The Annals of Statistics
1996, Vol. 24, No. 3, 952–974

ASYMPTOTICALLY OPTIMAL ESTIMATION IN MISSPECIFIED TIME SERIES MODELS

BY R. DAHLHAUS AND W. WEFELMEYER

Universität Heidelberg and Universität Siegen

A concept of asymptotically efficient estimation is presented when a misspecified parametric time series model is fitted to a stationary process. Efficiency of several minimum distance estimates is proved and the behavior of the Gaussian maximum likelihood estimate is studied. Furthermore, the behavior of estimates that minimize the h-step prediction error is discussed briefly. The paper answers to some extent the question what happens when a misspecified model is fitted to time series data and one acts as if the model were true.

1. Introduction. Let $X_1, \ldots, X_n$ be a sample from a real-valued stationary process X_t, $t \in \mathbb{Z}$, with mean 0, spectral density $f(\lambda)$, $\lambda \in [-\pi, \pi]$ and covariance function $c(u)$, $u \in \mathbb{Z}$. Suppose, for example, that we want to make a one-step-ahead prediction, and that we want to use for convenience an AR(p)-model (autoregressive model of order p—for this model the predictor has a simple form); that is, we use the *model*

$$X_t + a_1 X_{t-1} + \cdots + a_p X_{t-p} = \varepsilon_t,$$

where ε_t are iid with mean 0 and variance σ^2. The best linear predictor of X_{n+1} in an AR(p)-model is

$$\hat{X}_{n+1} = -\sum_{j=1}^{p} a_j X_{n+1-j} \tag{1.1}$$

[cf. Brockwell and Davis (1987), page 170, Example 5.3.1]. The mean square prediction error under the true distribution of the process is

$$\mathrm{PE}(\theta, f) = E\left(\sum_{j=0}^{p} a_j X_{t-j}\right)^2 = \sum_{j,k=0}^{p} a_j a_k c(k-j) = \int_{-\pi}^{\pi} f(\lambda)\left|A_\theta(\lambda)\right|^2 d\lambda,$$

where $a_0 = 1$, $\theta = (a_1, \ldots, a_p)$ and $A_\theta(\lambda) = \sum_{j=0}^{p} a_j \exp(-i\lambda j)$. Here $\mathrm{PE}(\theta, f)$ is a kind of distance between the true process and an AR(p)-process. Suppose now that we wish to estimate the parameter θ_0 which leads to the best mean square prediction error (note that the process is misspecified and hence there exists no "true" value θ_0); that is, we want to estimate

$$\theta_0 = \arg\min_{\theta} \mathrm{PE}(\theta, f).$$

Received October 1994; revised July 1995.

AMS 1991 *subject classifications*. Primary 62M10; secondary 62G20.

Key words and phrases. Time series, misspecified models, efficiency, minimum distance estimation, maximum likelihood, prediction.

A natural way to estimate θ_0 is to replace the covariance function $c(u)$ or the spectral density $f(\lambda)$ by a nonparametric estimate and to minimize the resulting empirical function. For example, we may take as an estimate of $f(\lambda)$ the periodogram

$$I_n(\lambda) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} X_t \exp(-i\lambda t)\right|^2$$

and minimize

$$\mathrm{PE}(\theta, I_n) = \int_{-\pi}^{\pi} I_n(\lambda)\left|A_\theta(\lambda)\right|^2 d\lambda.$$

This leads to the Yule–Walker estimate $\hat{\theta}_n$ of θ_0. (Estimates with better small-sample properties such as maximum likelihood estimates and tapered Yule–Walker estimates will be discussed in Section 4.) The innovation variance may be "estimated" by $\hat{\sigma}_n^2 = \mathrm{PE}(\hat{\theta}_n, I_n)$; the corresponding theoretical value is $\sigma_0^2 = \mathrm{PE}(\theta_0, f)$. We can proceed in a similar way if we wish to estimate those parameters of an AR(p)-model that lead to the best h-step-ahead prediction error (cf. Section 4).

An equivalent approach to the minimization of the one-step-ahead prediction error is to look for the model which is closest in the sense of the (asymptotic) Kullback–Leibler information divergence. For a Gaussian process and a Gaussian model, this divergence has the form

$$\frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\log\frac{f_\theta(\lambda)}{f(\lambda)} + \frac{f(\lambda)}{f_\theta(\lambda)} - 1\right\} d\lambda, \tag{1.2}$$

where $f_\theta(\lambda)$ is the spectral density of the model [cf. Pinsker (1963) and Parzen (1982, 1992)]. Thus, the best fit is achieved by

$$\theta_0 = \arg\min_\theta D(\theta, f), \tag{1.3}$$

where now

$$D(\theta, f) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\log f_\theta(\lambda) + \frac{f(\lambda)}{f_\theta(\lambda)}\right\} d\lambda. \tag{1.4}$$

It may be estimated by

$$\hat{\theta}_n = \arg\min_\theta D(\theta, I_n)$$

where $\hat{\theta}_n$ is called the Whittle estimate [Whittle (1952)]. For AR(p)-models we have $f_\theta(\lambda) = (\sigma^2/2\pi)|\sum_{j=0}^{p} a_j \exp(-i\lambda j)|^{-2}$ [we now set $\theta = (a_1, \ldots, a_p, \sigma^2)$]. Since Kolmogorov's formula gives

$$\frac{1}{4\pi}\int_{-\pi}^{\pi}\log f_\theta(\lambda)\, d\lambda = \frac{1}{2}\log\frac{\sigma^2}{2\pi}$$

[cf. Brockwell and Davis (1987), page 184, Theorem 5.8.1], the values θ_0 and $\hat{\theta}_n$ are exactly the same as the values obtained by minimization of the

one-step-ahead prediction error. In addition, σ_0^2 and $\hat{\sigma}_n^2$ are the same as the values obtained by minimization of the one-step-ahead prediction error. Now σ_0^2 and $\hat{\sigma}_n^2$ are also obtained as solutions of a minimization problem.

Since $D(\theta, I_n)$ converges uniformly to $D(\theta, f)$ [see (3.1)], we immediately get that $\hat{\theta}_n$ is a consistent estimate of θ_0. In this paper we prove that $\hat{\theta}_n$ is also efficient if the true underlying process is Gaussian (Theorem 3.2). The same holds for the Gaussian maximum likelihood estimate (Theorem 3.3). This is rather surprising since the MLE in a misspecified model is, in general, not an efficient estimate for the "best" approximating parameter (where "best" is meant in the sense of the Kullback–Leibler distance). As an example consider the situation where the true distribution of the process is AR(p) with ε_t following an unknown distribution.

Our efficiency result is proved by considering the best fit θ_0 as a functional $\theta_0 = T(f)$ of the unknown spectral density f and then applying a nonparametric version of the convolution theorem of Hájek (1970). Other functionals and estimates are treated by Hasminskii and Ibragimov (1986) and Ginovyan (1988). Their efficiency concept is based on a local asymptotic minimax theorem rather than a convolution theorem.

Note that in our setting the true spectral density lies outside the parametric model. Furthermore, it does *not* approach the parametric model asymptotically. Hence, our setting is not covered by the general results of Millar (1984) on optimality of minimum distance estimates. Efficiency in our sense is considered by Beran (1977) for iid observations and the Hellinger distance, and by Greenwood and Wefelmeyer (1993) for Markov chains and the Kullback–Leibler distance. Despite similar titles, there is little overlap with the latter paper. This has two reasons. In Greenwood and Wefelmeyer (1993), the true model is nonparametric. Hence, we expect the MLE in the misspecified model to be efficient, because it can be interpreted as a function of the empirical distribution. In the present paper, the true model is semiparametric, and a function of the empirical distribution is, in general, inefficient in such a model. Hence, our efficiency result is unexpected. The second reason is that the present paper deals with a specific parametrization, while the previous one leaves the parametrization unspecified, leading to technicalities that do not come up here.

More generally, we consider in this paper distance functions of the form

$$D(\theta, f) = \int_{-\pi}^{\pi} K(\theta, f(\lambda), \lambda)\, d\lambda.$$

We set

$$T(f) := \arg\min_{\theta} D(\theta, f)$$

and make the following assumption.

ASSUMPTION 1.1. $\Theta \subset \mathbb{R}^k$ is compact and $K\colon \Theta \times (0,\infty) \times [-\pi, \pi] \to \mathbb{R}$ is three times differentiable in (θ, x) with continuous derivatives in (θ, x, λ).

For the true spectral density f, we assume that $T(f)$ exists, is unique and lies in the interior of Θ.

One would usually consider functions with $T(f_\theta) = \theta$. However, we do not need this assumption. Taniguchi (1987) considers distance functions of the special type $K(\theta, f(\lambda), \lambda) = \overline{K}(f_\theta(\lambda)/f(\lambda))$, where $f_\theta(\lambda)$ is the spectral density of the model. An important distance function which is not of this form is the h-step prediction error (cf. Section 4). If we take $\overline{K}(f_\theta(\lambda)/f(\lambda))$ as the distance function, then Assumption 1.1 is fulfilled if $\overline{K}(x)$ is three times continuously differentiable with unique minimum at $x = 1$ and the model spectral densities f_θ fulfill the following assumption.

ASSUMPTION 1.2. $\Theta \subset \mathbb{R}^k$ is compact and the model spectral density f_θ is two times differentiable with respect to θ with continuous (in θ and λ) derivatives. f_θ and its derivatives are uniformly bounded and bounded away from 0.

