ASYMPTOTICALLY OPTIMAL ESTIMATION IN MISSPECIFIED TIME SERIES MODELS

By R. Dahlhaus and W. Wefelmeyer

Universität Heidelberg and Universität Siegen

Abstract

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 $X_1, ..., X_n$ 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), i.e. we use the model

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

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

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

(c.f. Brockwell and Davis, 1987, p.170, Example 5.3.1). The mean square prediction error under the true distribution of the process $(a_0 = 1, \theta = (a_1, ..., a_p))$ is

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

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

[0] *Key words and phrases.* Time series, misspecified models, efficiency, minimum distance estimation, maximum likelihood, predicition.

where $A_\theta(\lambda) = \sum_{j=0}^{p} a_j \exp(-i\lambda j)$. Here $PE(\theta, f)$ 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), i.e. we want to estimate

$$\theta_0 = \operatorname*{argmin}_\theta PE(\theta, f).$$

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

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

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 $\hat{\sigma}_n^2 = PE(\hat{\theta}_n, I_n)$; the corresponding theoretical value is $\sigma_0^2 = PE(\theta_0, f)$. 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 (cp. 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; Parzen, 1982, 1992). Thus, the best fit is achieved by

$$\theta_0 = \operatorname*{argmin}_\theta D(\theta, f) \tag{1.3}$$

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

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta} D(\theta, I_n).$$

$\hat{\theta}_n$ is called Whittle estimate (Whittle, 1952). For $AR(p)$-models we have $f_\theta(\lambda) = \frac{\sigma^2}{2\pi}\left|\sum_{j=0}^{p} a_j \exp(-i\lambda j)\right|^{-2}$ (we now set $\theta = (a_1, ..., a_p, \sigma^2)$). 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, p. 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 prediciton error. σ_0^2 and $\hat{\sigma}_n^2$ are now 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 is in general not an efficient estimate for the 'best' approximating parameter if the model is misspecified (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 i.i.d. observations and the Hellinger distance, and by Greenwood and Wefelmeyer (1993) for Markov chains and the Kullback-Leibler distance.

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

$$T(f) := \underset{\theta}{\operatorname{argmin}}\, D(\theta, f)$$

and make the following assumption.

(1.1) Assumption $\Theta \subset \mathbb{R}^k$ *is compact and* $K : \Theta \times (0, \infty) \times [-\pi, \pi] \to \mathbb{R}$ *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 $K(\theta, f(\lambda), \lambda) = \bar{K}\left(\frac{f_\theta(\lambda)}{f(\lambda)}\right)$, where $f_\theta(\lambda)$ is the spectral density of the model. An important distance function which is not of this form in the h-step prediction error (cf. section 4). If we take $\bar{K}\left(\frac{f_\theta(\lambda)}{f(\lambda)}\right)$ as the distance function, then Assumption 1.1 is fulfilled if $\bar{K}(x)$ is three times continuously differentiable with unique minimum at $x = 1$ and the model spectral densities f_θ fulfill

(1.2) Assumption $\Theta \subset \mathbb{R}^k$ *is compact and the model spectral density* $f_\theta(\lambda)$ *is three times continuously differentiable with respect to* θ *with continuous (in* θ *and* λ*) derivatives.*

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 is the case for the prediciton error where σ^2 is not estimated by minimizing a distance function).

The assumptions on the observed process are

(1.3) Assumption $X_t, t \in \mathbb{Z}$*, is a Gaussian stationary process with* $EX_t = 0$ *and spectral density* $f \in Lip_\kappa$ *(*κ *is specified below).*

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

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

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

(1.4) Assumption *$W_n(\alpha) = mW(m\alpha)$, where $n^{1/4} \ll m \ll n^{1/2}$ and W is bounded and non-negative 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), Hosoya (1989).

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).

