Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: http://www.tandfonline.com/loi/gnst20

Maximum likelihood estimation and model selection for locally stationary processes

R. Dahlhaus

To cite this article: R. Dahlhaus (1996) Maximum likelihood estimation and model selection for locally stationary processes , Journal of Nonparametric Statistics, 6:2-3, 171-191, DOI: 10.1080/10485259608832670

To link to this article: https://doi.org/10.1080/10485259608832670

Published online: 12 Apr 2007.

Nonparametric Statistics, Vol. 6, pp. 171–191
Reprints available directly from the publisher
Photocopying permitted by license only

MAXIMUM LIKELIHOOD ESTIMATION AND MODEL SELECTION FOR LOCALLY STATIONARY PROCESSES

R. DAHLHAUS

*Universität Heidelberg**

The Gaussian maximum likelihood estimate is investigated for time series models that have locally a stationary behaviour (e.g. for time varying autoregressive models). The asymptotic properties are studied in the case where the fitted model is either correct or misspecified. For example the behaviour of the maximum likelihood estimate is explained in the case where a stationary model is fitted to a nonstationary process. As a general model selection criterion the AIC is considered. It can for example automatically select between stationary models, nonstationary models and deterministic trends.

KEY WORDS: Gaussian likelihood, asymptotic properties, AIC, deterministic trends.

1. INTRODUCTION

Stationarity has always played a major role in time series analysis. One reason is that for stationary processes there exists a rich and elegant theory which allows for a detailed investigation of the different methods used in statistical inference. As a consequence practitioners very often try to use stationary methods even when the data clearly show a nonstationary behaviour, e.g., by taking differences or by looking at different segments separately.

One major difficulty in developing a general nonstationary theory is the problem of asymptotics. On the one hand an asymptotic theory is needed since an investigation of e.g., a maximum likelihood estimate for a fixed sample size is much too complicated and will not lead to any satisfactory results. On the other hand a classical asymptotic theory with the assumption that more and more observations of the future become available does not make sense since future observations of a general nonstationary process do not necessarily contain any information on the structure at present.

To overcome this problem we have suggested in Dahlhaus (1993) an asymptotic approach similar to nonparametric regression. Suppose for example that we observe

$$X_t = g(t) X_{t-1} + \varepsilon_t \text{ with } \varepsilon_t \text{ iid } \mathcal{N}(0, \sigma^2)$$

for $t = 1, \ldots, T$ where $g(t) = a + bt + ct^2$ with $|g(t)| < 1$ for $t \in [1, T]$. It is easy to construct different estimates for the parameters (e.g. a least squares estimate, a

* This work was supported by the Deutsche Forschungsgemeinschaft

maximum likelihood estimate, or a fit of $g(t)$ to locally estimated AR(1)-parameters). Therefore, also the need of describing the properties of such estimates arises. An asymptotic theory where T tends to infinity obviously makes no sense since $g(t)$ will finally leave the stability region $|g(t)| < 1$ and the future behaviour of the process therefore will be different from the behaviour at present. Analogously to non-parametric regression it therefore seems more natural to set down the asymptotic theory in a way that we "observe" $g(t)$ on a finer grid (but on the same interval), i.e. that we observe the process

$$X_{t,T} = g\left(\frac{t}{T}\right) X_{t-1,T} + \varepsilon_t \tag{1.1}$$

(where g is now rescaled to the interval $[0,1]$).

Letting T tend to infinity now means that we have in the sample $X_{1,T}, \ldots, X_{T,T}$ more and more "observations" for each value of g (if e.g. $g(u) = \chi_{(1/2,1]}(u)$ we have $T/2$ observations, both for the white noise situation ($g = 0$) and for the random walk case ($g = 1$)). Letting T tend to infinity no longer means extending the data to the future. Note, that it also does *not* mean that we sample more densely from a continuous time signal. Instead, it is an abstract setting for statistical inference leading e.g. to valuable results on the behaviour of our estimates. Note, that in the stationary case where g is constant $X_{t,T}$ is independent of T, and our asymptotics are identical to the classical asymptotics for stationary processes. Therefore, the asymptotic normality of the classical MLE for stationary processes follows as a special case from Theorem 2.4 below.

By using this asymptotic approach we investigate in this paper e.g. the behaviour of the MLE for a model of the form (1.1) where $g(u) = g_\theta(u)$ depends on a finite dimensional parameter. However, our goal is to derive results for more general parametric nonstationary models including models for the trend.