Note that we do not assume $f_{\theta_1} \neq f_{\theta_2}$ for $\theta_1 \neq \theta_2$. Instead we assume that $T(f)$ exists uniquely. This is of importance when only part of the parameters are estimated (as in the case of the prediction error where σ^2 is not estimated by minimizing a distance function).

The assumptions on the observed process are as follows.

ASSUMPTION 1.3. X_t, $t \in \mathbb{Z}$, is a Gaussian stationary process with $EX_t = 0$ and a Lipschitz-continuous spectral density f which is bounded and bounded away from 0.

As estimates of $T(f)$ we consider in this paper $T(I_n)$ and $T(\hat{f}_n)$, where

$$\hat{f}_n(\lambda) := \int_{-\pi}^{\pi} I_n(\lambda + \alpha) W_n(\alpha)\, d\alpha \tag{1.5}$$

is a kernel estimate of f. For the kernel we need the following assumption.

ASSUMPTION 1.4. $W_n(\alpha) = mW(m\alpha)$, where $n^{1/4} \ll m \ll n^{1/2}$ and W is bounded, symmetric and nonnegative with $W(x) = 0$ for $|x| > c$ and $\int_{-c}^{c} W(x)\, dx = 1$. The Fourier transform $\hat{W}(x)$ is assumed to be continuous with $\int_{-\infty}^{\infty} |\hat{W}(x)|\, dx < \infty$.

The above assumptions are discussed after Lemma A.7.

The use of $\hat{f}_n$ instead of I_n is important for distance functions that are not linear in f, since in this case $D(\theta, I_n)$ will usually not converge to $D(\theta, f)$, with the consequence that $T(I_n)$ is not a consistent estimate of $T(f)$. Taniguchi (1987) has proved asymptotic normality of $T(\hat{f}_n)$ for general stationary time series, and efficiency when the model is correctly specified. Asymptotic normality of $T(I_n)$ for the Whittle distance has been proved by several authors. We mention Whittle (1952), Walker (1964), Dzhaparidze

(1971), Hannan (1973), Hosoya and Taniguchi (1982) and Hosoya (1989). Asymptotic efficiency of $T(I_n)$ in correctly specified models has been proved by Dzhaparidze (1986).

We also study the efficiency of the exact Gaussian maximum likelihood estimate under model misspecification. The asymptotic distribution of this estimate was derived under model misspecification by Ogata (1980). Asymptotic efficiency of the MLE in correctly specified models has been proved by Dzhaparidze (1986) and Maliukevićius (1989).

In particular, we prove that $T(\hat{f}_n)$ is an efficient estimate of $T(f)$ even if the model is misspecified. $T(I_n)$ turns out to be efficient if the distance function is linear in f while the MLE is only efficient if the Kullback–Leibler divergence is used as a distance function.

The asymptotic variance bound is derived in Section 2. Efficiency of several estimates is proved in Section 3. The results are discussed together with some examples in Section 4. In particular, we consider the h-step prediction error in some detail. Some technical lemmas are given in the Appendix.

2. The asymptotic variance bound. In this section we derive a lower bound for the asymptotic variance of "regular" estimates of the functional $T(f)$ introduced in Section 1. Let $X_1, \dots, X_n$ be a sample from a real-valued stationary Gaussian process with mean 0 and unknown spectral density f. The distribution of the process is determined by the spectral density. Hence, we may consider f as an infinite-dimensional "parameter" of the distribution. We want to prove that certain estimates of $T(f)$ are efficient.

In a first step, we prove that the true model is locally asymptotically normal (LAN). If the spectral density were to depend on a finite-dimensional parameter θ, we would consider the likelihood ratio corresponding to θ and a nearby parameter $\theta + (1/\sqrt{n})h$, with h an arbitrary vector, the so-called *local* parameter. Local asymptotic normality in this case was first proved by Davies (1973). In our case, however, the spectral density is completely unknown. Hence, we fix a spectral density f and consider a nearby spectral density of the form $f(\lambda)(1 + (1/\sqrt{n})h(\lambda))$, with h an arbitrary (bounded) function. The function h now plays the role of local parameter.

Let L_2 denote the space of functions on $(-\pi, \pi)$ which are square-integrable with respect to Lebesgue measure. Introduce the inner product

$$\langle h, k \rangle = \frac{1}{4\pi} \int_{-\pi}^{\pi} h(\lambda) k(\lambda)\, d\lambda$$

and the norm $\|h\|^2 = \langle h, h \rangle$ on L_2. For a bounded function h set

$$f_{nh}(\lambda) = f(\lambda)\left(1 + \frac{1}{\sqrt{n}} h(\lambda)\right), \qquad \lambda \in (-\pi, \pi).$$

Furthermore, let

$$\Sigma_n(g) = \left\{ \int_{-\pi}^{\pi} g(\lambda) \exp(i\lambda(r-s))\, d\lambda \right\}_{r,s=1,\dots,n}$$

be the Toeplitz matrix of g and $\mathbf{X}_n = (X_1, \ldots, X_n)'$. If g is a vector function, $\Sigma_n(g)$ is the corresponding vector of matrices.

The following result is due to Dzhaparidze (1986).

THEOREM 2.1. *Let f be even, positive, bounded and bounded away from 0. Let P_{nh} be the distribution of the observations $X_1, \ldots, X_n$ of a Gaussian stationary process with mean 0 and spectral density $f_{nh} = f(1 + (1/\sqrt{n})h)$, where h is even and bounded. Then we have, under the law P_{n0},*

$$\log \frac{dP_{nh}}{dP_{n0}} - Z_n(h) + \frac{1}{2}\langle h, h\rangle \to_P 0,$$

where

$$Z_n(h) = \frac{1}{2\sqrt{n}}\left(\mathbf{X}_n'\Sigma_n(f)^{-1}\Sigma_n\left(\frac{h}{2\pi}\right)\mathbf{X}_n - \frac{n}{2\pi}\int_{-\pi}^{\pi} h(\lambda)\, d\lambda\right).$$

[*Note that $Z_n(h)$ can be written in the form $\langle h, Z_n'\rangle$.*] *Furthermore,*

$$Z_n(h) \to_D N(0, \langle h, h\rangle),$$

that is, the sequence P_{nh} is LAN.

PROOF. See Dzhaparidze [(1986), page 64, Section 1.3, Theorem 4(3), and page 155, Section 2, Theorem A1.2]. □

The following result is also due to Dzhaparidze [(1986), page 64, Section 1.3, Theorem 4(4)]. We prove it under slightly different conditions.

THEOREM 2.2. *Suppose Assumption* 1.3 *holds and h is even and bounded. Then*

$$\frac{\sqrt{n}}{4\pi}\int_{-\pi}^{\pi} \frac{I_n(\lambda) - f(\lambda)}{f(\lambda)} h(\lambda)\, d\lambda - Z_n(h) \to_P 0.$$

PROOF. Lemma A.7(iv) implies

$$\frac{\sqrt{n}}{4\pi}\int_{-\pi}^{\pi} \frac{EI_n(\lambda) - f(\lambda)}{f(\lambda)} h(\lambda)\, d\lambda = o(1).$$

Since

$$\int_{-\pi}^{\pi} I_n(\lambda) g(\lambda)\, d\lambda = \frac{1}{2\pi n}\mathbf{X}_n'\Sigma_n(g)\mathbf{X}_n,$$

the variance of the expression in Theorem 2.2 is equal to

$$\frac{1}{n}\operatorname{Var}\left[\mathbf{X}_n'\left\{\Sigma_n\left(\frac{h}{8\pi^2 f}\right) - \Sigma_n(f)^{-1}\Sigma_n\left(\frac{h}{4\pi}\right)\right\}\mathbf{X}_n\right].$$

If $\mathbf{X}$ is Gaussian with covariance matrix Σ, then $\operatorname{Var}(\mathbf{X}'A\mathbf{X}) = \operatorname{tr}(\Sigma A \Sigma A) + \operatorname{tr}(\Sigma A \Sigma A')$. Therefore, this variance is, with $\Sigma_f = \Sigma_n(f)$, equal to

$$\frac{2}{n}\left[\operatorname{tr}\left\{\Sigma_n\left(\frac{h}{8\pi^2 f}\right)\Sigma_f \Sigma_n\left(\frac{h}{8\pi^2 f}\right)\Sigma_f\right\} - \operatorname{tr}\left\{\Sigma_n\left(\frac{h}{8\pi^2 f}\right)\Sigma_n\left(\frac{h}{4\pi}\right)\Sigma_f\right\}\right.$$
$$\left. - \operatorname{tr}\left\{\Sigma_n\left(\frac{h}{4\pi}\right)\Sigma_f \Sigma_n\left(\frac{h}{8\pi^2 f}\right)\right\} + \operatorname{tr}\left\{\Sigma_n\left(\frac{h}{4\pi}\right)\Sigma_n\left(\frac{h}{4\pi}\right)\right\}\right],$$

which tends to 0 by Lemma A.3. □

We now derive a lower bound for the asymptotic variance of "regular" estimates of $T(f)$. Local asymptotic normality (see Theorem 2.1) induces the norm $\langle h, h\rangle$ on the local parameter space. The norm determines how difficult it is, asymptotically, to distinguish between f and $f(1 + (1/\sqrt{n})h)$ on the basis of a sample $X_1, \ldots, X_n$. Consider now the problem of estimating the functional $T(f)$. The convolution theorem says that a variance bound for "regular" estimates of $T(f)$ is given by the squared length of the gradient of the functional in terms of the inner product $\langle h, k\rangle$. The gradient is given in Corollary 2.4 below.