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 put into an 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 $X_1, ..., X_n$ be a sample form 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 $\theta + \frac{1}{\sqrt{n}} h$, 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 $f(\lambda)\left(1 + \frac{1}{\sqrt{n}} h(\lambda)\right)$, 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 $h(\lambda)$ set

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

Furthermore, let

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

be the Toeplitz matrix of g and $\underline{X}_n = (X_1, ..., X_n)'$. If g is a vector function, $\Sigma_n(g)$ is the corresponding vector of matrices. For example, by $\|\Sigma_n(\nabla f_\theta)\|$ we mean

$$\left(\sum_{i=1}^{k} \left\| \Sigma_n \left(\frac{\partial}{\partial \theta_i} f_\theta \right) \right\|^2 \right)^{1/2}$$

where $\|A\|$ is the spectral norm of A (cp. the appendix).

The following result is due to Dzhaparidze (1986).

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

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

where

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

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

$$Z_n(h) \xrightarrow{\mathcal{D}} N(0, \langle h, h \rangle),$$

i.e. the sequence P_{nh} is LAN.

PROOF. See Dzhaparidze (1986), p.64, Section I.3, Theorem 4(3) and p. 155, Section II, Theorem A1.2. □

The following result is also due to Dzhaparidze (1986), p.64, Section I.3, Theorem 4(4). We prove it under slightly different conditions.

(2.2) Theorem *Suppose h is bounded and $f \in Lip_\kappa$ with $\kappa > 1/2$. Then*

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

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 $\int_{-\pi}^{\pi} I_n(\lambda) g(\lambda) d\lambda = \frac{1}{2\pi n} \underline{X}_n' \Sigma_n(g) \underline{X}_n$ the variance of the above expression is equal to

$$\frac{1}{n} \mathrm{var} \underline{X}_n' \left\{ \Sigma_n \left(\frac{h}{4\pi f} \right) - \Sigma_n(f)^{-1} \Sigma_n \left(\frac{h}{4\pi} \right) \right\} \underline{X}_n.$$

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

$$\frac{2}{n} \Bigg[\mathrm{tr} \left\{ \Sigma_n \Big(\frac{h}{4\pi f} \Big) \Sigma_f \Sigma_n \Big(\frac{h}{4\pi f} \Big) \Sigma_f \right\} - \mathrm{tr} \left\{ \Sigma_n \Big(\frac{h}{4\pi f} \Big) \Sigma_n \Big(\frac{h}{4\pi} \Big) \Sigma_f \right\}$$
$$- \mathrm{tr} \left\{ \Sigma_n \Big(\frac{h}{4\pi} \Big) \Sigma_f \Sigma_n \Big(\frac{h}{4\pi f} \Big) \right\} + \mathrm{tr} \left\{ \Sigma_n \Big(\frac{h}{4\pi} \Big) \Sigma_n \Big(\frac{h}{4\pi} \Big) \right\} \Bigg]$$

which tends to zero with 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 $f\left(1+\frac{1}{\sqrt{n}}h\right)$ on the basis of a sample $X_1, ..., X_n$. 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 $K(\theta, x) = K(\theta, x, \cdot)$

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

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

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

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, Theorem 1(b) and Theorem 2 in the special case $K(\theta, f(\lambda), \lambda) = \bar{K}\left(\frac{f_\theta(\lambda)}{f(\lambda)}\right)$.)

(2.3) Theorem *Suppose that f_n is a sequence with $\|f_n - f\| \to 0$ and $0 < C_1 \leq f_n, f \leq C_2$. 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

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

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

Proof. (i) 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

$$\begin{aligned} T(f_n) - T(f) + O(|T(f_n) - T(f)|\,\|f_n - f\|) + O(|T(f_n) - T(f)|^2) \\ = \frac{1}{4\pi}\int_{-\pi}^{\pi} g_f(\lambda)\frac{f_n(\lambda) - f(\lambda)}{f(\lambda)}d(\lambda) + 0(\|f_n - f\|^2) \end{aligned}$$

which implies with part (i) the result. □

As a consequence we obtain

(2.4) Corollary *Let h be bounded and $f_{nh} = f\left(1 + \frac{1}{\sqrt{n}}h\right)$. 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:

$$\sqrt{n}(T_n - T(f_{nh})) \xrightarrow{\mathcal{D}} L \text{ for } h \text{ bounded.}$$

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:

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

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:

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

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 $T = (T_1, ..., T_k)'$ be a finite-dimensional functional of the spectral density. If g_j is the gradient of T_j, then $g_f = (g_1, ..., g_k)'$ 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 (1.4). As seen in Section 2, proving efficiency means proving the stochastic approximation (2.4).