To define a more general class of nonstationary processes which includes the above example we may try to take the time varying spectral representation

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} \exp(i\lambda t)\, A\left(\frac{t}{T}, \lambda\right) d\xi(\lambda). \tag{1.2}$$

(similar to the analogous representation for stationary processes). However, it turns out that the equation (1.1) has not exactly but only approximately a solution of the form (1.2). We therefore only require that (1.2) holds approximately which leads to the following definition.

DEFINITION 1.1. A sequence of stochastic processes $X_{t,T}$ ($t = 1, \ldots, T$; $T \geqslant 1$) is called locally stationary with transfer function A° and trend μ if there exists a representation

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} \exp(i\lambda t)\, A^\circ_{t,T}(\lambda)\, d\xi(\lambda) \tag{1.3}$$

where

(i) $\xi(\lambda)$ is a stochastic process on $[-\pi,\pi]$ with $\overline{\xi(\lambda)} = \xi(-\lambda)$ and

$$\operatorname{cum}\{d\xi(\lambda_1),\ldots,d\xi(\lambda_k)\} = \eta\left(\sum_{j=1}^{k}\lambda_j\right)h_k(\lambda_1,\ldots\lambda_{k-1})\,d\lambda_1\ldots d\lambda_k$$

where $\operatorname{cum}\{\ldots\}$ denotes the cumulant of k-th order, $h_1 = 0, h_2(\lambda) = 1, |h_k(\lambda_1,\ldots,\lambda_{k-1})| \leqslant \text{const}_k$ for all k and $\eta(\lambda) = \sum_{j=-\infty}^{\infty}\delta(\lambda + 2\pi j)$ is the period 2π extension of the Dirac delta function.

(ii) There exists a constant K and a 2π-periodic function $A:[0,1]\times\mathbb{R}\to\mathbb{C}$ with $A(u,-\lambda) = \overline{A(u,\lambda)}$ and

$$\sup_{t,\lambda}\left|A_{t,T}^{\circ}(\lambda) - A\left(\frac{t}{T},\lambda\right)\right| \leqslant KT^{-1} \tag{1.4}$$

for all T. $A(u,\lambda)$ and $\mu(u)$ are assumed to be continuous in u.

The smoothness of A in u guarantees that the process has locally a "stationary behaviour". Below we will require additional smoothness properties of A and μ.

Furthermore, we assume in this paper that the process $X_{t,T}$ is Gaussian, i.e., that $h_k(\lambda) = 0$ for all $k \geqslant 3$.

In the following we will denote by s and t always time points in the interval $[1,T]$ while u and v will denote time points in the rescaled interval $[0,1]$, i.e., $u = t/T$.

To illustrate the idea behind locally stationary processes we repeat some examples from Dahlhaus (1993).

EXAMPLES 1.2

(i) Suppose Y_t is a stationary process and $\mu,\sigma:[0,1]\to\mathbb{R}$ are continuous. Then

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \sigma\left(\frac{t}{T}\right)Y_t$$

is locally stationary with $A_{t,T}^{\circ}(\lambda) = A(t/T,\lambda)$. If Y_t is an AR (2)-process with (complex) roots close to the unit circle then Y_t shows a periodic behaviour and σ may be regarded as a time varying amplitude function of the process $X_{t,T}$. If T tends to infinity more and more cycles of the process with $u = t/T \in [u_0 - \varepsilon, u_0 + \varepsilon]$, i.e., with amplitude close to $\sigma(u_0)$ are observed.

(ii) Suppose ε_t is an iid sequence and

$$X_{t,T} = \sum_{j=0}^{\infty} a_j\left(\frac{t}{T}\right)\varepsilon_{t-j}$$

Then $X_{t,T}$ is locally stationary with $A_{t,T}^{\circ}(\lambda) = A(t/T,\lambda) = \sum_{j=0}^{\infty} a_j(t/T)\exp(-i\lambda j)$.

(iii) Autoregressive processes with time varying coefficients are locally stationary. This was proved in Dahlhaus (1995, Theorem 2.3). However, in this case we only have (1.4) instead of $A_{t,T}^{\circ}(\lambda) = A(t/T,\lambda)$.

$f(u,\lambda) := |A(u,\lambda)|^2$ is called the time-varying spectral density of the process. In Dahlhaus (1995) we have proved that $f(u,\lambda)$ is under certain conditions uniquely

determined by the process $X_{t,T}$.

$$c(u,k) := \int_{\pi}^{\pi} f(u,\lambda)\exp(i\lambda k)\,d\lambda$$

is the local covariance of lag k at time u. We have

$$\begin{aligned}&\operatorname{cov}\left(X_{[uT],T}, X_{[uT]+k,T}\right)\\ &= \int_{-\pi}^{\pi} \exp(i\lambda k)\, A^{\circ}_{[uT]+k,T}(\lambda)\, A^{\circ}_{[uT],T}(-\lambda)\,d\lambda\\ &= c(u,k) + O(T^{-1})\end{aligned}$$

uniformly in u and k (if e.g., $A(u,\lambda)$ is uniformly differentiable in u).