We need the following Taylor expansion which will also be used in Section 3. Let f_n be a spectral density which converges to f. Let us assume for the moment that $T(f_n)$ is also in the interior of Θ. Let ∇K denote the derivative of $K(\theta, x, \lambda)$ with respect to θ, and K' the derivative with respect to x. We obtain with $K(\theta, x) = K(\theta, x, \cdot)$

$$\begin{aligned} 0 &= \nabla D(T(f_n), f_n) \\ &= \nabla D(T(f), f_n) + \left\{\int_{-\pi}^{\pi} \nabla^2 K(T(f), f(\lambda), \lambda)\, d\lambda\right\}(T(f_n) - T(f)) \\ &\quad + \left\{\int_{-\pi}^{\pi} \nabla^2 K'\big(T(f), \tilde{f}(\lambda), \lambda\big)(f_n(\lambda) - f(\lambda))\, d\lambda\right\}(T(f_n) - T(f)) \\ &\quad + \tfrac{1}{2}(T(f_n) - T(f))'\left\{\int_{-\pi}^{\pi} \nabla^3 K\big(\tilde{t}, f_n(\lambda), \lambda\big)\, d\lambda\right\}(T(f_n) - T(f)), \end{aligned} \tag{2.1}$$

where $|\tilde{f}(\lambda) - f(\lambda)| \le |f_n(\lambda) - f(\lambda)|$ and $|\tilde{t} - T(f)| \le |T(f_n) - T(f)|$. Furthermore,

$$\begin{aligned} \nabla D(T(f), f_n) &= \nabla D(T(f), f) \\ &\quad + \int_{-\pi}^{\pi} \nabla K'(T(f), f(\lambda), \lambda)\big(f_n(\lambda) - f(\lambda)\big)\, d\lambda \\ &\quad + \tfrac{1}{2}\int_{-\pi}^{\pi} \nabla K''\big(T(f), \tilde{f}(\lambda), \lambda\big)\big(f_n(\lambda) - f(\lambda)\big)^2\, d\lambda. \end{aligned} \tag{2.2}$$

Note that $\nabla D(T(f), f) = 0$. As a first consequence we obtain the following result. [A similar result was proved by Taniguchi (1987), Theorems 1(b) and 2 in the special case $K(\theta, f(\lambda), \lambda) = \bar{K}(f_\theta(\lambda)/f(\lambda))$.]

THEOREM 2.3. *Suppose that* f_n *is a sequence with* $\|f_n - f\|_2 \to 0$ *and* $0 < C_1 \le f_n,\ f \le C_2$. *Then we have*:

(i) $T(f_n) \to T(f)$.

(ii) *If*

$$H_f := \int_{-\pi}^{\pi} \nabla^2 K(T(f), f(\lambda), \lambda)\, d\lambda$$

is a nonsingular matrix, then we have with

$$g_f(\lambda) = \tfrac{1}{2}\big(h_f(\lambda) - h_f(-\lambda)\big)$$

and

$$h_f(\lambda) = -4\pi H_f^{-1}\, \nabla K'(T(f), f(\lambda), \lambda) f(\lambda),$$

$$T(f_n) - T(f) = \frac{1}{4\pi}\int_{-\pi}^{\pi} g_f(\lambda)\frac{f_n(\lambda) - f(\lambda)}{f(\lambda)}\, d\lambda + O\big(\|f_n - f\|_2^2\big).$$

PROOF. (i) $T(f_n) \to T(f)$ is exactly analogous to Theorem 1(b) in Taniguchi (1987).

(ii) Since $T(f_n) \to T(f)$, the value $T(f_n)$ lies in the interior of Θ for n large enough. Furthermore, f_n and f are bounded from above and below. Therefore, all derivatives of K in the above Taylor expansion are bounded, and we obtain

$$\begin{aligned} T(f_n) - T(f) + O\big(|T(f_n) - T(f)|\,\|f_n - f\|_2\big) + O\big(|T(f_n) - T(f)|^2\big) \\ = \frac{1}{4\pi}\int_{-\pi}^{\pi} g_f(\lambda)\frac{f_n(\lambda) - f(\lambda)}{f(\lambda)}\, d\lambda + O\big(\|f_n - f\|_2^2\big), \end{aligned}$$

which, with part (i), implies the result. □

As a consequence we obtain the following corollary.

COROLLARY 2.4. *Let* h *be bounded and* $f_{nh} = f(1 + (1/\sqrt{n})h)$. *Then*

$$\sqrt{n}\,(T(f_{nh}) - T(f)) \to \frac{1}{4\pi}\int_{-\pi}^{\pi} g_f(\lambda) h(\lambda)\, d\lambda.$$

The function g_f is called the *gradient* of T at f. We are now in a position to formulate our efficiency concept for estimates of $T(f)$. An estimate T_n is *regular* for T at f with *limit* L if its distribution converges continuously to L in the following sense:

$$\sqrt{n}\,(T_n - T(f_{nh})) \to_D L \quad \text{under } P_{nh} \text{ for all bounded } h.$$

The convolution theorem says that the limit L is the convolution of some distribution M with a normal distribution the variance of which equals the squared length of the gradient:

$$L = M * N\big(0, \langle g_f, g_f\rangle\big). \tag{2.3}$$

By a well-known result of Anderson (1955), L is less concentrated in symmetric intervals than $N(0, \langle g_f, g_f \rangle)$. This justifies calling T_n *efficient* for T at f if its limit distribution is $N(0, \langle g_f, g_f \rangle)$. We say briefly that $\langle g_f, g_f \rangle$ is a *variance bound* for regular estimates. Note, however, that the optimality result is much stronger: it holds for all (bounded) symmetric bowl-shaped loss functions, not just for the (truncated) quadratic loss function.

We also have the following useful characterization: an estimate T_n is regular and efficient for T at f if and only if it admits the following stochastic approximation:

$$\sqrt{n}\,(T_n - T(f)) - Z_n(g_f) \to_P 0. \tag{2.4}$$

A convenient reference for the above version of the convolution theorem, and the characterization, is Greenwood and Wefelmeyer (1990). There it is also pointed out that the convolution theorem implies its own multivariate version. Specifically, let $T = (T_1, \ldots, T_k)'$ be a finite-dimensional functional of the spectral density. If g_j is the gradient of T_j, then $g_f = (g_1, \ldots, g_k)'$ is called the gradient of T. The convolution theorem (2.3) is true with a k-dimensional normal distribution with covariance matrix $\langle g_f, g_f' \rangle$. The characterization (2.4) is true with vectors T and g_f.

3. Efficient estimates. In this section we study several estimates of $T(f)$. We start with $T(\hat{f}_n)$, where $\hat{f}_n$ is a kernel estimate as in (1.5). As seen in Section 2, proving efficiency means proving the stochastic approximation (2.4).

THEOREM 3.1. *Suppose Assumptions* 1.1, 1.3 *and* 1.4 *hold and* H_f *is nonsingular. Then we have*

$$\sqrt{n}\left(T\left(\hat{f}_n\right) - T(f)\right) - Z_n(g_f) \to_P 0,$$

that is, $T(\hat{f}_n)$ *is an efficient estimate of* $T(f)$.

PROOF. We start by proving consistency. We cannot apply Theorem 2.3 directly, since $\hat{f}_n$ is not necessarily bounded. f is bounded and bounded away from 0, say $0 < 2\gamma_1 \le f \le \gamma_2$. Then, on the set $B_n = \{\|\hat{f}_n - f\|_\infty \le \gamma_1\} \cap \{\|\hat{f}_n - f\|_2 < n^{-1/4}\}$ we have $0 < \gamma_1 \le f$, $\hat{f}_n \le \gamma_1 + \gamma_2$. Lemma A.7 implies $P(B_n) \to 0$ and we therefore obtain as in the proof of Theorem 2.3 that

$$\sqrt{n}\left(T\left(\hat{f}_n\right) - T(f)\right) = \frac{\sqrt{n}}{4\pi}\int_{-\pi}^{\pi} g_f(\lambda)\frac{\hat{f}_n(\lambda) - f(\lambda)}{f(\lambda)}\,d\lambda + O_p\left(n^{1/2}\|\hat{f}_n - f\|_2^2\right),$$

which by using Lemma A.7(i) and (iii) is equal to

$$\frac{\sqrt{n}}{4\pi}\int_{-\pi}^{\pi} g_f(\lambda)\frac{I_n(\lambda) - f(\lambda)}{f(\lambda)}\,d\lambda + o_p(1)$$

[cf. Taniguchi (1987), proof of Theorem 2]. Theorem 2.2 implies that this is $Z_n(g_f) + o_p(1)$. □

We now study the estimate $T(I_n)$ with distance functions that are linear in $f(\lambda)$. An example is the Kullback–Leibler distance as in (1.4).