(3.1) Theorem *Suppose Assumptions 1.1, 1.3 (with $\kappa = 1$) and 1.4 hold and H_f is nonsingular. Then we have*

$$\sqrt{n}(T(\hat{f}_n) - T(f)) - Z_n(g_f) \xrightarrow{P} 0,$$

i.e. $T(\hat{f}_n)$ is an efficient estimate of $T(f)$.

Proof. We start by proving consistency. We cannot apply Theorem 2.3 directly, since $\hat{f}_n(\lambda)$ is not necessarily bounded. However, if $\sup_\lambda |\hat{f}_n(\lambda) - f(\lambda)| < m/2$ we know that $\hat{f}_n$ is bounded. We therefore obtain from Theorem 2.3 (1) that there exists for all $\varepsilon > 0$ a $\delta > 0$ with

$$P(|T(\hat{f}_n) - T(f)| \geq \varepsilon) \leq P(\|\hat{f}_n - f\| \geq \delta) + P(\sup_\lambda |\hat{f}_n(\lambda) - f(\lambda)| \geq m/2)$$

which implies consistency with Lemma A.7. Since $P(\sup_\lambda |\hat{f}_n(\lambda) - f(\lambda)| \geq m/2) \to 0$ we obtain as in the proof of Theorem 2.3

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

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

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

(cp. also 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).

(3.2) Theorem *Suppose Assumptions 1.1 and 1.3 (with $\kappa > 1/2$) hold, where $K(\theta, x, \lambda) = a_\theta(\lambda) + b_\theta(\lambda)x$. If H_f is nonsingular then*

$$\sqrt{n}(T(I_n) - T(f)) - Z_n(g_f) \xrightarrow{P} 0$$

i.e. $T(I_n)$ is an efficient estimate of $T(f)$. Here

$$H_f = \int_{-\pi}^{\pi} \left(\nabla^2 a_{T(f)} + \nabla^2 b_{T(f)} f\right)$$

and

$$g_f = -4\pi H_f^{-1} \nabla b_{T(f)} f.$$

Proof. Lemma A.7(v) implies that

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

Since

$$D(T(I_n), I_n) \leq D(T(f), I_n)$$

and

$$D(T(f), f) \leq D(T(I_n), f),$$

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 $|\widetilde{T(f)} - T(f)| \leq |T(I_n) - T(f)|$:

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

Since

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

$$\int_{-\pi}^{\pi}\nabla^2 K(\widetilde{T(f)},f) \xrightarrow{P} H_f \qquad \text{(smoothness of } K\text{)},$$

and

$$\int_{-\pi}^{\pi}\nabla^2 b_{\widetilde{T(f)}}(I_n-f)\xrightarrow{P} 0 \qquad \text{(Lemma A.7(v))}$$

the result is proved. □

In the next theorem we consider the Kullback-Leibler distance $D(\theta, f)$ as in (1.3) (i.e. $a_\theta(\lambda) = \frac{1}{4\pi}\log f_\theta(\lambda)$ and $b_\theta(\lambda) = \frac{1}{4\pi}f_\theta^{-1}$ with $b_\theta(\lambda)$ as in Theorem 3.2). We then have

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

As a consequence of Theorem 2.1 and Theorem 3.2 we now obtain for the Whittle estimate $\hat\theta_n = T(I_n)$

$$\sqrt{n}(\hat\theta_n-\theta_0)\xrightarrow{\mathcal{D}} W\left(0,\frac{1}{4\pi}H_f^{-1}\left\{\int_{-\pi}^{\pi} f^2(\nabla f_{\theta_0}^{-1})(\nabla f_{\theta_0}^{-1})'d\lambda\right\}H_f^{-1}\right),$$

a result already proved by Taniguchi (1979). However, we now know that this limit variance is the smallest which can be achieved under model misspecification.

