimation of nonparametric functions, and a standard proposal to deal with it is to replace f and $f^{(\omega)}$ by 'pilot' estimates based on either a simple parametric model or on an initial choice of bandwidth, in the hope that M^* will not be too sensitive to the design of the pilot estimates, though it can be very hard to accurately estimate b^*. Andrews (1991) has developed this approach in case of pilot AR spectrum estimates, and also showed that the eventual spectrum estimates are still consistent in the presence of the data-dependent M. Newey and West (1994) justified the optimality of methods using an initial choice of bandwidth.

The preceding discussion is relevant to the JHB approach, in which f is to be estimated at a single, zero, frequency only, and extends readily to vector cases. In the E approach, say in the linear regression context of Section 3, f has to be estimated across the Nyquist band $(-\pi, \pi]$. It would be possible to determine suitable M as in the previous paragraph, in a frequency-dependent way. A simpler approach is to obtain a single M which reflects characteristics of the data across all frequencies. An informal procedure, called "window closing", entails estimating f across $(-\pi, \pi]$ for various M and choosing the one which suggests the best mix of resolution and stability. To seek an optimal choice, consider the integrated MSE

$$\int_{-\pi}^{\pi} \mathrm{E}\{\hat{f}(\lambda) - f(\lambda)\}^2 \chi(\lambda) d\lambda \quad , \tag{6.7}$$

for a weight function $\chi(\lambda)$. Lomnicki and Zaremba (1957) suggested $\chi(\lambda) \equiv 1$, and Jenkins and Watts (1968) suggested $\chi(\lambda) = 1/f(\lambda)^2$, because $\hat{f}(\lambda)$ has asymptotic variance proportional to $f(\lambda)^2$. We minimize (6.7) asymptotically by

$$M^* = \left\{ 2\omega N \kappa^{(\omega)2} \int_{-\pi}^{\pi} f^{(\omega)}(\lambda)^2 \chi(\lambda) d\lambda \Big/ \int_{-\pi}^{\pi} f(\lambda)^2 \chi(\lambda) d\lambda \int_{-\infty}^{\infty} k^2(x) dx \right\}^{1/(2\omega+1)} . \tag{6.8}$$

Again this can be estimated using pilot estimates of f and $f^{(\omega)}$. Pickands (1970) proposed a related partly automatic method. Hurvich (1985), Beltrao and Bloomfield (1987) and Robinson (1991a) considered a fully automatic cross validation method. One version of this is as follows. Introduce the leave-two-out spectrum estimate

$$\hat{f}_{pj}(\lambda_j) = \frac{2\pi}{N} \sum_{\substack{l=1 \\ l \neq j, N-j}} K_M(\lambda_j - \lambda_l) I(\lambda_l) , \tag{6.9}$$

cf (2.18), and the pseudo log-likelihood criterion

$$\sum_{j=1}^{N} \{ \log \hat{f}_{pj}(\lambda_j) + I(\lambda_j) / \hat{f}_{pj}(\lambda_j) \} .$$

Then M minimizing this is consistent for (6.5) with $\chi(\lambda) = 1/f(\lambda)^2$. An alternative approach for choosing M, based on stochastic complexity, was proposed by Hannan and Rissanen (1988), but they gave no theoretical justification for it.

There is a much larger literature on choice of AR order, for use in AR spectrum estimates, particularly by penalized likelihood methods such as AIC. Choice of order in ARMA models has also been studied. Because AR spectrum estimates are less often used in the JHB or E approaches than $\hat{f}_c$ or $\hat{f}_p$, and because an adequate discussion of this well developed field would take up a good deal of space, we refer the reader to the survey article of Hannan (1987).

7. Higher-order asymptotics and the bootstrap

In this section we discuss for simplicity only inference based on the scalar sample mean, following Section 2. The statistic (2.12) was proposed in view of its asymptotic standard normal distribution. However this may well be a poor approximation to the actual distribution in finite samples, indeed there is evidence that serial dependence can slow convergence. In principle, representations of the exact distribution of (2.12) can be obtained, at least under the presumption of Gaussianity of X_t, but these are likely to be complicated and difficult to use in practice. A compromise may be provided by higher-order asymptotics.

One interesting question which readily lends itself to higher-order asymptotic study is the cost of correcting for autocorrelation in estimating $V(\overline{X})$ when none exists. This is a special case of a more general problem, that of over-specifying M in $\hat{f}_c$ or $\hat{f}_p$ relative to an actual MA order less than M, or overspecifying the AR order. Albers (1978) considered the case when the prescribed MA or AR order is

fixed relative to N, and (2.4) holds. He found that while there is no asymptotic loss of power of (2.12) relative to the ordinary t-ratio, the deficiency measure of Hodges and Lehmann (1970) – the difference between the numbers of observations required to achieve the same power – is non-zero to order $O(N^{-1})$, and he estimated the deficiency.