THEOREM 3.2. *Suppose Assumptions* 1.1 *and* 1.3 *hold, where* $K(\theta, x, \lambda) = a_\theta(\lambda) + b_\theta(\lambda)x$. *If* H_f *is nonsingular, then*

$$\sqrt{n}\,(T(I_n) - T(f)) - Z_n(g_f) \to_P 0,$$

that is, $T(I_n)$ *is an efficient estimate of* $T(f)$. *Here*

$$g_f(\lambda) = -2\pi H_f^{-1}\,\nabla\big(b_{T(f)}(\lambda) + b_{T(f)}(-\lambda)\big)f(\lambda)$$

and

$$H_f = \int_{-\pi}^{\pi}\big(\nabla^2 a_{T(f)}(\lambda) + \nabla^2 b_{T(f)}(\lambda)f(\lambda)\big)\,d\lambda.$$

PROOF. Lemma A.7(v) implies that

$$\sup_\theta |D(\theta, I_n) - D(\theta, f)| \to_P 0. \tag{3.1}$$

Since $D(T(I_n), I_n) \le D(T(f), I_n)$ and $D(T(f), f) \le D(T(I_n), f)$, it follows that $D(T(I_n), f) \to D(T(f), f)$ in probability and therefore also $T(I_n) \to T(f)$ in probability. A modification of the Taylor expansion (2.1) yields with $|\tilde t - T(f)| \le |T(I_n) - T(f)|$:

$$\begin{aligned}\frac{\sqrt{n}}{4\pi}&\int_{-\pi}^{\pi} g_f(\lambda)\frac{I_n(\lambda) - f(\lambda)}{f(\lambda)}\,d\lambda\\ &= H_f^{-1}\bigg\{\int_{-\pi}^{\pi}\nabla^2 K(\tilde t, f(\lambda), \lambda)\,d\lambda\\ &\qquad + \int_{-\pi}^{\pi}\nabla^2 b_{\tilde t}(\lambda)(I_n(\lambda) - f(\lambda))\,d\lambda\bigg\}\sqrt{n}\,(T(I_n) - T(f)).\end{aligned}$$

Since

$$\frac{\sqrt{n}}{4\pi}\int_{-\pi}^{\pi} g_f(\lambda)\frac{I_n(\lambda) - f(\lambda)}{f(\lambda)}\,d\lambda = Z_n(g_f) + o_p(1) \qquad \text{(Theorem 2.2)},$$

$$\int_{-\pi}^{\pi}\nabla^2 K(\tilde t, f(\lambda), \lambda)\,d\lambda \to_P H_f \qquad \text{(smoothness of } K\text{)}$$

and

$$\int_{-\pi}^{\pi}\nabla^2 b_{\tilde t}(\lambda)(I_n(\lambda) - f(\lambda))\,d\lambda \to_P 0 \qquad [\text{Lemma A.7(v)}],$$

the result is proved. □

At this point we also want to mention another optimality property of the above estimate in correctly specified models which was first proved by Whittle (1953) [for generalizations see Kabaila (1980) and Dzhaparidze

(1984)]: consider the class of all estimates $T(I_n)$, where $K(\theta, x, \lambda) = a_\theta(\lambda) + b_\theta(\lambda)x$ with $T(f_\theta) = \theta$ (examples for different K with this property are the one- and three-step-ahead prediction errors—cf. Section 4). Then $T^*(I_n)$ with $K^*(\theta, x, \lambda) = (1/4\pi)\log f_\theta(\lambda) + (1/4\pi)f_\theta^{-1}(\lambda)I_n(\lambda)$ is the estimate of θ_0 with the smallest variance among all estimates of this class. This also holds for non-Gaussian processes (the MLE may have a smaller variance but it is of a different form). $\hat{\theta}_n = T^*(I_n)$ is the Whittle estimate (cf. Section 1).

In the next theorem we consider the Kullback–Leibler distance $D(\theta, f)$ as in (1.4) [i.e., we have $K(\theta, x, \lambda)$ as in Theorem 3.2 with $a_\theta(\lambda) = (1/4\pi)\log f_\theta(\lambda)$ and $b_\theta(\lambda) = (1/4\pi)f_\theta^{-1}(\lambda)$]. We then have

$$H_f = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\big(f(\lambda) - f_{\theta_0}(\lambda)\big)\nabla^2 f_{\theta_0}^{-1}(\lambda) + \big(\nabla \log f_{\theta_0}(\lambda)\big)\big(\nabla \log f_{\theta_0}(\lambda)\big)'\right\} d\lambda.$$

As a consequence of Theorems 2.1 and 3.2, we now obtain for the Whittle estimate $\hat{\theta}_n = T(I_n)$:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to_D N\left(0, \frac{1}{4\pi}H_f^{-1}\left\{\int_{-\pi}^{\pi} f^2(\lambda)\big(\nabla f_{\theta_0}^{-1}(\lambda)\big)\big(\nabla f_{\theta_0}^{-1}(\lambda)\big)' d\lambda\right\}H_f^{-1}\right),$$

a result already proved by Taniguchi (1979). In the special case of an AR(p)-model, this result was proved by Bhansali (1981) (in this case the Whittle estimate is identical to the Yule–Walker estimate). In addition, we now know that this limit variance is the smallest which can be achieved by regular estimators under model misspecification.

We now study the behavior of the Gaussian maximum likelihood estimate $\tilde{\theta}_n$ *of the fitted model*, that is,

$$\tilde{\theta}_n = \arg\min \mathscr{L}_n(\theta),$$

where

$$\mathscr{L}_n(\theta) = \frac{1}{2}\log(2\pi) + \frac{1}{2n}\log\det \Sigma_n(f_\theta) + \frac{1}{2}\mathbf{X}_n'\Sigma_n(f_\theta)^{-1}\mathbf{X}_n.$$

Below we prove that $\tilde{\theta}_n$ is also an asymptotically efficient estimate of θ_0. It is obvious that $\tilde{\theta}_n$ cannot be efficient for any point different from θ_0. Thus, if we choose a distance function different from the Kullback–Leibler divergence, the maximum likelihood estimate will usually not be consistent. In particular, this holds if θ_0 is the parameter that gives the best h-step prediction error for $h \geq 2$ (cf. Section 4).

THEOREM 3.3. *Suppose θ_0 is the unique solution of* (1.3) *and lies in the interior of Θ. Suppose further that Assumptions* 1.2 *and* 1.3 *hold. If H_f is nonsingular, then we have, with $g_f = -H_f^{-1} f\, \nabla f_{\theta_0}^{-1}$,*

$$\sqrt{n}\left(\tilde{\theta}_n - \theta_0\right) - Z_n(g_f) \to_P 0,$$

that is, the Gaussian maximum likelihood estimate is efficient for the point θ_0 which minimizes the asymptotic Kullback–Leibler information divergence.

PROOF. Unfortunately, it is much more difficult to prove the analogous result to (3.1) with $\mathscr{L}_n(\theta)$ instead of $D(\theta, I_n)$. Therefore, we follow the method of proof of Walker (1964), Section 2, to prove consistency of $\tilde{\theta}_n$. We start by proving that for all $\theta_1 \in \Theta$, $\theta_1 \neq \theta_0$ there exists a constant $c(\theta_1) > 0$ with

$$\lim_{n\to\infty} E\{\mathscr{L}_n(\theta_1) - \mathscr{L}_n(\theta_0)\} \geq c(\theta_1).$$

Let $\Sigma_\theta = \Sigma_n(f_\theta)$, $\Sigma_f = \Sigma_n(f)$ and $\Sigma_\nabla = \Sigma_n(\nabla f_{\theta_0})$. We obtain

$$E\{\mathscr{L}_n(\theta_1) - \mathscr{L}_n(\theta_0)\} = \frac{1}{2n}\log\det\Sigma_{\theta_1}\Sigma_{\theta_0}^{-1} + \frac{1}{2n}\operatorname{tr}\left\{\Sigma_f\left(\Sigma_{\theta_1}^{-1} - \Sigma_{\theta_0}^{-1}\right)\right\},$$

which tends with Szegö's identity [cf. Grenander and Szegö (1958), page 64, Section 5.2] and Lemma A.5 to

$$D(\theta_1, f) - D(\theta_0, f) =: c(\theta_1) > 0$$

due to the uniqueness of θ_0. Furthermore,

$$\operatorname{Var}(\mathscr{L}_n(\theta_1) - \mathscr{L}_n(\theta_0)) = \frac{1}{2n^2}\operatorname{tr}\left\{\left[\Sigma_f\left(\Sigma_{\theta_1}^{-1} - \Sigma_{\theta_0}^{-1}\right)\right]^2\right\}$$

tends to 0 with Lemma A.5, which implies

$$\lim_{n\to\infty} P\left(\mathscr{L}_n(\theta_1) - \mathscr{L}_n(\theta_0) < c(\theta_1)/2\right) = 0.$$

Using Lemma A.1, we obtain, with $|\theta - \theta_1| \leq |\theta_2 - \theta_1|$,

$$\begin{aligned}\mathscr{L}_n(\theta_2) &- \mathscr{L}_n(\theta_1)\\ &= \frac{1}{2n}\sum_{i=1}^{k}(\theta_2 - \theta_1)_i\left[\operatorname{tr}\left\{\Sigma_\theta^{-1}\Sigma_n\left(\frac{\partial}{\partial\theta_i}f_\theta\right)\right\} - \mathbf{X}_n'\Sigma_\theta^{-1}\Sigma_n\left(\frac{\partial}{\partial\theta_i}f_\theta\right)\Sigma_\theta^{-1}\mathbf{X}_n'\right].\end{aligned}$$

Lemmas A.1 and A.2 now imply, with $U_\delta(\theta_1) := \{\theta_2 \in \Theta: |\theta_2 - \theta_1| < \delta\}$,

$$\sup_{\theta_2\in U_\delta(\theta_1)}\left|\mathscr{L}_n(\theta_2) - \mathscr{L}_n(\theta_1)\right| \leq K\delta\left\{1 + \frac{1}{n}\mathbf{X}_n'\mathbf{X}_n\right\}.$$

Since $En^{-1}\mathbf{X}_n'\mathbf{X}_n = c(0) = \int_{-\pi}^{\pi} f(\lambda)\,d\lambda$ and $\operatorname{Var}((1/n)\mathbf{X}_n'\mathbf{X}_n) = O(n^{-1})$, it follows that there exists for all $\theta_1 \neq \theta_0$ a $c(\theta_1) > 0$ with

$$\lim_{n\to\infty} P\left(\inf_{\theta_2\in U_\delta(\theta_1)}(\mathscr{L}_n(\theta_2) - \mathscr{L}_n(\theta_0)) \geq c(\theta_1)/4\right) = 1$$

for sufficiently small δ. With a compactness argument we obtain as in Walker (1964) that $\tilde{\theta}_n \to_P \theta_0$.