We now study the behavior of the Gaussian maximum likelihood estimate $\tilde\theta_n$ of the fitted model, i.e.

$$\tilde\theta_n = \operatorname{argmin}\mathcal{L}_n(\theta),$$

where

$$\begin{aligned}\mathcal{L}_n(\theta) &= -\frac{1}{n}\log\ \text{likelihood}\\ &= \frac{1}{2}\log(2\pi)+\frac{1}{2n}\log|\Sigma_n(f_\theta)|+\frac{1}{2}\underline{X}_n'\Sigma_n(f_\theta)^{-1}\underline{X}_n.\end{aligned}$$

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 $h \geq 2$ (cp. Section 4).

(3.3) Theorem *Suppose θ_0 is the unique solution of (1.3) and lies in the interior of Θ. Suppose further that Assumption 1.2 and 1.3 (with $\kappa > 1/2$) hold and in addition $f_{\theta_0} \in Lip_\kappa, \kappa > 1/2$. If H_f is nonsingular, then we have with $g_f = -H_f^{-1} f \nabla f_{\theta_0}^{-1}$*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) - Z_n(g_f) \xrightarrow{P} 0,$$

i.e. 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 $\mathcal{L}_n(\theta)$ 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

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

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

$$E\left\{\mathcal{L}_n(\theta_1) - \mathcal{L}_n(\theta_0)\right\} = \frac{1}{2n} \log |\Sigma_{\theta_1} \Sigma_{\theta_0}^{-1}| + \frac{1}{2n} \mathrm{tr}\left\{\Sigma_f(\Sigma_{\theta_1}^{-1} - \Sigma_{\theta_0}^{-1})\right\}$$

which tends with Szegö's identity (cf. Grenander and Szegö, 1958, p.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,

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

tends to zero with Lemma A.5 which implies

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

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

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

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

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

Since $E\frac{1}{n}\underline{X}_n'\underline{X}_n = c(0) = \int_{-\pi}^{\pi} f(\lambda)d\lambda$ and $\mathrm{var}\left(\frac{1}{n}\underline{X}_n'\underline{X}_n\right) = O(T^{-1})$, it follows that there exists for all $\theta_1 \neq \theta_0$ a $c(\theta_1) > 0$ with

$$\lim_{n\to\infty} P\Big(\inf_{\theta_2\in U_\delta(\theta_1)} (\mathcal{L}_n(\theta_2) - \mathcal{L}_n(\theta_o)) \geq c(\theta_1)/4\Big) = 1$$

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

We now obtain with a Taylor argument

$$\nabla\mathcal{L}_n(\tilde{\theta}_n)_i - \nabla\mathcal{L}_n(\theta_0)_i = \Big\{\nabla^2\mathcal{L}_n(\theta_n^{(i)})(\tilde{\theta}_n - \theta_0)\Big\}_i$$

where $|\theta_n^{(i)} - \theta_0| \leq |\tilde{\theta}_n - \theta_0| \, (i = 1, ..., k)$. If $\tilde{\theta}_n$ is an interior point of Θ we have $\nabla\mathcal{L}_n(\tilde{\theta}_n) = 0$. 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 $\delta > 0$, i.e. we obtain

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

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

$$\nabla^2\mathcal{L}_n(\theta_n^{(i)}) \xrightarrow{P} H_f \tag{3.2}$$

and

$$\sqrt{n}\nabla\mathcal{L}_n(\theta_0) - Z_n\left(f\nabla f_{\theta_0}^{-1}\right) \xrightarrow{P} 0. \tag{3.3}$$