Edgeworth expansions of probability distribution functions and densities have been extensively studied in many contexts. Götze and Hipp (1983) developed a valid Edgeworth expansion for $N^{1/2}(\overline{X} - \mu)$ in case of fairly generally weakly dependent X_t. They showed that the distribution function of $N^{1/2}(\overline{X} - \mu)$ is uniformly approximated, to order $o(N^{-(s-2)/2})$, by

$$\Psi_{ns} = \sum_{r=0}^{s-2} N^{-r/2} p_r \ , \tag{7.1}$$

where under stationarity p_0 is the $N(0, 2\pi f(0))$ distribution function and for $r \geq 1$ p_r is a finite signed measure with p_0-density q_r, where q_r is a polynomial with coefficients uniquely determined by the moments of X_t up to order $r+2$. In fact the q_r can be written in terms of higher-order spectra which can in principle be estimated in a manner which generalizes ones for $f(\lambda)$. The form of (7.1) simplifies in case X_t is Gaussian, and here Taniguchi (e.g. 1984, 1987) discussed the problem in detail, and for other important statistics besides $\overline{X}$. Götze and Hipp (1994) have recently carried out extensions for the non-Gaussian case.

Even in the Gaussian case higher-order theory becomes more difficult in case of the studentized statistic (2.12). In general even the goodness of the usual normal approximation (represented, for example, by a Berry–Esseen bound) will depend on the rate of increase of the bandwidth number M. Bentkus et al. (1994) obtained Berry–Esseen bounds for the studentized sample mean and other statistics, assuming exponentially decaying autocorrelation. Götze and Künsch (1995), Lahiri (1991) established the validity of a one-term Edgeworth expansion (i.e. $s = 3$ in (7.1)) under conditions of a similar character to Götze and Hipp's. One needs also the validity of the estimated Edgeworth expansion, where the unknown features of the $r = 1$ term are estimated from the data, though even the one-term approximation is liable to be complicated.

An alternative means of improving on the normal approximation in many statistics, in particular of matching the one-term Edgeworth expansion, is the bootstrap. The bootstrap as originally developed for independent observations, by Efron (1979), does not work under dependence, but valid extensions to dependent data have been developed. Künsch (1989), Liu and Singh (1992) proposed the following. Let ℓ and m be integers such that $N = \ell m$. Consider blocks $x_t = (X_t, ..., X_{t+\ell-1})$, $t = 1, ..., N - \ell + 1$. Draw $x_1^*, ..., x_m^*$ randomly with replacement from $x_1, ..., x_{N-\ell+1}$, and denote by $X^*_{\ell(t-1)+i}$ the i-$^{\text{th}}$ element of x_t^*, $i = 1, ..., \ell$, $t = 1, ..., m$. The bootstrapped sample mean is

$$\overline{X}^* = \frac{1}{N} \sum_{t=1}^{N} X_t^* \ . \tag{7.2}$$

To studentize $\overline{X}^*$, note that $x_1^*, ..., x_m^*$ are independent so a bootstrap γ_j estimate is

$$c_j^* = \frac{1}{N} \sum_{t=1}^{m} \sum_{i=1}^{l-j} X^*_{(t-1)l+i} X^*_{(t-1)l+i+j} \ . \tag{7.3}$$

Write

$$V^* = \sum_{j=-M}^{M} k\left[\frac{j}{M}\right] c_j^* \ , \tag{7.4}$$

which by comparison with (2.16) is a bootstrap version of $2\pi f(0)$, with bandwidth M and lag window k. Then the bootstrap distribution for $\overline{X}$ is the distribution of

$$\frac{n^{1/2}(\overline{X}^* - \overline{X})}{V^{*1/2}} \tag{7.5}$$

conditional on $X_1, X_2, ..., X_n$. Moreover, under suitable conditions (see Götze and Künsch (1995), Lahiri (1991)) this distribution differs from that of (2.12) by $o(N^{-1/2})$, so that the bootstrap approximation is as good as the one-term Edgeworth approximation.

There have been numerous other recent contributions concerning the bootstrap for dependent variables, including modifications of the above procedure (e.g. Politis and Romano, 1994), data-driven choice of ℓ (Bühlmann and Künsch, 1994), extensions to the problem of M-estimation (Lahiri, 1994) and spectral density estimation (e.g. Franke and Härdle, 1992, Politis and Romano, 1992), as well as versions of the bootstrap based on parametric assumptions about the autocorrelation (e.g. Freedman, 1984, Bose, 1988).

8. Inference under long range dependence

The methods of Section 2 rely crucially on assumption (2.6), and other previous sections take similar assumptions for granted. However, there has been considerable interest in the possibility that (2.6) does not hold, with the main concern that $f(0) = \infty$, though $f(0) = 0$ has also been considered. Both possibilities are covered by the model

$$f(\lambda) \sim G\lambda^{1-2H}, \text{ as } \lambda \to 0^+, \quad 0 < H < 1 \ , \tag{8.1}$$

for $0 < G < \infty$. The case $H = 1/2$ includes (2.6). When $1/2 < H < 1$, the case of most interest, we say, there is 'long range dependence'. When $0 < H < 1/2$ there is 'anti-persistence'. Thus (2.6) seems a rather specialised assumption. Condition (8.1) is closely related to

$$\gamma_j \sim g j^{2H-2}, \text{ as } j \to \infty \ , \tag{8.2}$$

where

$$g = 2G\Gamma(2-2H)\cos \pi H \ , \tag{8.3}$$

so γ_j is not summable when $H > 1/2$. Examples of (8.1) and (8.2) are given by fractional ARIMA models, in which