Strictly speaking local stationarity is a property of the whole triangular array $X_{t,T}(t = 1,\ldots,T;\ T \geqslant 1)$ and $f(u,\lambda)$ is only uniquely determined by the whole array (or by $X_{t,T}$ $t = 1,\ldots,T$; as T tends to infinity). In a practical situation we observe only one process $X_1,\ldots,X_T$ with a fixed T. As pointed out above our approach is an abstraction which helps to explain e.g., the behaviour of estimates in this situation.

We frequently use the spectral density as a tool in this paper (e.g., in the specification of the asymptotic covariance matrix or in our proofs). However, the parametric models we have in mind are mainly models in the time domain (cp. 3.1). For specific models such as time varying autoregressions all results can probably be formulated and proved without using the spectral representation.

We want to mention that evolutionary spectral representations have first been considered by Priestley (1965). However, our approach is different.

In Section 2 we prove consistency and asymptotic normality of the MLE in Gaussian locally stationary models. Furthermore, we discuss the issue of model misspecification and consider the AIC as a model selection criterion for nonstationary models. In Section 3 state space representations for time varying ARMA-models are discussed. In Section 4 we prove efficiency of the estimate by showing that the sequence of experiments is locally asymptotically normal. For the proofs we need several results on norms and traces of products of covariance-matrices. These results are proved in Dahlhaus (1995) and summarized in the appendix.

2. THE ASYMPTOTIC BEHAVIOUR OF THE MAXIMUM LIKELIHOOD ESTIMATE

Suppose we fit a Gaussian locally stationary time series model with transfer function A°_{θ} and mean function μ_{θ} to the observed data $\underline{X} = (X_{1,T},\ldots,X_{T,T})'$ (e.g., a time varying AR-model–cp. Section 3) and estimate the parameters of the model by maximizing the likelihood, i.e. we consider

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta} \mathscr{L}_T(\theta)$$

where

$$\mathscr{L}_T(\theta) = -\frac{1}{T}\text{Gaussian log likelihood}$$
$$= \frac{1}{2}\log(2\pi) + \frac{1}{2T}\log\det\sum_\theta + \frac{1}{2T}(X-\mu_\theta)'\sum_\theta^{-1}(X-\mu_\theta).$$
$$\sum_\theta = \left\{\int_{-\pi}^{\pi}\exp(i\lambda(r-s))\,A^{\circ}_{\theta,r,T}(\lambda)\,A^{\circ}_{\theta,s,T}(-\lambda)\,d\lambda\right\}_{r,s=1,\ldots,t}$$

is the model covariance matrix for the observations and $\mu_\theta = (\mu_\theta(1/T),\ldots,\mu_\theta(T/T))'$ is the mean vector. An important special case is where the parameter separates, i.e., where $\theta = (\tau,\nu)$ with $A^{\circ}_\theta = A^{\circ}_\tau$ and $\mu_\theta = \mu_\nu$. This will be investigated in Remark 2.5.

In Theorem 2.3 and 2.4 we study the convergence of $\hat{\theta}_T$. If the model is correct, i.e., the observations are coming from a locally stationary process with true transfer function $A^{\circ}_{\theta_0}$ and mean function μ_{θ_0}, then Theorem 2.3 shows that $\hat{\theta}_T \to \theta_0$ in probability. If the model is not correct, then $\hat{\theta}_T$ will converge to

$$\theta_0 := \arg\min \mathscr{L}(\theta)$$

where

$$\mathscr{L}(\theta) := \lim_{T\to\infty} E\,\mathscr{L}_T(\theta).$$

It is shown in Dahlhaus (1995, Theorem 3.4) that

$$\mathscr{L}(\theta) = \frac{1}{4\pi}\int_0^1\left\{\int_{-\pi}^{\pi}\left[\log 4\pi^2 f_\theta(u,\lambda) + \frac{f(u,\lambda)}{f_\theta(u,\lambda)}\right]d\lambda + \frac{(\mu_\theta(u)-\mu(u))^2}{f_\theta(u,0)}\right\}du$$

where $f(u,\lambda) = |A(u,\lambda)|^2, f_\theta(u,\lambda) = |A_\theta(u,\lambda)|^2$ are the time varying spectral densities of the true process and the model process, respectively, and $\mu(u)$ and $\mu_\theta(u)$ are the true mean function and the model mean function. In the case of model misspecification θ_0 is the value that gives theoretically the best fit in the sense of the Kullback Leibler information divergence.