We now obtain with a Taylor argument

$$\nabla\mathscr{L}_n(\tilde{\theta}_n)_i - \nabla\mathscr{L}_n(\theta_0)_i = \left\{\nabla^2\mathscr{L}_n(\theta_n^{(i)})(\tilde{\theta}_n - \theta_0)\right\}_i,$$

where $|\theta_n^{(i)} - \theta_0| \leq |\tilde{\theta}_n - \theta_0|$, $i = 1, \ldots, k$. If $\tilde{\theta}_n$ is an interior point of Θ, we have $\nabla\mathscr{L}_n(\tilde{\theta}_n) = 0$. If $\tilde{\theta}_n$ lies on the boundary of Θ, then the assumption that θ_0 is in the interior implies $|\tilde{\theta}_n - \theta_0| \geq \delta$ for some $\delta > 0$, that is, we obtain

$$P\left(\sqrt{n}\left|\nabla\mathscr{L}_n(\tilde{\theta}_n)\right| \geq \varepsilon\right) \leq P\left(|\tilde{\theta}_n - \theta_0| \geq \delta\right) \to 0$$

for all $\varepsilon > 0$. Therefore, it is sufficient to prove

$$\nabla^2 \mathscr{L}_n\big(\theta_n^{(i)}\big) \to_P H_f \tag{3.2}$$

and

$$\sqrt{n}\ \nabla \mathscr{L}_n(\theta_0) - Z_n\big(f\ \nabla f_{\theta_0}^{-1}\big) \to_P 0. \tag{3.3}$$

We have with Lemma A.1

$$\nabla \mathscr{L}_n(\theta) = \frac{1}{2n}\operatorname{tr}\{\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\} - \frac{1}{2n}\mathbf{X}_n'\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\Sigma_\theta^{-1}\mathbf{X}_n$$

and

$$\begin{aligned}\nabla^2 \mathscr{L}_n(\theta) = &-\frac{1}{2n}\operatorname{tr}\Big\{\big(\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\big)^2\Big\} + \frac{1}{2n}\operatorname{tr}\{\Sigma_\theta^{-1}\Sigma_n(\nabla^2 f_\theta)\}\\ &+\frac{1}{n}\mathbf{X}_n'\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\Sigma_\theta^{-1}\mathbf{X}_n\\ &-\frac{1}{2}\mathbf{X}_n'\Sigma_\theta^{-1}\Sigma_n(\nabla^2 f_\theta)\Sigma_\theta^{-1}\mathbf{X}_n.\end{aligned}$$

Lemma A.6 implies

$$\begin{aligned}&E\Big(\sqrt{n}\ \nabla \mathscr{L}_n(\theta_0) - Z_n\big(f\ \nabla f_{\theta_0}^{-1}\big)\Big)\\ &\quad= \frac{1}{2\sqrt{n}}\Big(\operatorname{tr}\{\Sigma_{\theta_0}^{-1}\Sigma_\nabla\} - \operatorname{tr}\{\Sigma_f\Sigma_{\theta_0}^{-1}\Sigma_\nabla\Sigma_{\theta_0}^{-1}\}\Big)\\ &\quad= \frac{\sqrt{n}}{4\pi}\int_{-\pi}^{\pi}\big(f(\lambda) - f_{\theta_0}(\lambda)\big)\nabla f_{\theta_0}^{-1}(\lambda)\,d\lambda + o(1)\\ &\quad= \sqrt{n}\ \nabla D(\theta_0, f) + o(1) = o(1).\end{aligned}$$

Furthermore,

$$\begin{aligned}&\operatorname{Var}\Big(\sqrt{n}\ \nabla \mathscr{L}_n(\theta_0) - Z_n\big(f\ \nabla f_{\theta_0}^{-1}\big)\Big)\\ &\quad= \frac{1}{4n}\Bigg[2\operatorname{tr}\Big\{\big(\Sigma_f\Sigma_{\theta_0}^{-1}\Sigma_\nabla\Sigma_{\theta_0}^{-1}\big)^2\Big\} - 2\operatorname{tr}\Big\{\Sigma_f\Sigma_{\theta_0}^{-1}\Sigma_\nabla\Sigma_{\theta_0}^{-1}\Sigma_n\Big(\frac{f}{2\pi}\nabla f_{\theta_0}^{-1}\Big)\Big\}\\ &\qquad -2\operatorname{tr}\Big\{\Sigma_{\theta_0}^{-1}\Sigma_\nabla\Sigma_{\theta_0}^{-1}\Sigma_f\Sigma_n\Big(\frac{f}{2\pi}\nabla f_{\theta_0}^{-1}\Big)\Big\} + \operatorname{tr}\Big\{\Sigma_n\Big(\frac{f}{2\pi}\nabla f_{\theta_0}^{-1}\Big)^2\Big\}\\ &\qquad\quad +\operatorname{tr}\Big\{\Sigma_f^{-1}\Sigma_n\Big(\frac{f}{2\pi}\nabla f_{\theta_0}^{-1}\Big)\Sigma_f\Sigma_n\Big(\frac{f}{2\pi}\nabla f_{\theta_0}^{-1}\Big)\Big\}\Bigg].\end{aligned}$$

Lemma A.5 implies that this tends to 0 which proves (3.3). We only sketch the proof of (3.2). By using the smoothness properties of f_θ, we can prove with Lemmas A.1, A.2 and A.5 that

$$\nabla^2 \mathscr{L}_n\big(\theta_n^{(i)}\big) - \nabla^2 \mathscr{L}_n(\theta_0) \to_P 0$$

[cf. Dahlhaus (1988), proof of Theorem 3.3]. With Lemma A.5 it then follows

$$E\,\nabla^2 \mathscr{L}_n(\theta_0) \to H_f$$

and

$$\operatorname{Var}\bigl(\nabla^2 \mathscr{L}_n(\theta_0)\bigr) = O(n^{-1}),$$

which implies the result. □

We now discuss the question whether the above estimates remain efficient when the model is correctly specified, that is, when $f = f_{\theta_0}$. In this case

$$I(\theta) := \frac{1}{4\pi}\int_{-\pi}^{\pi} (\nabla \log f_\theta(\lambda))(\nabla \log f_\theta(\lambda))'\, d\lambda$$

is the Fisher information matrix. If the model is correct, then the MLE is known to be efficient (this also follows in the same way as in Theorem 3.4). For the other estimates we obtain the following result.

THEOREM 3.4. *Let $\hat{\theta}_n$ be one of the estimates of Theorems* 3.1 *and* 3.2. *Suppose that the conditions of the corresponding theorem hold. If the model is correctly specified ($f = f_{\theta_0}$), then $\hat{\theta}_n$ is efficient for θ_0 if and only if*

$$\frac{1}{4\pi} I(\theta_0)^{-1} \nabla f_{\theta_0}(\lambda)^{-1} = H_f^{-1}\, \nabla K'\bigl(\theta_0, f_{\theta_0}(\lambda), \lambda\bigr)$$

for all $\lambda \in [-\pi, \pi]$.

PROOF. Theorems 4.2 and 4.4 of Davies (1973) imply that the sequence of experiments $\{P_{nh}\colon h \in \mathbb{R}^k\}$, where P_{nh} is the Gaussian distribution of n observations with spectral density $f_{\theta_0 + h/\sqrt{n}}$, is locally asymptotically normal with central sequence

$$\overline{Z}_n = \frac{1}{2\sqrt{n}} I(\theta_0)^{-1}\Bigl(\mathbf{X}_n' \Sigma_n(f_{\theta_0})^{-1} \Sigma_n(\nabla f_{\theta_0}) \Sigma_n(f_{\theta_0})^{-1} \mathbf{X}_n - \operatorname{tr}\Bigl\{\Sigma_n(f_{\theta_0})^{-1} \Sigma_n(\nabla f_{\theta_0})\Bigr\}\Bigr).$$

We therefore have efficiency if and only if $\overline{Z}_n - Z_n(g_{f_{\theta_0}}) \to_P 0$. We have $E(\overline{Z}_n - Z_n(g_{f_{\theta_0}})) = 0$ and, with $\Sigma_0 = \Sigma(f_{\theta_0})$, $I_0 = I(\theta_0)$,

$$E\Bigl(\overline{Z}_n - Z_n(g_{f_{\theta_0}})\Bigr)'\Bigl(\overline{Z}_n - Z_n(g_{f_{\theta_0}})\Bigr) = \frac{1}{4n}\sum_{i=1}^{k} \operatorname{Var}\Biggl(\mathbf{X}_n'\Biggl\{\Sigma_0^{-1}\Sigma_n\Bigl(\bigl(I_0^{-1}\nabla f_{\theta_0}\bigr)_i\Bigr)\Sigma_0^{-1} - \Sigma_0^{-1}\Sigma_n\Bigl(\frac{1}{2\pi}\bigl(g_{f_{\theta_0}}\bigr)_i\Bigr)\Biggr\}\mathbf{X}_n\Biggr),$$

which tends by similar arguments as in the proof of Theorem 2.2 or 3.3 to

$$4\pi\int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^2\left|\frac{1}{4\pi}I_0^{-1}\,\nabla f_{\theta_0}(\lambda)^{-1} - H_f^{-1}\,\nabla K'\big(\theta_0, f_{\theta_0}(\lambda), \lambda\big)\right|^2 d\lambda,$$

where $|\cdot|$ is the Euclidean norm. This implies the result. □

We now give an important class of estimates that are also efficient if the model is correctly specified.