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left|1 - e^{i\lambda}\right|^{1-2H}\left|\frac{a(e^{i\lambda})}{b(e^{i\lambda})}\right|^2, \quad -\pi < \lambda \leq \pi \ , \tag{8.4}$$

where $\sigma^2 > 0$ and a and b are polynomials of finite degree having no zeros in or on the unit circle, and the fractional noise model where

$$\gamma_j = \frac{\gamma_0}{2}\left(|j+1|^{2H} - 2|j|^{2H} + |j-1|^{2H}\right), \ j = \pm 1, \ldots \tag{8.5}$$

Consider a covariance stationary sequence $\{X_t\}$, with mean μ and spectrum and autocovariance satisfying (8.4) and (8.5). Now $\overline{X}$ is no longer asymptotically a BLUE of μ when $H \neq 1/2$, see Adenstedt (1979). Samarov and Taqqu (1988) found that for $0 < H < 1/2$ the efficiency can be poor, though for $1/2 < H < 1$ it is at least 0.98. In any case $\overline{X}$ is still unbiased and consistent for μ and is a computationally simple candidate for use in inference, especially in the Gaussian case.

Consider now the statistic on the left side of (2.8). In case of (8.2) or (8.3) the standard normal limit in (2.8) may or may not hold, but in any case neither Conditions *A* nor *B* is appropriate when $1/2 < H < 1$. For example the summability conditions on the mixing numbers α_j and the MA weights β_j each imply f is bounded, violating (8.1). Instead, consider the following alternative conditions:

Condition E $X_t = P(V_t)$, where V_t is a stationary Gaussian sequence with autocovariance

$$\delta_j \sim g' j^{2h-2}, \text{ as } j \to \infty, \ h < 1 \ . \tag{8.6}$$

for $g' > 0$, and

$$m = \min_{i \geq 0}\{i : J(i) \neq 0\} \ , \tag{8.7}$$

where $J(i) = E(X_t H_i'(X_t))$, and $H_i'(x) = (-1)^i e^{x^2/2}(d^i/dx^i)e^{-x^2/2}$.

Condition F X_t is as in Condition *B*, but with the restriction $\sum_{-\infty}^{\infty}|\beta_j| < \infty$ relaxed to $\sum_{j=-\infty}^{\infty}\beta_j^2 < \infty$, and $\sum_{j=-\infty}^{\infty}(\beta_{j-1} + \ldots + \beta_{j-n})^2 \to \infty$ as $n \to \infty$.

The number m in Condition *E* is called the Hermite rank of P, and was introduced by Taqqu (1975), who showed that it implies (8.2) for $1/2 < H < 1$ when $h = 1 + (H-1)/m$ satisfies $1 - 1/2m < h < 1$, and that (2.8) holds if and only if $m = 1$, the limit distribution being a nonstandard functional of Brownian motion for $m > 1$. Condition *F*, due to Eicker (1967), Ibragimov and Linnik (1971), Hannan (1979), avoids the Gaussianity component, but corresponds only to $m = 1$ in Condition *E* in that it implies (2.8).

No explicit direction has yet been given as to how to conduct inference using the non-normal limit distribution under Condition E for $m > 1$, so we focus on studentizing in (2.8). Under (8.2)

$$V(\overline{X}) \sim \frac{gN^{2(H-1)}}{H(2H-1)} \ , \tag{8.8}$$

so $\overline{X}$ is less than $N^{1/2}$-consistent when $1/2 < H < 1$. (8.2) holds for all $H \in (0,1)$ where H, as well as G, is likely to be unknown in practice. The studentization in (2.12) will not produce consistent interval estimates or asymptotically valid hypothesis tests if $H \neq 1/2$. Given (2.8) and (8.8), Robinson (1994a) observed that

$$N^{1-\hat{H}} \left\{ \frac{\hat{H}(2\hat{H}-1)}{2\hat{G}\Gamma(2-2\hat{H})\cos(\pi\hat{H})} \right\}^{1/2} (\overline{X} - \mu) \to_d N(0,1) \tag{8.9}$$

if the estimates $\hat{G}$, $\hat{H}$ satisfy

$$\hat{G} \to_p G, \quad (\log N)(\hat{H} - H) \to_p 0 \ . \tag{8.10}$$

Robinson (1994b, 1995a, 1995b) has verified these properties for three different types of estimate, all of which are based only on local assumptions on $f(\lambda)$ near $\lambda = 0$, and which do not require parameterization of f across all frequencies. A correctly specified parametric f or γ_j (as in (8.4) or (8.5)) can yield $N^{1/2}$-consistent estimates (see e.g. Fox and Taqqu, 1986), and Beran (1989) considered studentization based on such estimates, with an alternative type of approximate distribution to that arising in (8.9). However, if f is incorrectly specified, such estimates will be inconsistent, indicating a cost to modelling high frequency behaviour in a situation in which only low frequency behaviour is of real interest. In case of the estimates in Robinson (1995a, 1995b) the desired properties were established for any $H \in (0,1)$, so that prior knowledge that $f(0)$ is infinite, finite/positive, or zero, is not needed, thereby considerably generalizating the "$H = 1/2$" inference discussed in Section 2 (in case $H = 1/2$, G corresponds to $f(0)$).