We have with Lemma A.1

$$\nabla\mathcal{L}_n(\theta) = \frac{1}{2n}\mathrm{tr}\left\{\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\right\} - \frac{1}{2n}\underline{X}_n'\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\Sigma_\theta^{-1}\underline{X}_n$$

and

$$\begin{aligned}\nabla^2\mathcal{L}_n(\theta) \;=\; & -\frac{1}{2n}\mathrm{tr}\left\{\left(\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\right)^2\right\} + \frac{1}{2n}\mathrm{tr}\left\{\Sigma_\theta^{-1}\Sigma_n(\nabla^2 f_\theta)\right\} \\ & + \frac{1}{n}\underline{X}_n'\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\Sigma_\theta^{-1}\Sigma_n(\nabla f_\theta)\Sigma_\theta^{-1}\underline{X}_n \\ & - \frac{1}{2}\underline{X}_n'\Sigma_\theta^{-1}\Sigma_n(\nabla^2 f_\theta)\Sigma_\theta^{-1}\underline{X}_n.\end{aligned}$$

Lemma A.6 implies

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

Furthermore,

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

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

$$\nabla^2\mathcal{L}_n(\theta_n^{(i)}) - \nabla^2\mathcal{L}_n(\theta_0) \xrightarrow{P} 0$$

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

$$E\nabla^2\tilde{\mathcal{L}}_n(\theta_0) \to H_f$$

and

$$\operatorname{var}(\nabla^2\mathcal{L}_n(\theta_0)) = O(n^{-1})$$

which implies the result. □

We now discuss the question whether the above estimates remain efficient when the model is correctly specified, i.e. 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

(3.4) Theorem *Let $\hat{\theta}_n$ be one of the estimates of Theorem 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*

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

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

Proof. Theorem 4.2 and Theorem 4.4 of Davies (1973) imply that the sequence of experiments $\{P_{n,h} : h \in \mathbb{R}^k\}$, 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

$$\bar{Z}_n = \frac{1}{2\sqrt{n}} I(\theta_0)^{-1} \left(\underline{X}_n' \Sigma_n(f_{\theta_0})^{-1} \Sigma_n(\nabla f_{\theta_0}) \Sigma_n(f_{\theta_0})^{-1} \underline{X}_n - \operatorname{tr}\left\{ \Sigma_n(f_{\theta_0})^{-1} \Sigma_n(\nabla f_{\theta_0}) \right\} \right).$$

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

$$\begin{aligned} & E\left(\bar{Z}_n - Z_n(g_{f_{\theta_0}})\right)' \left(\bar{Z}_n - Z_n(g_{f_{\theta_0}})\right) \\ = \; & \frac{1}{4n} \sum_{i=1}^{k} \operatorname{var}\left(\underline{X}_n' \left\{ \Sigma_0^{-1} \Sigma_n\Big((I_0^{-1} \nabla f_{\theta_0})_i \Big) \Sigma_0^{-1} - \Sigma_0^{-1} \Sigma_n\Big(\frac{1}{2\pi} (g_{f_{\theta_0}})_i \Big) \right\} \underline{X}_n \right) \end{aligned}$$

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

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

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.

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

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

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

and

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

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 $X_1, ..., X_n$ 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 $X_1, ..., X_N$. 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 $c_j := c_j(a_1, ..., a_p)$ are certain functions of the parameters.

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

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

This is an example for a distance function which is linear in f and different from the Kullback-Leibler distance. Theorem 3.1 and 3.2 imply that $\hat{\theta}_n = \operatorname{argmin} PE_h(\theta, \hat{f}_n)$ and $\hat{\theta}'_n = \operatorname{argmin} PE_h(\theta, I_n)$ are efficient estimates of $\theta_0 = \operatorname{argmin} PE_h(\theta, f)$. 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 $h = 2$ and $p = 1$. Then $\theta = a_1$, $c_1 = -a_1^2$, and

$$PE_2(\theta, f) = \int_{-\pi}^{\pi} f(\lambda)(1 - 2\theta^2 \cos 2\lambda + \theta^4) d\lambda$$