COROLLARY 3.5. *Let* $K(\theta, f(\lambda), \lambda) = \overline{K}(f_\theta(\lambda)/f(\lambda))$, *where* $\overline{K}$ *is three times differentiable with unique minimum at* $x = 1$. *Suppose the model is correctly specified. Then the estimates from Theorems* 3.1 *and* 3.2 *are efficient for* θ_0.

PROOF. Let $c = K''(1)$. Direct calculation gives

$$H_f = \int_{-\pi}^{\pi} \nabla^2 K\big(\theta_0, f_{\theta_0}(\lambda), \lambda\big)\, d\lambda = 4\pi c I(\theta_0)$$

and

$$\nabla K\big(\theta_0, f_{\theta_0}(\lambda), \lambda\big) = c\,\nabla f_{\theta_0}^{-1}(\lambda),$$

which implies the result. □

The above result has been derived directly by Taniguchi (1987), Theorem 5. An example for an estimate that is not efficient when the model is correctly specified will be given in the next section.

4. Discussion, extensions and examples.

Minimizing the linear h-step prediction error. Suppose that we have observed $X_1, \ldots, X_n$ and wish to make a linear prediction of X_{n+h}. If the process is an AR(p)-process, the best linear predictor is given by

$$\hat{X}_{N+h} = -\sum_{j=1}^{p} a_j \hat{X}_{N+h-j},$$

where $\hat{X}_{N+h-j}$ is the best linear predictor of X_{N+h-j} given $X_1, \ldots, X_N$. This means that we can start with (1.1) and calculate $\hat{X}_{N+h}$ iteratively. In particular,

$$\hat{X}_{N+h} = -\sum_{j=1}^{p} c_j X_{N+1-j},$$

where $c_j := c_j(a_1, \ldots, a_p)$ are certain functions of the parameters.

If we proceed as if the process were AR(p), the mean square prediction error is given by [$\theta = (a_1, \ldots, a_p)$]

$$\begin{aligned}\mathrm{PE}_h(\theta, f) &= E\big(X_{N+h} - \hat{X}_{N+h}\big)^2 \\ &= \int_{-\pi}^{\pi} f(\lambda)\left|\exp(i\lambda(h-1)) + \sum_{j=1}^{p} c_j \exp(-i\lambda j)\right|^2 d\lambda.\end{aligned}$$

This is an example for a distance function which is linear in f and different from the Kullback–Leibler distance. Theorems 3.1 and 3.2 imply that $\hat{\theta}_n = \arg\min \mathrm{PE}_h(\theta, \hat{f}_n)$ and $\hat{\theta}'_n = \arg\min \mathrm{PE}_h(\theta, I_n)$ are efficient estimates of $\theta_0 = \arg\min \mathrm{PE}_h(\theta, f)$. As indicated in the discussion prior to Theorem 3.3, the MLE $\tilde{\theta}_n$ will, in general, not even be consistent.

To be specific, let $h = 3$ and $p = 1$. Then $\theta = a_1$, $c_1 = -a_1^3$ and

$$\mathrm{PE}_3(\theta, f) = \int_{-\pi}^{\pi} f(\lambda)(1 - 2\theta^3 \cos 3\lambda + \theta^6)\, d\lambda,$$

which leads with $c(u) = \mathrm{Var}(X_t, X_{t+u})$ to

$$\nabla \mathrm{PE}_3(\theta, f) = \int_{-\pi}^{\pi} f(\lambda)[-6\theta^2 \cos 3\lambda + 6\theta^5]\, d\lambda$$

and

$$\theta_0 = \left[\frac{\int_{-\pi}^{\pi} f(\lambda)\cos(3\lambda)\, d\lambda}{\int_{-\pi}^{\pi} f(\lambda)\, d\lambda}\right]^{1/3} = \left[\frac{c(3)}{c(0)}\right]^{1/3}.$$

The corresponding efficient estimate obtained, for example, by minimizing $\mathrm{PE}_3(\theta, I_n)$ is

$$\hat{\theta}_n = \left[\frac{c_n(3)}{c_n(0)}\right]^{1/3},$$

where

$$c_n(u) = \frac{1}{n}\sum_{t=1}^{n-u} X_t X_{t+u} = \int_{-\pi}^{\pi} I_n(\lambda)\exp(i\lambda u)\, d\lambda$$

is the empirical covariance.

If we take instead $h = 1$ we obtain

$$\theta_0^* = -\frac{c(1)}{c(0)}.$$

If the true process is AR(1), then $\theta_0^* = \theta_0$ while, in general, $\theta_0^* \neq \theta_0$. The MLE $\tilde{\theta}_n$ is an efficient estimate for θ_0^* but not for θ_0.

Since

$$\mathrm{PE}_3(\theta, f) = \int_{-\pi}^{\pi} K(\theta, f(\lambda), \lambda)\, d\lambda,$$

with $K(\theta, x, \lambda) = x[1 - 2\theta^3 \cos 3\lambda + \theta^6]$, we have

$$\nabla K'\big(\theta_0, f_{\theta_0}(\lambda), \lambda\big) = -6\theta_0^2 \cos 3\lambda + 6\theta_0^5.$$

Since

$$\nabla f_{\theta_0}^{-1}(\lambda) = \frac{2\pi}{\sigma^2}(2\cos\lambda + 2\theta_0),$$

the condition of Theorem 3.4 is not fulfilled, and $\hat{\theta}_n$ is therefore *not* efficient if the model is correctly specified.

Note that $\theta_0 = \arg\min \mathrm{PE}_h(\theta, f)$ is not always uniquely determined, that is, Assumption 1.1 may be violated. For example, for $h = 2$ and $p = 1$ only $\theta_0^2 = a_1^2$ is uniquely determined. We conjecture that in this situation $c_n(2)/c_n(0)$ is an efficient estimate of θ_0^2.

From a practical point of view, the situation becomes difficult when the AR(p)-model is "close" to the true f (which usually is the case when the order is selected by an information criterion). Then $\theta_0^* \approx \theta_0$, and it will depend on the (unknown) difference between θ_0^* and θ_0 whether $\hat{\theta}_n$ or $\tilde{\theta}_n$ will lead to the better estimate of θ_0.

The fitting of time series models by minimizing multi-step-ahead prediction errors has recently been discussed under more practical aspects by Haywood and Tunnicliffe Wilson (1993). They also use the frequency domain approach.

Distances between spectral densities. As in Corollary 3.5, Taniguchi (1987) has considered several distances of the form $K(\theta, f(\lambda), \lambda) = \overline{K}(f_\theta(\lambda)/f(\lambda))$. Examples are $\overline{K}(x) = \log x + 1/x$ (Kullback–Leibler distance), $\overline{K}(x) = -\log x + x$ or $\overline{K}(x) = (x^\alpha - 1)^2$.

Taniguchi (1987), Section 4, recommends choosing the distance function $\overline{K}(x)$ dependent on the parameter space to obtain noniterative efficient estimates (e.g., for MA-models $\overline{K} = -\log x + x$ instead of $\log x + 1/x$).

We mention that different $\overline{K}$ lead in the misspecified case to different

$$\theta_0 = \arg\min \int \overline{K}\left(\frac{f_\theta(\lambda)}{f(\lambda)}\right) d\lambda$$

[e.g., for an MA(1)-model $\overline{K}(x) = -\log x + x$ leads to a different θ_0 from $\overline{K}(x) = \log x + 1/x$]. This means that one is estimating efficiently different values of the parameter space. Therefore, the above-mentioned advice has to be handled with care.

Small-sample effects. It is well known that estimates based on the nontapered periodogram $I_n(\lambda)$ have a poor small-sample behavior. The small-sample behavior of $I_n(\lambda)$ and of $\hat{\theta}_n$ may be drastically improved by applying a data taper [cf. Dahlhaus (1988)]. If the data taper stays constant with increasing sample size, the tapered estimates are no longer efficient due to an increase of the asymptotic variance. However, if the proportion of tapered data tends to 0 as $n \to \infty$, the resulting estimates $T(I_n)$ and $T(\hat{f}_n)$ will be efficient as well. This can be proved by suitable modifications of the above results. In order not to complicate the calculations, we have omitted these results.