Consider now, briefly, the case of regression estimation, in model (3.1) with long range dependent X_t. Suppose we assume that Z_t is nonstochastic, for example that it satisfies Condition C. Here, the OLSE is not asymptotically efficient when $H \neq 1/2$, but Eicker (1967) justified the approximation (3.10) under a somewhat strengthened version of Condition F, which allows for (8.1) or (8.2) for any $H \in (0,1)$. However he discussed studentization only under (3.2), which can apply only when $H = 1/2$. Yajima (1988, 1991) discussed the structure of the asymptotic variance of $\hat{\beta}$ for polynomial-in-t and other Z_t, and Robinson (1994b) explicitly justified a suitable studentization in the polynomial-in-t case. Dahlhaus (1992) established the asymptotic normality of a studentized GLSE in case of Gaussian X_t. Now suppose that Z_t is stochastic, in particular that it satisfies Condition D. Then the implications for the OLSE of slope coefficients are rather different, if there is an intercept in the regression. As noted by Robinson (1994a)

the process $(Z_t - EZ_t)X_t$ can have a bounded spectral density matrix $f_u(\lambda)$ at $\lambda = 0$ even if both X_t and Z_t exhibit long range dependence, that is collectively not too strong, in particular when $f_u(\lambda)$ increases at rate λ^{1-2L} as $\lambda \to 0+$, and $H + L < 3/2$. Thus the JHB approach can be used as described near the end of Section 3 (see (3.22)), though existing proofs of consistency of spectrum estimates, for example under the mixing conditions stressed in the econometric literature (see Section 5), do not apply. On the other hand, if $H + L \geq 3/2$ the limit distribution of (3.3) is non-normal. This problem does not arise for the GLSE. Robinson and Hidalgo (1997) established (3.19) allowing for an arbitrary degree of long range dependence in X_t and Z_t, for the class of ϕ including $\phi = f^{-1}$, and went on to justify feasible estimates in case of a parametric f. Künsch et al. (1993) discussed the effect of long range dependence in errors on standard independence-based inference rules in the context of certain experimental designs.

M-estimates of location and linear regression models with long-range dependent errors X_t have been considered by Beran (1986), Koul (1992). In case of independent X_t, M-estimates will typically be less efficient than the OLSE when X_t is Gaussian. However under long range dependence there is no efficiency loss, because the M-estimate will be dominated by a linear term. Moreover, when $m = 1$ in Condition E but X_t is non-Gaussian, the OLSE can be less efficient than some M-estimates. Inference in nonlinear models, including nonlinear regressions, (5.1), with long range dependent errors, has been discussed by Robinson (1994a), Robinson and Hidalgo (1997).

Long range dependence can be much more broadly defined to include also forms of nonstationarity. There is any number of such forms that might be studied, including the I(d) class, for real $d \geq 1/2$, where the d-th differences have spectrum satisfying (2.6). (For noninteger d, d-$^{\text{th}}$ differences can be defined via the binomial theorem and a suitable start-up condition.) The special case $d = 1$, unit root behaviour, has been greatly stressed of late, especially in the econometric literature. Dickey and Fuller (1979) derived a representation for the limiting distribution of the OLSE of a first order AR coefficient when the true process is a random walk, and tabulated the distribution. Phillips (1987) found the limiting distribution when the errors X_t in the unit root AR are autocorrelated but satisfy (2.6). This limiting distribution differs from Dickey and Fuller's in a way that depends simply on $f(0)$, such that Dickey and Fuller's tables can still be used with a minor modification and an estimate of $f(0)$ of the type discussed in Section 2, as Phillips explicitly proposed. There have been many elaborations on this theme in the econometric literature.

9. Inference on nonparametric probability density and regression functions

So far the prospect of serial dependence has considerably changed and complicated inference, by comparison with independence-based rules. There is an important class of problems where the latter rules continue to apply at least in case of weak dependence, and so we discuss this in a section to itself.

Let X_t now be strictly stationary with marginal probability density $p(x)$. It is desired to estimate $p(x)$ without assuming that its functional form is known up to finitely many parameters. There are many nonparametric estimates of $p(x)$, but they have similar asymptotic properties, and we shall consider only kernel estimates. Let

$$\hat{p}(x) = \frac{1}{Na} \sum_{t=1}^{N} k\left[\frac{x - X_t}{a}\right] , \tag{9.1}$$

where the given function k, called the kernel, integrates to one, and the given positive scalar a, called the bandwidth, is regarded as converging to zero as $N \to \infty$, while $aN \to \infty$.