In the case where the model is correctly specified, i.e., where $A(u,\lambda) = A_{\theta^*}(u,\lambda)$, $\mu(u) = \mu_{\theta^*}(u)$ with some $\theta^* \in \Theta$ we get from the inequality $x \geq 1 + \log x$

$$\mathscr{L}(\theta) \geq \frac{1}{4\pi}\int_0^1\int_{\pi}^{\pi}(\log 4\pi^2 f(u,\lambda) + 1)\,d\lambda\,du$$

with equality if and only if $f(u,\lambda) = f_\theta(u,\lambda)$ and $\mu(u) = \mu_\theta(u)$, i.e., we obtain $\theta_0 = \theta^*$ and the uniqueness of θ_0 if θ^* can be uniquely identified from f_θ and μ_θ. Let

$$\nabla_i = \frac{\partial}{\partial\theta_i}, \nabla = (\nabla_1,\ldots,\nabla_p)', \nabla_{ij} = \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j} \quad \text{and} \quad \nabla^2 = (\nabla_{ij})_{ij=1,\ldots,p}.$$

We will prove the results under the following assumptions.

Assumption 2.1

(i) We observe a realisation $X_{1,T},\ldots,X_{T,T}$ of a locally stationary Gaussian process with true mean function μ and transfer function A° and fit a class of locally

stationary Gaussian processes with mean function μ_θ and transfer function $A_\theta, \theta \in \Theta \subset \mathbb{R}^p, \Theta$ compact.

(ii) $\theta_0 = \arg\min \mathcal{L}(\theta)$ exists uniquely and lies in the interior of Θ.

(iii) There exists a constant K with

$$\sup_{t,\lambda} \left| \tilde{\nabla} \left\{ A_{\theta,t,T}(\lambda) - A_\theta\left(\frac{t}{T}, \lambda\right) \right\} \right| \leq KT^{-1}$$

where $\tilde{\nabla} = \nabla_i$ or $\tilde{\nabla} = \nabla_{ij}$. The $A_\theta(u,\lambda)$ are uniformly bounded from above and below. The components of $A_\theta(u,\lambda), \nabla A_\theta(u,\lambda)$ and $\nabla^2 A_\theta(u,\lambda)$ are differentiable in u and λ with uniformly continuous derivatives $\partial/\partial u\, \partial/\partial\lambda$. The same holds for the true $A(u,\lambda)$.

(iv) The components of $\mu(u)$, $\nabla\mu_\theta(u)$ and $\nabla^2\mu_\theta(u)$ are differentiable in u with uniformly continuous derivatives.

(v) $A_{\theta_0}(u,\lambda)$ is twice differentiable in u with uniformly bounded derivative.

The above conditions can be relaxed. For example, condition (v) is only needed if the model is not correct. For the proof of Theorem 2.3 we need much less. However, we do not want to complicate the paper by another set of conditions.

We set

$$\sum_T(A,B) = \left\{ \int_{-\pi}^{\pi} \exp(i\lambda(r-s))\, A^{\circ}_{r,T}(\lambda)\, B^{\circ}_{s,T}(-\lambda)\, d\lambda \right\}_{r,s=1,\ldots,T},$$

i.e., we have $\sum_\theta = \sum_T(A_\theta, A_\theta)$. Furthermore, let

$$\sum = \sum_T(A,A)$$

be the true covariance matrix of the process and

$$C^{(i)}_\theta := \nabla_i \sum_\theta = \sum_T(\nabla_i A_\theta, A_\theta) + \sum_T(A_\theta, \nabla_i A_\theta).$$

By $\|A\|$ we denote the spectral norm and by $|A|$ the Euclidean norm of a matrix (cp. the appendix); $\|\underline{\mu}\|_2$ is the Euclidean norm of a vector.

We start by proving a frequently used lemma.

LEMMA 2.2. Suppose Assumption 2.1(i)–(iv) holds. Let B_T be a sequence of $T \times T$ matrices with uniformly bounded norm $\|B_T\|$. Then

$$Y_T := \frac{1}{T}(\underline{X} - \underline{\mu}_\theta)' B_T (\underline{X} - \underline{\mu}_\theta)$$

is bounded in probability. If $B_T = \sum_\theta^{-1}$ then we have $\operatorname{var} Y_T = O(T^{-1})$.