For linear distance functions we recommend using $T(I_n)$ instead of $T(\hat{f}_n)$ since the convolution in $\hat{f}_n$ may lead to a loss of sharpness of the peaks in $\hat{f}_n$ and therefore also in $f_{T(\hat{f}_n)}$.

APPENDIX

In this Appendix we briefly summarize some properties of matrix norms and Toeplitz matrices [cf. Grenander and Szegö (1958), Davies (1973), Azencott and Dacunha-Castelle (1986), Dzhaparidze (1986) and Taniguchi (1991)]. Furthermore, we prove some convergence results for spectral estimates.

Suppose A is an $n \times n$ matrix. We denote

$$\|A\| = \sup_{x \in \mathbb{C}^n} \frac{|Ax|}{|x|} = \sup_{x \in \mathbb{C}^n} \left(\frac{x^* A^* A x}{x^* x} \right)^{1/2}$$

$$= [\text{maximum characteristic root of } A^* A]^{1/2},$$

where A^* denotes the conjugate transpose of A, and

$$|A| = [\mathrm{tr}(AA^*)]^{1/2}.$$

If A is a real nonnegative symmetric matrix, that is, $A = P'DP$ with $PP' = P'P = I$ and $D = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$, where $\lambda_i \geq 0$, then we define $A^{1/2} = P'D^{1/2}P$, where $D^{1/2} = \mathrm{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\}$. Thus, $A^{1/2}$ is also nonnegative definite and symmetric with $A^{1/2}A^{1/2} = A$. Furthermore, $A^{-1/2} = (A^{1/2})^{-1}$ if A is positive definite.

The following results are well known [see, e.g., Davies (1973), Appendix II, or Graybill (1983), Section 5.6].

LEMMA A.1. *Let A, B be $n \times n$ matrices. Then*:

(a) $|\mathrm{tr}(AB)| \leq |A|\,|B|$,
(b) $|AB| \leq \|A\|\,|B|$,
(c) $|AB| \leq |A|\,\|B\|$,
(d) $\|A\| \leq |A| < \sqrt{n}\,\|A\|$,
(e) $\|AB\| \leq \|A\|\,\|B\|$,
(f) $\|A\| = \|A^*\|$,
(g) $|\mathrm{tr}(A)| \leq \sqrt{n}\,|A|$,
(h) $|x^*Ax| \leq x^*x\|A\|$, $x \in \mathbb{C}^n$,
(i) $\log\det A \leq \mathrm{tr}\{A - I\}$, $A \geq 0$.

Suppose now that the elements of A are continuously differentiable functions of θ. Then:

(j) $$\frac{\partial}{\partial\theta} A^{-1} = -A^{-1}\left(\frac{\partial}{\partial\theta} A\right) A^{-1},$$

(k) $$\frac{\partial}{\partial\theta} \log\det A = \mathrm{tr}\left\{A^{-1}\,\frac{\partial}{\partial\theta} A\right\},$$

(l) $$|A(\theta_1) - A(\theta_2)| \leq \sum_i |\theta_{1i} - \theta_{2i}| \left|\frac{\partial}{\partial\theta_i} A(\bar{\theta})\right| \quad \textit{with a mean value } \bar{\theta},$$

(m) $$\|A(\theta_1) - A(\theta_2)\| \leq \sum_i |\theta_{1i} - \theta_{2i}| \left\|\frac{\partial}{\partial\theta_i} A(\bar{\theta})\right\| \quad \textit{with a mean value } \bar{\theta}.$$

LEMMA A.2. *Suppose h is a real, symmetric function such that there exist constants with $0 < c_1 \le h(\lambda) \le c_2$. Then*

$$\left\|\Sigma_n(h)^{1/2}\right\| \le \sqrt{2\pi c_2} \quad \text{and} \quad \left\|\Sigma_n(h)^{-1/2}\right\| \le 1/\sqrt{2\pi c_1}$$

and, as a consequence,

$$\left\|\Sigma_n(h)\right\| \le 2\pi c_2 \quad \text{and} \quad \left\|\Sigma_n(h)^{-1}\right\| \le 1/(2\pi c_1).$$

Lemma A.2 follows, for example, from Proposition 4.5.3 of Brockwell and Davis (1987).

LEMMA A.3. *Suppose that $g_j \in \mathscr{L}_{p_j}$ with $1 \le p_j \le \infty$, $j = 1, \dots, k$, are symmetric functions with $\sum_{j=1}^k p_j^{-1} \le 1$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{tr}\left\{ \prod_{j=1}^k \Sigma_n(g_j) \right\} = (2\pi)^{k-1} \int_{-\pi}^{\pi} \left\{ \prod_{j=1}^k g_j(\lambda) \right\} d\lambda.$$

Lemma A.3 was first obtained by Grenander and Szegö (1958), Section 8.1. The above version is due to Avram (1988), Theorem 1.

LEMMA A.4. *Suppose $f \in \mathscr{L}_4$ is a real, symmetric function with $f^{-1} \in \mathscr{L}_4$. Then we have, with I the identity matrix,*

$$\text{(A.1)} \qquad \frac{1}{\sqrt{n}} \left| I - \Sigma_n\left(\frac{f}{2\pi}\right)^{1/2} \Sigma_n\left(\frac{f^{-1}}{2\pi}\right) \Sigma_n\left(\frac{f}{2\pi}\right)^{1/2} \right| = o(1),$$

that is, $\Sigma_n(f^{-1}/2\pi)$ is an approximate inverse of $\Sigma_n(f/2\pi)$.

PROOF. Let

$$\Delta_n(x) = \sum_{j=1}^n \exp(-ijx).$$

Since $\int_{-\pi}^{\pi} \Delta_n(x-y)\Delta_n(y-z)\,dy = 2\pi\Delta_n(x-z)$ the square of the left-hand side of (A.1) is equal to

$$\begin{aligned} 1 - \frac{2}{n} \operatorname{tr}\left\{ \Sigma_n\left(\frac{f}{2\pi}\right) \Sigma_n\left(\frac{f^{-1}}{2\pi}\right) \right\} + \frac{1}{n} \operatorname{tr}\left\{ \left(\Sigma_n\left(\frac{f}{2\pi}\right) \Sigma_n\left(\frac{f^{-1}}{2\pi}\right) \right)^2 \right\} \\ = (2\pi)^{-4} n^{-1} \int_{[-\pi,\pi]^4} \left(\frac{f(x_1)}{f(x_2)} - 1 \right) \left(\frac{f(x_3)}{f(x_4)} - 1 \right) \\ \times \Delta_n(x_1 - x_2)\, \Delta_n(x_2 - x_3)\, \Delta_n(x_3 - x_4)\, \Delta_n(x_4 - x_1)\, d\mathbf{x} \\ = \int_{[-\pi,\pi]^3} G(x_1, x_2, x_3)\, \phi_n(x_1, x_2, x_3)\, d\mathbf{x}, \end{aligned}$$

where

$$\phi_n(x_1, x_2, x_3) = (2\pi)^{-3} n^{-1} \Delta_n(x_1)\, \Delta_n(x_2)\, \Delta_n(x_3)\, \Delta_n(-x_1 - x_2 - x_3)$$

and

$$G(x_1, x_2, x_3) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\frac{f(x + x_1 + x_2 + x_3)}{f(x + x_1 + x_2)} - 1 \right) \left(\frac{f(x + x_1)}{f(x)} - 1 \right) dx.$$

G is continuous in 0 with $G(0, 0, 0) = 0$ and ϕ_n is an approximate convolution identity [cf. Dahlhaus (1983), Lemma 3]. This implies the result. □

LEMMA A.5. *Suppose that g_j are real, symmetric functions with $0 < c_1 \le g_j(\lambda) \le c_2$. Let*

$$\sigma_j = \begin{cases} 1, & j \in P_1, \\ -1, & j \in P_{-1}, \end{cases}$$

where $\{P_1, P_{-1}\}$ is a partition of $\{1, \ldots, k\}$. Then we have

$$\lim_{n \to \infty} \frac{1}{n} \left\{ \prod_{j=1}^{k} \Sigma_n \left(\frac{g_j}{2\pi} \right)^{\sigma_j} \right\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \prod_{j=1}^{k} g_j(\lambda)^{\sigma_j} \right\} d\lambda.$$

PROOF. We have

$$\begin{aligned} &\frac{1}{n} \operatorname{tr} \left\{ \prod_{j=1}^{k} \Sigma_n \left(\frac{g_j}{2\pi} \right)^{\sigma_j} - \prod_{j=1}^{k} \Sigma_n \left(\frac{g_j^{\sigma_j}}{2\pi} \right) \right\} \\ &\quad = \sum_{i=1}^{k} \frac{1}{n} \operatorname{tr} \left\{ \left[\prod_{j=1}^{i-1} \Sigma_n \left(\frac{g_j^{\sigma_j}}{2\pi} \right) \right] \left[\Sigma_n \left(\frac{g_i}{2\pi} \right)^{\sigma_i} - \Sigma_n \left(\frac{g_i^{\sigma_i}}{2\pi} \right) \right] \left[\prod_{j=i+1}^{k} \Sigma_n \left(\frac{g_j}{2\pi} \right)^{\sigma_j} \right] \right\} \\ &\quad \le \sum_{i \in P_{-1}} \frac{1}{n} \left| \prod_{j=1}^{i-1} \Sigma_n \left(\frac{g_j^{\sigma_j}}{2\pi} \right) \right| \left\| \Sigma_n \left(\frac{g_i}{2\pi} \right)^{\sigma_i} - \Sigma_n \left(\frac{g_i^{\sigma_i}}{2\pi} \right) \right| \left\| \prod_{j=i+1}^{k} \Sigma_n \left(\frac{g_j}{2\pi} \right)^{\sigma_j} \right\|. \end{aligned}$$

Since $|A^{-1} - B| \le \|A^{-1/2}\|^2 |I - A^{1/2} B A^{1/2}|$, Lemmas A.2, A.3 and A.4 imply that this converges to 0. □

For the expectation of the maximum likelihood estimate in Theorem 3.3, we need the convergence of Lemma A.5 with rate $o(n^{-1/2})$. We state the result as we need it in Theorem 3.3.