In case the X_t are independent, asymptotic properties of $\hat{p}(x)$ were studied by Parzen (1962b) and others. In particular, under suitable additional conditions, for fixed points $x_1, ..., x_s$, as $N \to \infty$

$$(Na)^{1/2}\{\hat{p}(x_j) - p(x_j)\} \to_d \text{NID}\left[0, \int_{-\infty}^{\infty} k_2(u)du\, p(x_j)\right], \; j = 1, ..., s \; , \tag{9.2}$$

so that the $\hat{p}(x_j)$ have a simple approximate distribution and are asymptotically independent. From experience with many statistics (such as those discussed previously in the paper) it is no surprise that $\hat{p}(x)$ will still be consistent and asymptotically normal when independence of X_t is significantly relaxed to allow for various forms of weak dependence. However, the asymptotic variance, and the asymptotic independence, in (9.2) also continue to hold, as shown by Roussas (1969) and Rosenblatt (1970) in case of Markov X_t, by Robinson (1983b) in case of α-mixing X_t, and by subsequent authors under various other conditions. Notice that the finite sample distribution of $\hat{p}(x)$ will certainly be affected by the dependence, for example if k has compact support then the $\hat{p}(x_j)$ will, for a sufficiently small, be independent for finite N, if the X_t are independent, but not if they are dependent, unless the dependence eventually dies out completely. Thus (9.2) can perhaps be taken less seriously as a finite-sample approximation when there is dependence. Nevertheless for suitably large N (9.2) is an attractive source of simple confidence regions; we treat the

$$\left\{Na/\int_{-\infty}^{\infty} k^2(u)du\, p(x_j)\right\}^{1/2} \{\hat{p}(x_j) - p(x_j)\}, \; j = 1, ..., s \; , \tag{9.3}$$

as approximately independent standard normal variates.

The level of the bandwidth a significantly affects $\hat{p}(x)$. As with the bandwidth in spectral estimation (see Section 6), there are rules for optimally choosing it. It turns out that the a that asymptotically minimize MSE criteria in case of independence continue to be valid under weak dependence, see Prakasa Rao (1978),

Ahmad (1982). Moreover automatic data-dependent methods under independence can carry through to weak dependence.

Long range dependence, as stressed in the previous section, can however lead to considerably different asymptotic distributional behaviour from that under independence. In the simple case of Gaussian X_t, Robinson (1991b) showed that while $\hat{p}$ may still be asymptotically normal, the rate of convergence is affected, as are the variances in the limiting normal distribution, and far from being asymptotically independent across fixed points x_j, they are perfectly correlated. Robinson's analysis disallowed any of the x_j from equalling EX_t, but here Cheng and Robinson (1991) found the asymptotic distribution to be a nonstandard functional of Brownian motion; they also generalized Robinson's results in other directions. Hall and Hart (1990), Robinson (1991b) also looked at the MSE of $\hat{p}(x)$ and found that for suitably strong dependence the optimal a can be affected.

The density estimate $\hat{p}(x)$ can be extended to estimate the joint density of a multivariate X_t, conditional densities, and their derivatives, and analogous conclusions follow. Similar methods are also used to estimate the nonparametric regression function r in

$$Y_t = r(Z_t) + X_t \ , \tag{9.4}$$

where Y_t and Z_t (assumed scalar for simplicity) are observable and X_t is unobservable, with zero mean. An M-estimate (z) of $r(z)$ satisfies

$$\sum_{t=1}^{N} \psi(Y_t - \hat{r}(z)) k\left(\frac{Z_t - z}{a}\right) = 0 \ , \tag{9.5}$$

giving the familiar Nadaraya–Watson estimate when $\psi(\mathrm{x}) = \mathrm{x}$. In case Z_t, like X_t, is stochastic and stationary, the $\hat{r}(z_j)$ at fixed points z_j have the same simple limiting normal distribution (with diagonal covariance matrix) under weak dependence of X_t and Z_t as they do under independence, and so the estimates can still be simply studentized, see e.g. Roussas (1969), Robinson (1983b, 1984), Boente and Fraiman (1990). However, a fixed-design version of (9.4) is also of importance, where, in the simplest case,

$$Z_t = t/N \ , \tag{9.6}$$

so that the Z_t are equally spaced points in (0,1) that actually get closer as N increases. Here, the asymptotic variance of the estimates of r is affected even by weak dependence of X_t, for example when (2.6) holds but (2.4) does not, see Roussas, Tran and Ioannides (1993), Csörgo and Mielniczuk (1995b). Their distributional behaviour under long range dependence has also been studied by Csörgo and Mielniczuk (1995a) and Robinson (1995c), the latter author showing how they can be studentized.

References

Adenstedt, R. (1974). On large-sample estimation for the mean of a stationary random sequence. *Ann. Statist.* **2**, 1095–1107.

Ahmad, I. A. (1982). Integrated mean square properties of density estimation by orthogonal series methods for dependent variables. *Ann. Inst. Statist. Math.* **34**, 339–350.

Albers, W. (1978). Testing the mean of a normal population under dependence. *Ann. Statist.* **6**, 1337–1399.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–858.

Andrews, D. W. K. and J. C. Monahan (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica* **60**, 953–966.

Beltrao, K.I. and P. Bloomfield (1987). Determining the bandwidth of a kernel spectrum estimate. *J. Time Ser. Anal.* **8**, 21–38.

Bentkus, V., F. Götze and A. Tikhomirov (1995). Berry-Esseen bounds for statistics of weakly dependent samples. Preprint.

Beran, J. (1986). Estimation, testing and prediction for self-similar and related processes. *Ph.D. thesis*, ETH Zurich.

Beran, J. (1989). A test of location for data with slowing decaying serial correlation. *Biometrika* **76**, 261–269.

Beran, R. (1976). Adaptive estimates for autoregressive processes. *Ann. Inst. Statist. Math.* **28**, 77–99.