Proof. Lemma A.1 (*h*) and Lemma A.4 imply

$$\frac{1}{T}(\underline{X} - \underline{\mu}_\theta)' B_T (\underline{X} - \underline{\mu}_\theta) \leq \frac{1}{T}(\underline{X} - \underline{\mu}_\theta) \sum^{-1} (\underline{X} - \underline{\mu}_\theta).$$

Furthermore,

$$\frac{1}{T}(X-\mu_\theta)\sum^{-1}(\underline{X}-\underline{\mu}_\theta)=\frac{1}{T}(\underline{X}-\underline{\mu})'\sum^{-1}(\underline{X}-\underline{\mu})$$

$$+\frac{2}{T}(\underline{X}-\underline{\mu})'\sum^{-1}(\underline{\mu}-\underline{\mu}_\theta)+\frac{1}{T}(\underline{\mu}-\underline{\mu}_\theta)'\sum^{-1}(\underline{\mu}-\underline{\mu}_\theta).$$

Since the first two terms of the right hand side are independent for a Gaussian $\underline{X}$ and

$$E\frac{1}{T}(\underline{X}-\underline{\mu})'\sum^{-1}(\underline{X}-\underline{\mu})=1,$$

$$\operatorname{var}\frac{1}{T}(\underline{X}-\underline{\mu})'\sum^{-1}(\underline{X}-\underline{\mu})=\frac{2}{T},$$

$$\operatorname{var}\frac{1}{T}(\underline{X}-\underline{\mu})'\sum^{-1}(\underline{\mu}-\underline{\mu}_\theta)=\frac{1}{T^2}(\underline{\mu}-\underline{\mu}_\theta)'\sum^{-1}(\underline{\mu}-\underline{\mu}_\theta)$$

the first result follows from Lemma A.5. If $B_T=\sum_\theta{}^{-1}$ we have the same decomposition (with $\sum_\theta$ instead of $\sum$) and

$$\operatorname{var}\frac{1}{T}(\underline{X}-\underline{\mu})'\sum_\theta^{-1}(\underline{X}-\underline{\mu})=\frac{2}{T}\operatorname{tr}\left\{\sum\sum_\theta^{-1}\sum\sum_\theta^{-1}\right\}$$

$$\operatorname{var}\frac{1}{T}(\underline{X}-\underline{\mu})\sum_\theta^{-1}(\underline{\mu}-\underline{\mu}_\theta)'=\frac{1}{T^2}(\underline{\mu}-\underline{\mu}_\theta)'\sum_\theta^{-1}\sum\sum_\theta^{-1}(\underline{\mu}-\underline{\mu}_\theta).$$

The second result now follows again with Lemma A.5.

THEOREM 2.3. Suppose that Assumption 2.1(i) – (iv) holds. Then

$$\hat{\theta}_T\to\theta_0$$

in probability.

Proof. The basic idea is taken from Walker (1964), Section 2. We have

$$E\,\mathscr{L}_T(\theta)\to\mathscr{L}(\theta)$$

and from Lemma 2.2

$$\operatorname{var}\mathscr{L}_T(\theta)=O(T^{-1}).$$

Therefore,

$$\mathscr{L}_T(\theta)\to\mathscr{L}(\theta)$$

in probability. Since θ_0 is assumed to be unique it follows that for all $\theta_1\neq\theta_0$ there exists a constant $c(\theta_1)>0$ with

$$\lim_{T\to\infty}P(\mathscr{L}_T(\theta_1)-\mathscr{L}_T(\theta_0)<c(\theta_1))=0.$$

Furthermore, we have with a mean value $\bar{\theta}$

$$\mathscr{L}_T(\theta_2)-\mathscr{L}_T(\theta_1)=(\theta_2-\theta_1)'\nabla\mathscr{L}_T(\bar{\theta})$$

where

$$\nabla \mathscr{L}_T(\theta)_i = \frac{1}{2T}\operatorname{tr}\left\{\sum_\theta^{-1} C_\theta^{(i)}\right\} - \frac{1}{2T}(\underline{X}-\underline{\mu}_\theta)'\sum_\theta^{-1} C_\theta^{(i)}\sum_\theta^{-1}(\underline{X}-\underline{\mu}_\theta) \tag{2.1}$$

$$-\frac{1}{T}(\nabla_i \underline{\mu}_\theta)'\sum_\theta^{-1}(\underline{X}-\underline{\mu}_\theta)$$

$$= -\frac{1}{2T}(\underline{X}-\underline{\mu})'\sum_\theta^{-1} C_\theta^{(i)}\sum_\theta^{-1}(\underline{X}-\underline{\mu}) \tag{2.2}$$

$$-\frac{1}{T}[(\nabla_i \underline{\mu}_\theta)' + (\underline{\mu}-\underline{\mu}_\theta)'\sum_\theta^{-1} C_\theta^{(i)}]\sum_\theta^{-1}(\underline{X}-\underline{\mu})$$