LEMMA A.6. *Suppose g, f and f_0 are real, symmetric functions, bounded from above and below with $f, f_0 \in \operatorname{Lip}_\kappa$ with $\kappa > 1/2$. Then*

$$\frac{1}{n} \operatorname{tr} \left\{ \Sigma_n(f_0)^{-1} \Sigma_n(g) \right\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{g(\lambda)}{f_0(\lambda)}\, d\lambda + o(n^{-1/2})$$

and

$$\frac{1}{n}\operatorname{tr}\left\{\Sigma_n(f)\Sigma_n(f_0)^{-1}\Sigma_n(g)\Sigma_n(f_0)^{-1}\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{f(\lambda)g(\lambda)}{f_0(\lambda)^2}\,d\lambda + O(n^{-1/2}).$$

The proof is omitted. It is quite technical and uses calculations in the frequency domain similar to the proof of Lemma A.4. Under stronger conditions the result would follow, for example, from Theorem 2.1.1 of Taniguchi (1991) or from Lemma 4.5 in Azencott and Dacunha-Castelle (1986), Chapter 13.

LEMMA A.7. *Suppose* X_t, $t \in \mathbb{Z}$, *is a fourth-order stationary process with* $EX_t = 0$, *Lipschitz-continuous spectral density* f *and bounded fourth-order spectrum. Under Assumption* 1.4 *we have*:

(i) $E\int_{-\pi}^{\pi}(\hat{f}_n(\lambda) - f(\lambda))^2\,d\lambda = O(m^{-2}) + O(m/n) = o(n^{-1/2})$,
(ii) $P(\sup_\lambda|\hat{f}_n(\lambda) - f(\lambda)| \geq \varepsilon) = o(1)$,
(iii) $\sqrt{n}\int_{-\pi}^{\pi}\psi(\lambda)\{\hat{f}_n(\lambda) - I_n(\lambda)\}\,d\lambda \to_P 0$ *for continuous* ψ,
(iv) $\sqrt{n}\int_{-\pi}^{\pi}\psi(\lambda)\{EI_n(\lambda) - f(\lambda)\}\,d\lambda = o(1)$ *for bounded* ψ,
(v) $\sup_{\theta\in\Theta}|\int_{-\pi}^{\pi}h(\theta,\lambda)\{I_n(\lambda) - f(\lambda)\}\,d\lambda| \to_P 0$ *for* Θ *compact and* h *continuous on* $\Theta \times [-\pi, \pi]$.

PROOF. (i) and (iv) are standard. (iii) is contained in the proof of Theorem 3 in Taniguchi (1987). (v) is proved by approximating $h(\theta, \lambda)$ by the Cesaro sum of its Fourier series [cf. Hannan (1973), Lemma 1]. (ii) follows since

$$\sup_\lambda\left|\hat{f}_n(\lambda) - E\hat{f}_n(\lambda)\right| \leq \frac{1}{2\pi}\sum_{|u|\leq n-1}|c_n(u) - Ec_n(u)|\left|\hat{w}\left(\frac{u}{n}\right)\right|$$

and

$$\operatorname{Var} c_n(u) = O(u^{-1})$$

uniformly in u. □

If we make the stronger assumption that f is differentiable with Lipschitz-continuous derivative, then we get in (i) the stronger result $O(m^{-4}) + O(m/n)$. We then can relax the conditions in Assumption 1.4 to $n^{1/8} \ll m \ll n^{1/2}$.

Acknowledgments. We are grateful to David F. Findley and the referees for correcting some mistakes and suggesting improvements in the presentation.

REFERENCES

ANDERSON, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.* **6** 170–176.

AVRAM, F. (1988). On bilinear forms in Gaussian random variables and Toeplitz matrices. *Probab. Theory Related Fields* **79** 37–45.

AZENCOTT, R. and DACUNHA-CASTELLE, D. (1986). *Series of Irregular Observations. Forecasting and Model Building*. Springer, New York.

BERAN, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463.

BHANSALI, R. J. (1981). Effects of not knowing the order of an autoregressive process on the mean squared error of prediction—I. *J. Amer. Statist. Assoc.* **76** 588–597.

BROCKWELL, P. J. and DAVIS, R. A. (1987). *Time Series: Theory and Methods*. Springer, New York.

DAHLHAUS, R. (1983). Spectral analysis with tapered data. *J. Time Ser. Anal.* **4** 163–175.

DAHLHAUS, R. (1988). Small sample effects in time series analysis: a new asymptotic theory and a new estimate. *Ann. Statist.* **16** 808–841.

DAVIES, R. B. (1973). Asymptotic inference in stationary Gaussian time series. *Adv. in Appl. Probab.* **5** 469–497.

DZHAPARIDZE, K. (1971). On methods for obtaining asymptotically efficient spectral parameter estimates for a stationary Gaussian process with rational spectral density. *Theory Probab. Appl.* **16** 550–554.

DZHAPARIDZE, K. (1984). On asymptotically efficient estimation of spectrum parameters. Preprint. Centrum Wisk. Inform., Amsterdam.

DZHAPARIDZE, K. (1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. Springer, New York.

GINOVYAN, M. S. (1988). Asymptotically efficient nonparametric estimation of functionals of a spectral density having zeros. *Theory Probab. Appl.* **33** 296–303.

GRAYBILL, F. A. (1983). *Metrics with Applications in Statistics*, 2nd ed. Wadsworth, Belmont, CA.

GREENWOOD, P. E. and WEFELMEYER, W. (1990). Efficiency of estimators for partially specified filtered models. *Stochastic Process. Appl.* **36** 353–370.

GREENWOOD, P. E. and WEFELMEYER, W. (1993). Maximum likelihood estimator and Kullback–Leibler information in misspecified Markov chain models. Unpublished manuscript.

GRENANDER, U. and SZEGÖ, G. (1958). *Toeplitz Forms and Their Applications*. Univ. California Press, Berkeley.

HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330.

HANNAN, E. J. (1973). The asymptotic theory of linear time series models. *J. Appl. Probab.* **10** 130–145.

HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1986). Asymptotically efficient nonparametric estimation of functionals of a spectral density function. *Probab. Theory Related Fields* **73** 447–461.

HAYWOOD, J. and TUNNICLIFFE WILSON, G. (1993). Fitting time series models by minimising multi-step ahead errors: a frequency domain approach. Preprint, Lancaster Univ.

HOSOYA, Y. and TANIGUCHI, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *Ann. Statist.* **10** 132–153. [Correction (1993) *Ann. Statist.* **21** 1115–1117.]

HOSOYA, Y. (1989). The bracketing condition for limit theorems on stationary linear processes. *Ann. Statist.* **17** 401–418.

KABAILA, P. V. (1980). An optimality property of the least-squares estimate of the parameter of the spectrum of a purely nondeterministic time series. *Ann. Statist.* **8** 1082–1092.

MALIUKEVIČIUS, R. (1989). Maximum likelihood estimation of the spectral density parameter. *Lithuanian Math. J.* **28** 353–364.

MILLAR, P. W. (1984). A general approach to the optimality of minimum distance estimators. *Trans. Amer. Math. Soc.* **286** 377–418.

OGATA, Y. (1980). Maximum likelihood estimates for incorrect Markov models for time series and the derivation of AIC. *J. Appl. Probab.* **17** 59–72.

PARZEN, E. (1982). Maximum entropy interpretation of autoregressive spectral densities. *Statist. Probab. Lett.* **1** 7–11.

PARZEN, E. (1992). Time series, statistics, and information. In *New Directions in Time Series Analysis, Part I* (D. Brillinger et al., eds.) 265–286. Springer, New York.

PINSKER, M. (1963). *Information and Information Stability of Random Variables*. Holden-Day, San Francisco.

TANIGUCHI, M. (1979). On estimation of parameters of Gaussian stationary processes. *J. Appl. Probab.* **16** 575–591.

TANIGUCHI, M. (1987). Minimum contrast estimation for spectral densities of stationary processes. *J. Roy. Statist. Soc. Ser. B* **49** 315–325.

TANIGUCHI, M. (1991). *Higher Order Asymptotic Theory for Time Series Analysis*. *Lecture Notes in Statist.* Springer, Berlin.

WALKER, A. M. (1964). Asymptotic properties of least squares estimates of the parameters of the spectrum of a stationary non-deterministic time series. *J. Austral. Math. Soc.* **4** 363–384.

WHITTLE, P. (1952). Some results in time series analysis. *Skand. Aktuarietidskr.* **35** 48–60.

WHITTLE, P. (1953). Estimation and information in stationary time series. *Ark. Mat.* **2** 423–434.

INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY

FB 6 MATHEMATIK
UNIVERSITÄT SIEGEN
HÖLDERLINSTRASSE 3
57068 SIEGEN
GERMANY