Bertail, P., D. N. Politis and J. P. Romano (1995). On subsampling estimators with unknown rate of convergence. Preprint.

Bickel, P. J. and A. M. Herzberg (1979). Robustness of design autocorrelation in time I: Asymptotic theory, optimality for location and linear regression. *Ann. Statist.* **7**, 77–95.

Boente, G. and R. Fraiman (1990). Asymptotic distribution of robust estimates for nonparametric models from mixing processes. *Ann. Statist.* **18**, 891–906.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *Ann. Statist.* **16**, 1709–1722.

Bustos, O. H. (1982). General M-estimates for contaminated p^{th}-order autoregressive processes: Consistency and asymptotic normality. Z. Wahrsch. verw. Gebiete **59**, 491–509.

Brillinger, D. R. (1973). Estimation of the mean of a stationary time series by sampling. *J. Appl. Prob.* **10**, 419–431.

Brillinger, D. R. (1975). Time Series, Data Analysis and Theory. Holden-Day, San Francisco.

Brillinger, D. R. (1979). Confidence intervals for the crosscovariance function. *Selecta Statist. Canad.* **5**, 3–16.

Bühlmann, P. and H. R. Künsch (1994). Block length selection in the bootstrap for time series. Preprint.

Burg, J. P. (1975). Maximum entropy spectral analysis. Ph.D. dissertation, Stanford University, Stanford, California.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14**, 1171–1179.

Cheng, B. and P. M. Robinson (1991). Density estimation in strongly dependent non-linear time-series. *Statist. Sinica* **1**, 335–359.

Cochrane, D. and G. H. Orcutt (1949). Application of least squares regression to relationships containing autocorrelated error terms. *J. Amer. Statist. Assoc.* **44**, 32–61.

Csörgo, S. and J. Mielniczuk (1995a). Close short-range dependent sums and regression estimation. *Acta Sci. Math.* (Szeged), **60**, 177–196.

Csörgo, S. and J. Mielniczuk (1995b). Nonparametric regression under long-range dependent normal errors. *Ann. Statist.* **23**, 1000–1014.

Dalhaus, R. (1992). Efficient location and regression estimation for long range dependent regression models. Preprint.

Dickey, D. A. and W. A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* **74**, 427–431.

Domowitz, I. and H. White (1982). Misspecified models with dependent observations. *J. Econometrics* **20**, 35–58.
Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1–26.
Eicker, F. (1967). Limit theorems for regression with unequal and dependent errors. In: Proc. Fifth Berkeley Symp. Math. Statist. Prob., 59–82, University of California Press, Berkeley, California.
Engle, R. F. (1974). Band spectrum regression. *Internat. Econom. Rev.* **15**, 1–11.
Fox, R. and M. S. Taqqu (1986). Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Statist.* **14**, 517–32.
Franke, J. and W. Härdle (1992). On bootstrapping kernel spectral estimates. *Ann. Statist.* **20**, 121–145.
Freedman, D. (1984). On bootstrapping two-stage least-squares estimates in stationary linear models. *Ann. Statist.* **12**, 827–842.
Gallant, A. R. and H. White (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Basil Blackwell, Oxford.
Gastwirth, J. L. and H. Rubin (1971). Effects of dependence on the level of some one-sample tests. *J. Amer. Statist. Assoc.* **66**, 816–820.
Gastwirth, J. L. and H. Rubin (1975). The behaviour of robust estimators on dependent data. *Ann. Statist.* **3**, 1070–1107.
Gastwirth, J. L., H. Rubin and S. S. Wolff (1967). The effect of autoregressive dependence on a nonparametric test. *IEEE Trans. Information Theory* **IT-13**, 311–313.
Götze, F. and C. Hipp (1983). Asymptotic expansions for sums of weakly dependent random vectors. *Z. Wahrsch. verw. Gebiete* **64**, 211–239.
Götze, F. and C. Hipp, (1994). Asymptotic distribution of statistics in time series. *Ann. Statist.* **22**, 2062–2088.
Götze, F. and H. R. Künsch (1995). Second-order correctness of the blockwise bootstrap for stationary observations. Preprint.
Grenander, U. (1954). On the estimation of regression coefficients in the case of an autocorrelated disturbance. *Ann. Math. Statist.* **25**, 252–272.
Grenander, U. and M. Rosenblatt (1957). *Statistical Analysis of Stationary Time Series*. Wiley, New York.
Hall, P. and J. Hart (1990). Convergence rates in density estimation for data from infinite-order moving average processes. *Probab. Theory Relat. Fields* **87**, 253–274.
Hannan, E. J. (1957). The variance of the mean of a stationary process. *J. Roy. Statist. Soc. Ser. B*, **19**, 282–285.
Hannan, E. J. (1963a). Regression for time series with errors of measurement. *Biometrika* **50**, 293–302.
Hannan, E. J. (1963b). Regression for time series. In: M. Rosenblatt, ed., Time Series Analysis, 17–37. Wiley, New York.
Hannan, E. J. (1967). The estimation of a lagged regression relation. *Biometrika* **54**, 409–418.
Hannan, E. J. (1970). Multiple Time Series. Wiley, New York.
Hannan, E. J. (1971). Non-linear time series regression. *J. Appl. Prob.* **8**, 767–780.
Hannan, E. J. (1973). Central limit theorems for time series regression. Z. Wahrsch. verw. Gebiete **26**, 157–170.
Hannan, E. J. (1979). The central limit theorem for time series regression. *Stoch. Proc. Appl.* **9**, 281–289.
Hannan, E. J. (1987). Rational transfer function approximation. *Statist. Sci.* **2**, 135–161.
Hannan, E. J. and J. Rissanen, (1988). The width of a spectral window. *J. Appl. Prob.* **25A**, 301–307.
Hannan, E. J. and P.M. Robinson, (1973). Lagged regression with unknown lags. *J. Roy. Statist. Soc. Ser. B* **35**, 252–267.
Hansen, B. E. (1992). Consistent covariance matrix estimation for dependent heterogeneous processes. *Econometrica* **60**, 967–972.
Hansen, L. P. (1982). Large sample properties of generalized method of moments estimates. *Econometrica* **50**, 1029–1059.
Hodges, J. L. Jr and E. L. Lehmann (1970). Deficiency. *Ann. Math. Statist.* **41**, 783–801.
Huber, P. J. (1964). Estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Huber, P. J. (1972). The behaviour of maximum likelihood estimates under non-standard conditions. *Proc. Fifth Berkeley Sympos. Math. Statist. Probab.* **1**, 221–233. University of California Press.