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

$$\nabla PE_2(\theta, f) = \int_{-\pi}^{\pi} f(\lambda) \left[-4\theta \cos 2\lambda + 4\theta^3\right] d\lambda$$

and

$$\theta_0 = \left[\frac{\int_{-\pi}^{\pi} f(\lambda) cos(2\lambda) d\lambda}{\int_{-\pi}^{\pi} f(\lambda) d\lambda}\right]^{1/2} = \left[\frac{c(2)}{c(0)}\right]^{1/2}.$$

The corresponding efficient estimate obtained e.g. by minimizing $PE_2(\theta, I_n)$ is

$$\hat{\theta}_n = \left[\frac{c_n(2)}{c_n(0)}\right]^{1/2}$$

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

$$PE_2(\theta, f) = \int_{-\pi}^{\pi} K(\theta, f(\lambda), \lambda)d\lambda$$

with $K(\theta, x, \lambda) = x[1 - 2\theta^2\cos 2\lambda + \theta^4]$, we have

$$\nabla K'(\theta_0, f_{\theta_0}(\lambda), \lambda) = -4\theta_0\cos 2\lambda + 4\theta_0^3.$$

Since

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

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

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.

<u>Distances between spectral densities</u>

As in Corollary 3.5 Taniguchi (1987) has considered several distances of the form $K(\theta, f(\lambda), \lambda) = \bar{K}\left(\frac{f_\theta(\lambda)}{f(\lambda)}\right)$. Examples are $\bar{K}(x) = \log x + \frac{1}{x}$ (Kullback-Leibler distance), $\bar{K}(x) = -\log x + x$ or $\bar{K}(x) = (x^\alpha - 1)^2$.

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

We mention that different $\bar{K}$ lead in the misspecified case to different $\theta_0 = \operatorname{argmin} \int \bar{K}\left(\frac{f_\theta(\lambda)}{f(\lambda)}\right) d\lambda$ (e.g. for an $MA(1)$-model $\bar{K}(x) = -\log x + x$ leads to another θ_0 as $\bar{K}(x) = \log x + \frac{1}{x}$). 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 (cp. 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 zero 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 (cp. also Grenander und Szegö, 1955; Davies, 1973; Azencott and Dacunha-Castelle, 1986; Dzhaparidze, 1986; Taniguchi, 1991). Furthermore, we prove some convergence results for spectral estimates.

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

$$\begin{aligned} \|A\| &= \sup_{x \in \mathbb{C}^n} \frac{|Ax|}{|x|} = \sup_{x \in \mathbb{C}^n} \left(\frac{x^* A^* A x}{x^* x}\right)^{1/2} \\ &= [\text{maximum characteristic root of } A^* A]^{1/2}, \end{aligned}$$

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

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

If A is a real nonnegative symmetric matrix, i.e., $A = P'DP$ with $PP' = P'P = I$ and $D = \operatorname{diag}\{\lambda_1, ..., \lambda_n\}$, where $\lambda_i \geq 0$, then we define $A^{1/2} = P'D^{1/2}P$, where

$D^{1/2} = \text{diag}\{\sqrt{\lambda_1}, ..., \sqrt{\lambda_n}\}$. 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].

(A.1) Lemma *Let A, B be $n \times n$ matrices. Then*

(a) $|\text{tr}(AB)| \leq |A||B|$,

(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) $|\text{tr}(A)| \leq \sqrt{n}|A|$,

(h) $|x^*Ax| \leq x^*x\|A\|$, $x \in \mathbb{C}^n$,

(i) $\log\det A \leq \text{tr}\{A - I\}$, $A \geq 0$.

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

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

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

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

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

and, as a consequence,

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

Proof. Since

$$x^*\Sigma_n(h)x = \int_{-\pi}^{\pi} h(\lambda)\left|\sum_{t=1}^{n} x_t \exp(-i\lambda t)\right|^2 d\lambda$$

we obtain the result. □

(A.3) Lemma *Suppose that $g_j \in \mathcal{L}_{p_j}$ with $1 \leq p_j \leq \infty \quad (j = 1, ..., k)$ are symmetric functions with $\sum_{j=1}^k p_j^{-1} \leq 1$. Then*

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

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

(A.4) Lemma *Suppose $f \in \mathcal{L}_4$ is a real symmetric function with $f^{-1} \in \mathcal{L}_4$. Then we have*

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

i.e. $\Sigma_n\left(\frac{f^{-1}}{2\pi}\right)$ is an approximate inverse of $\Sigma_n\left(\frac{f}{2\pi}\right)$.