$$+\text{const.}$$

with a constant independent of $\underline{X}$ (but dependent of T.) With the Cauchy-Schwarz inequality and Lemma A.1(h) we get

$$\frac{1}{T}(\nabla_i \underline{\mu}_\theta)'\sum_\theta^{-1}(\underline{X}-\underline{\mu}_\theta)$$

$$\leqslant \frac{1}{T}\left\{(\nabla_i \underline{\mu}_\theta)'\sum_\theta^{-1}(\nabla_i \underline{\mu}_\theta)\cdot(\underline{X}\quad \underline{\mu}_\theta)'\sum_\theta^{-1}(\underline{X}-\underline{\mu}_\theta)\right\}^{1/2}$$

$$\leqslant \left\{\frac{1}{T}\|\nabla_i \underline{\mu}_\theta\|_2^2\right\}^{1/2}\left\{\frac{2}{T}\|\underline{X}\|_2^2 + \frac{2}{T}\|\underline{\mu}_\theta\|_2^2\right\}^{1/2}\|\sum_\theta^{-1}\|$$

which by Assumption 2.1 and Lemma A.4 is uniformly bounded by

$$K + K\frac{1}{T}\|\underline{X}\|_2^2.$$

Similarly, we can estimate the other terms in (2.1) which leads to

$$\sup_{\theta_2 \in U_\delta(\theta_1)} |\mathscr{L}_T(\theta_2) - \mathscr{L}_T(\theta_1)|$$

$$\leqslant K\delta\left(1 + \frac{1}{T}\underline{X}'\underline{X}\right)$$

with some constant K. With Lemma 2.2 we now obtain that there exists for all $\theta_1 \neq \theta_0$ a $c(\theta_1) > 0$ with

$$\lim_{T\to\infty} P\left(\inf_{\theta_2 \in U_\delta(\theta_1)} \mathscr{L}_T(\theta_2) - \mathscr{L}_T(\theta_0) \geqslant c(\theta_1)/2\right)$$

$$\geqslant 1 - \lim_{T\to\infty} P\left(\mathscr{L}_T(\theta_1) - \mathscr{L}_T(\theta_0) < c(\theta_1)\right)$$

$$- \lim_{T\to\infty} P\left(\sup_{\theta_2 \in U_\delta(\theta_1)} |\mathscr{L}_T(\theta_2) - \mathscr{L}_T(\theta_1)| \geqslant c(\theta_1)/2\right)$$

for sufficiently small δ. A compactness argument as in Walker (1964) implies the result.

THEOREM 2.4. Suppose that Assumption 2.1 holds. Then we have

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Gamma^{-1} V \Gamma^{-1})$$

with

$$\Gamma = \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} (f - f_{\theta_0}) \nabla^2 f_{\theta_0}^{-1} \, d\lambda \, du + \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} (\nabla \log f_{\theta_0})(\nabla \log f_{\theta_0})' \, d\lambda \, du$$
$$+ \frac{1}{4\pi}\int_0^1 \nabla^2 \frac{(\mu_{\theta_0}(u) - \mu(u))^2}{f_{\theta_0}(u,0)} \, du$$

$$V = \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} f^2 \nabla f_{\theta_0}^{-1} \nabla f_{\theta_0}^{-1} \, d\lambda \, du$$
$$+ \frac{1}{2\pi}\int_0^1 f(u,0) \left(\nabla \frac{\mu_{\theta_0}(u) - \mu(u)}{f_{\theta_0}(u,0)}\right)\left(\nabla \frac{\mu_{\theta_0}(u) - \mu(u)}{f_{\theta_0}(u,0)}\right)' du.$$

Proof. We obtain with the mean value theorem

$$\nabla \mathcal{L}_T(\hat{\theta}_T)_i \quad \nabla \mathcal{L}_T(\theta_0)_i = \{\nabla^2 \mathcal{L}_T(\theta_T^{(i)})(\hat{\theta}_T - \theta_0)\}_i$$

with $|\theta_T^{(i)} - \theta_0| \leqslant |\hat{\theta}_T - \theta_0| (i = 1, \ldots p)$. If $\hat{\theta}_T$ lies in the interior of Θ, we have $\nabla \mathcal{L}_T(\hat{\theta}_T) = 0$. If $\hat{\theta}_T$ lies on the boundary of Θ, then the assumption that θ_0 is in the interior implies $|\hat{\theta}_T - \theta_0| \geqslant \delta$ for some $\delta > 0$, i.e., we obtain $P(\sqrt{N}|\nabla \mathcal{L}_T(\hat{\theta}_T)| \geqslant \varepsilon) \leqslant P(|\hat{\theta}_T - \theta_0| \geqslant \delta) \to 0$ for all $\varepsilon > 0$. Thus, the result follows if we prove

$$\nabla^2 \mathcal{L}_T(\theta_T^{(i)}) - \nabla^2 \mathcal{L}_T(\theta_0) \xrightarrow{p} 0 \tag{i}$$

$$\nabla^2 \mathcal{L}_T(\theta_0) \xrightarrow{p} \Gamma \tag{ii}$$

$$\sqrt{T} \nabla \mathcal{L}_T(\theta_0) \xrightarrow{\mathcal{L}} N(0, V). \tag{iii}$$