Hurvich, C. M. (1985). Data driven choice of a spectrum estimate: extending the applicability of cross-validation methods. *J. Amer. Statist. Assoc.* **80**, 933–950.

Ibragimov, I. A. and Y. V. Linnik (1971). Independent and Stationary Sequences of Random Variables. Wolters-Noordhoff, Groningen.

Jenkins, G. M. and D. G. Watts (1968). Spectrum Analysis and its Applications. Holden-Day, San Francisco.

Jowett, G. H. (1955). The comparison of means of sets of observations from sections of independent stochastic series. *J. Roy. Statist. Soc. Ser. B* **17**, 208–227.

Koul, H. L. (1977). Behaviour of robust estimates in the regression model with dependent errors. *Ann. Statist.* **5**, 681–699.

Koul, H. L. (1992). *M*-estimators in linear models with long range dependent errors. *Statist. Probab. Lett.* **14**, 153–164.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217–1241.

Künsch, H. R., J. Beran and F. Hampel (1993). Contrasts under long-range dependence. *Ann. Statist.* **21**, 943–964.

Lahiri, S. N. (1991). Second order optimality of stationary bootstrap. *Statist. Probab. Lett.* **11**, 335–341.

Lahiri, S. N. (1994). On Edgeworth expansion and moving block bootstrap for studentized *M*-estimators in multiple linear regression models. Preprint.

Levine, D. (1983). A remark on serial correlation in maximum likelihood. *J. Econometrics* **23**, 337–342.

Liu, R. Y. and K. Singh (1992). Moving block jackknife and bootstrap capture weak dependence. In: R. LePage and L. Billard, eds., *Exploring the Limits of Bootstrap*, 225–258. Wiley, New York.

Lomnicki, Z. A. and S.K. Zaremba (1957). On estimating the spectral density of a stochastic process. *J. Roy. Statist. Soc. Ser. B*, **19**, 13–37.

Martin, R. D. and V. J. Yohai (1986). Influence functionals for time series (with discussion). *Ann. Statist.* **14**, 781–818.

Newey, W. K. and K. D. West (1987). A simple, positive semidefinite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–708.

Newey, W. K. and K. D. West (1994). Automatic lag selection in covariance matrix estimation. *Rev. Econom. Stud.* **61**, 631–653.

Parzen, E. (1962a). Spectral analysis of asymptotically stationary time series. *Bull. Int. Statist. Inst.* **39**, 87–103.

Parzen, E. (1962b). On estimation of a probability density and mode. *Ann. Math. Statist.* **38**, 1065–1076.

Parzen, E. (1969). Multiple time series modelling. In: P. R. Krishnaiah, ed., *Multivariate Analysis - II*, 383–509. Academic Press, New York.

Phillips, P.C. B. (1987). Time series regression with a unit root. *Econometrica* **55**, 277–301.

Pickands, J. III (1970). Spectral estimation with random truncation. *Ann. Math. Statist.* **41**, 44–58.

Pisarenko, V. F. (1972). On the estimation of spectra by means of non-linear functions of the covariance matrix. *Geophys. J. Roy. Astron. Soc.* **28**, 511–531.

Politis, D. N. and J. P. Romano, (1992). A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation. *Ann. Statist.* **20**, 1985–2007.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *J. Amer. Statist. Assoc.* **89**, 1303–1313.

Politis, D. N. and J. P. Romano (1995). A general theory for large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.*, to appear.

Portnoy, S. L. (1977). Robust estimation in dependent situations. *Ann. Statist.* **5**, 22–43.

Portnoy, S. L. (1978). Further remarks on robust estimation in dependent situations. *Ann. Statist.* **7**, 224–231.

Prakasa Rao, B. L. S. (1978). Density estimation for Markov processes using delta sequence. *Ann. Inst. Statist. Mat.* **30**, 321–328.