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

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

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. □

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

$$\sigma_j = \begin{cases} 1, & j \in I_1 \\ -1, & j \in I_{-1} \end{cases}$$

where $\{1,...,k\} = I_1 + I_{-1}$. 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

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

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

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.

(A.6) Lemma *Suppose g, f and f_o are real symmetric functions, bounded from above and below with $f, f_o \in \mathrm{Lip}_\kappa$ with $\kappa > 1/2$. Then*

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

and

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

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 e.g. follow from Theorem 2.1.1 of Taniguchi (1991) or from Lemma 4.5 in Azencott and Dacunha-Castelle (1986, Chapter XIII).

(A.7) Lemma *Suppose X_t, $t \in \mathbb{Z}$ is a fourth order stationary process with $EX_t = 0$, Lipschitz- continuous spectral density $f(\lambda)$ and bounded fourth order spectrum. Under Assumption 1.2 we have*

(i) $E \int_{-\pi}^{\pi} \left(\hat{f}_n(\lambda) - f(\lambda) \right)^2 d\lambda = O(m^{-2}) + O\left(\frac{m}{n}\right) = o(n^{-1/2})$,

(ii) $P\left(\sup_\lambda |\hat{f}_n(\lambda) - f(\lambda)| \geq \varepsilon \right) = o(1)$,

(iii) $\sqrt{n} \int_{-\pi}^{\pi} \psi(\lambda) \left\{ \hat{f}_n(\lambda) - I_n(\lambda) \right\} d\lambda \xrightarrow{P} 0$ *for* $\psi(\lambda)$ *continuous,*

(iv) $\sqrt{n} \int_{-\pi}^{\pi} \psi(\lambda) \left\{ EI_n(\lambda) - f(\lambda) \right\} d\lambda = o(1)$ *for* $\psi(\lambda)$ *bounded,*

(v) $\sup_{\theta \in \Theta} \left| \int_{-\pi}^{\pi} h_\theta(\lambda) \left\{ I_n(\lambda) - f(\lambda) \right\} d\lambda \right| \xrightarrow{P} 0$ *for* Θ *compact and* h_θ *continuous on* $\Theta \times [-\pi, \pi]$.

Proof. (i) and (iv) are standard. (iii) is contained in the proof of Theorem 3 in Taniguchi (1987). (iv) is proved by approximating $h_\theta(\lambda)$ by the Cesaro sum of its Fourier series (cp. 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

$$\text{var} c_n(u) = O(u^{-1})$$

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 $O(m^{-4}) + O\left(\frac{m}{n}\right)$. We then can relax the conditions in Assumption 1.2 to $n^{1/8} \ll m \ll n^{1/2}$.

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. Th. Rel. Fields 79, 37-45.

Azencott, R. and Dacunha-Castelle, D. (1986). Series of Irregular Observations. Forecasting and Model Building. Springer- Verlag. New York.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. Ann. Statist. 5, 445-463.

Brockwell, P.J. and Davis, R.A. (1987). Time Series: Theory and Methods. Springer-Verlag, 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. 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. Theor. Probab. Appl. 16, 550-554.

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

Ginovyan, M.S. (1988). Asymptotically efficient nonparametric estimation of functionals of a spectral density having zeros. Theor. Probab. Appl. 33, 296-303.

Graybill, F.A. (1983). Metrics with Applications in Statistics, 2nd ed. Wedsworth, Belmont, Calif.

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. Preprint.

Grenander, U. and Szegö, G. (1958). Toeplitz Forms and their Applications. University of 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 University.

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.

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

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-Verlag, 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. Prob. 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 Statistics, Springer-Verlag, 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.

Whitle, P. (1952). Some results in time series analysis. Skand. Aktuarietidskr. 35, 48-60.