With $D_\theta^{(i,j)} := \partial^2 / \partial\theta_i \partial\theta_j \Sigma_\theta$ we obtain from (2.1)

$$\nabla^2 \mathcal{L}_T(\theta)_{ij} = -\frac{1}{2T}\mathrm{tr}\left\{\sum_\theta^{-1} C_\theta^{(i)} \sum_\theta^{-1} C_\theta^{(j)}\right\} + \frac{1}{2T}\mathrm{tr}\left\{\sum_\theta^{-1} D_\theta^{(i,j)}\right\} \tag{2.3}$$
$$+ \frac{1}{2T}(\underline{X} - \underline{\mu}_\theta)' \left(\nabla_{ij}^2 \sum_\theta^{-1}\right)(\underline{X} - \underline{\mu}_\theta)$$
$$+ \frac{1}{T}(\nabla_i \underline{\mu}_\theta)' \sum_\theta^{-1} (\nabla_j \underline{\mu}_\theta)$$
$$- \frac{1}{T}(\nabla_i \underline{\mu}_\theta)' \left(\nabla_j \sum_\theta^{-1}\right)(\underline{X} - \underline{\mu}_\theta)$$
$$- \frac{1}{T}(\nabla_j \underline{\mu}_\theta)' \left(\nabla_i \sum_\theta^{-1}\right)(\underline{X} - \underline{\mu}_\theta)$$
$$- \frac{1}{T}(\nabla_{ij}^2 \underline{\mu}_\theta)' \sum_\theta^{-1} (\underline{X} - \underline{\mu}_\theta).$$

(i) To prove (i), i.e., to estimate the difference $\nabla^2 \mathscr{L}_T(\theta_T^{(i)}) - \nabla^2 \mathscr{L}_T(\theta_0)$ we have to consider the above terms separately. The difference of the first two terms is with $\theta_1 - \theta_T^{(i)}$ less than

$$\left|\frac{1}{2T}\operatorname{tr}\left\{\left(\sum_{\theta_0}^{-1} - \sum_{\theta_1}{}^{-1}\right) C_{\theta_0}^{(i)} \sum_{\theta_0}{}^{-1} C_{\theta_0}^{(j)}\right\}\right|$$

$$+ \left|\frac{1}{2T}\operatorname{tr}\left\{\sum_{\theta_1} (C_{\theta_0}^{(i)} - C_{\theta_1}^{(i)}) \sum_{\theta_0}{}^{-1} C_{\theta_0}^{(j)}\right\}\right| + 2 \text{ similar terms}$$

$$\leqslant \frac{1}{2} \| \sum_{\theta_0}^{-1} \| \, \| \sum_{\theta_1}^{-1} - \sum_{\theta_0} \| \, \| \sum_{\theta_T}^{-1} \| \, \| \sum_{\theta_0}^{-1} \| \, \| C_{\theta_0}^{(j)} \|$$

$$+ \frac{1}{2} \| \sum_{\theta_1}^{-1} \| \, \| C_{\theta_0}^{(i)} - C_{\theta_1}^{(i)} \| \, \| \sum_{\theta_0}^{-1} \| \, \| C_{\theta_0}^{(j)} \| + 2 \text{ similar terms.}$$

The matrix norms are uniformly bounded in θ by Lemma A.4. Furthermore, $\| \sum_{\theta_1} - \sum_{\theta_0} \| \leqslant 2\pi \sup_{u,\lambda} \| A_{\theta_T}(u,\lambda)|^2 - |A_{\theta_0}(u,\lambda)|^2| + o(1)$ which tends to zero in probability. The same holds for $\| C_{\theta_0}^{(i)} - C_{\theta_1}^{(i)} \|$ which implies that the difference of the first two terms tends to zero. The same holds for the second two terms.