Press, H. and J. W. Tukey (1956). Power spectral methods of analysis and their applications to problems in airplane dynamics. *Bell Systems Monograph* 2606.

Robinson, P. M. (1972). Non-linear regression for multiple time-series. *J. Appl. Prob.* **9**, 758–768.

Robinson, P. M. (1976). The estimation of linear differential equations with constant coefficients. *Econometrica* **44**, 751–764.

Robinson, P. M. (1983a). Review of various approaches to power spectrum estimation. In: D. R. Brillinger and P. R. Krishnaiah, eds., *Handbook of Statistics*, Vol.3, 343–368. Elsevier Science Publishers, New York.

Robinson, P. M. (1983b). Nonparametric estimators for time series. *J. Time Ser. Anal.* **4**, 185–207.

Robinson, P. M. (1984). Robust nonparametric autoregression. In: J. Franke et al. eds., *Robust and nonlinear time series analysis*. Lecture Notes in Statistics, Vol. 26, 247–255. Springer-Verlag.

Robinson, P. M. (1986). On the errors-in-variables problem for time series. *J. Multivar. Anal.* **19**, 1986, 240–250.

Robinson, P. M. (1989). Hypothesis testing in semiparametric and nonparametric models for econometric time series. *Rev. Econom. Stud.* **56**, 511–534.

Robinson, P. M. (1991a). Automatic frequency domain inference on semiparametric and nonparametric models. *Econometrica* **59**, 1329–1364.

Robinson, P. M. (1991b). Nonparametric function estimates for long memory time series. In: W.A. Barnett, J. Powell and G.E. Tauchen, eds., Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics, 437–457. Cambridge University Press, Cambridge.

Robinson, P. M. (1994a). Time series with strong dependence. In Advances in Econometrics, Sixth World Congress, Vol.1 (C.A. Sims, Ed.). Cambridge University Press, Cambridge.

Robinson, P. M. (1994b). Semiparametric analysis of long memory time series. *Ann. Statist.* **22**, 515–539.

Robinson, P. M. (1995a). Log-periodogram regression of time series with long range dependence. *Ann. Statist.* **23**, 1048–1072.

Robinson, P. M. (1995b). Gaussian semiparametric estimation of long range dependence. *Ann. Statist.* **23**, 1630–1661.

Robinson, P. M. (1995c). Large-sample inference for nonparametric regression with dependent errors. Under revision for Ann. Statist..

Robinson, P. M. and J. Hidalgo (1997). Time series regression with long range dependence. Forthcoming, Ann. Statist..

Rosenblatt, M. (1970). Density estimates and Markov processes. In: M.L. Puri, ed., *Nonparametric techniques in statistical inference*, 199–210. Cambridge University Press, Cambridge.

Roussas, G. (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Ann. Math. Statist.* **40**, 1386–1900.

Roussas, G., L. Tran and J. A. Ioannides (1993). Fixed design regression for time series: Asymptotic normality. *J. Multivar. Anal.* **40**, 262–291.

Samarov, A. M. (1987). Robust spectral regression. *Ann. Statist.* **15**, 99–111.

Samarov, A. and M. S. Taqqu 1988). On the efficiency of the sample mean in long memory noise. *J. Time Ser. Anal.* **9**, 191–200.

Serfling, R. J. (1968). The Wilcoxon two-sample statistic on strongly mixing processes. *Ann. Math. Statist.* **39**, 1202–1209.

Song, W. T. and B. W. Schmeiser (1992). Variance of the sample mean: Properties and graphs of quadratic-form estimators. *Operations Res.* **41**, 501–517.

Stone, C. J. (1975). Adaptive maximum likelihood estimation of a location parameter. *Ann. Statist.* **3**, 267–284.

Taniguchi, M. (1984). Validity of Edgeworth expansions for statistics of time series. *J. Time Ser. Anal.* **5**, 37–51.

Taniguchi, M. (1987). Third order asymptotic properties of BLUE and LSE for a regression model with ARMA residuals. *J. Time Ser. Anal.* **8**, 111–114.

Taqqu, M. S. (1975). Weak convergence to fractional Brownian motion and to Rosenblatt process. *Z. Wahrsch. Verw. Gebiete* **31**, 287–302.

Tukey, J. W. (1967). An introduction to the calculations of numerical spectrum analysis. In: B. Harris, ed., *Advanced Seminar on Spectral Analysis of Time Series*, 25–46. Wiley, New York.

White, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press, New York.

White, H. and I. Domowitz (1984). Nonlinear regression with dependent observations. *Econometrica* **52**, 143–161.

Wu, C. F. J. (1990). On the asymptotic properties of the jackknife histogram. *Ann. Statist.* **18**, 1438–1452.

Yajima, Y. (1988). On estimation of a regression model with long-memory stationary errors. *Ann. Statist.* **16**, 791–807.

Yajima, Y. (1991). Asymptotic properties of the LSE in a regression model with long-memory stationary errors. *Ann. Statist.* **19**, 158–177.

Zamar, R. H. (1990). Robustness against unexpected dependence in the location model. *Statist. Probab. Lett.* **9**, 367–374.

Zygmund, A. (1977). *Trigonometric Series*. Cambridge University Press, Cambridge.