The remaining terms of (2.3) can all be written as sums of expressions of the form

$$\frac{1}{T} \underline{X}' A_\theta \underline{X}, \quad \frac{1}{T} \underline{y}_\theta{}' A_\theta \underline{X} \quad \text{or} \quad \frac{1}{T} \underline{y}_{1\theta}' A_\theta \underline{y}_{2\theta} \tag{2.4}$$

with $\| A_{\theta_T} - A_{\theta_0} \| \to 0$ and $1/T \| \underline{y}_{\theta_T} - \underline{y}_{\theta_0} \|_2^2 \to 0$ in probability (this follows as above). Furthermore, $\| A_\theta \|$ and $1/T \| \underline{y}_\theta \|_2^2$ are uniformly bounded. This implies for example with the Cauchy-Schwarz inequality

$$\left|\frac{1}{T} \underline{y}_{\theta_1}' A_{\theta_1} \underline{X} - \frac{1}{T} \underline{y}_{\theta_0}' A_{\theta_0} \underline{X}\right|$$

$$\leqslant \frac{1}{T} |(\underline{y}_{\theta_T} - \underline{y}_{\theta_0})' A_{\theta_T} \underline{X}| + \frac{1}{T} |\underline{y}_{\theta_0}' (A_{\theta_1} - A_{\theta_0}) \underline{X}|$$

$$\leqslant \frac{1}{T} \{ \| \underline{y}_{\theta_T} - \underline{y}_{\theta_0} \|_2^2 \, \| \underline{X} \|_2^2 \}^{1/2} \| A_{\theta_1} \| + \frac{1}{T} \{ \| \underline{y}_{\theta_0} \|_2^2 \, \| \underline{X} \|_2^2 \}^{1/2} \| A_{\theta_1} - A_{\theta_0} \|$$

Lemma 2.2 implies that $1/T \| \underline{X} \|_2^2$ is bounded in probability. Therefore, the above expression tends to zero in probability. The other two expressions of (2.4) can be handled similarly, which implies (i).

(ii) It follows as in the proof of Lemma 2.2 that

$$\operatorname{var} \nabla^2 \mathscr{L}_T(\theta_0)_{ij} = O(T^{-1}).$$

To calculate $E \nabla^2 \mathscr{L}_T(\theta_0)_{ij}$ we consider again all terms of (2.3) separately. The expectation of the first three terms of (2.3) is

$$-\frac{1}{2T} \operatorname{tr}\left\{\sum_{\theta_0}^{-1} C_{\theta_0}^{(i)} \sum_{\theta_0}^{-1} C_{\theta_0}^{(j)}\right\} + \frac{1}{2T} \operatorname{tr}\left\{\sum_{\theta_0}^{-1} D_{\theta_0}^{(i,j)}\right\} \tag{2.5}$$

$$-\frac{1}{2T}\operatorname{tr}\left\{\sum\sum_{\theta_0}^{-1} D_{\theta_0}^{(i,j)} \sum_{\theta_0}^{-1}\right\}$$

$$+\frac{1}{T}\operatorname{tr}\left\{\sum\sum_{\theta_0}^{-1} C_{\theta_0}^{(i)} \sum_{\theta_0}^{-1} C_{\theta_0}^{(j)} \sum_{\theta_0}^{-1}\right\}$$

$$+\frac{1}{2T}(\underline{\mu}-\underline{\mu}_\theta)'\left(\nabla_{ij}^2 \sum_{\theta_0}^{-1}\right)(\underline{\mu}-\underline{\mu}_{\theta_0}).$$

The first four terms of this tend with Lemma A.5 to

$$\frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\left\{\left(\frac{1}{f_{\theta_0}}-\frac{f}{f_{\theta_0}^2}\right)\left(\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}f_{\theta_0}\right)+\left(\frac{2f}{f_{\theta_0}^3}-\frac{1}{f_{\theta_0}^2}\right)\left(\frac{\partial}{\partial\theta_i}f_{\theta_0}\right)\left(\frac{\partial}{\delta\theta_j}f_{\theta_0}\right)\right\}d\lambda\,du.$$

Since

$$\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}f_{\theta_0}^{-1}=\frac{2}{f_{\theta_0}^3}\left(\frac{\partial}{\partial\theta_i}f_{\theta_0}\right)\left(\frac{\partial}{\partial\theta_j}f_{\theta_0}\right)-\frac{1}{f_{\theta_0}^2}\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}f_{\theta_0}$$

this is equal to the first and the second summand of Γ_{ij}. Similarly, it can be shown with Lemma A.5 that the last term of (2.5) and the expectation of the last four terms of (2.3) converge to the last summand of Γ_{ij}, which proves (ii).

(iii) We use the method of cumulants. We have

$$0=\nabla\mathscr{L}(\theta_0)=\frac{1}{4\pi}\int_0^1\left\{\int_{-\pi}^{\pi}(f-f_{\theta_0})\nabla f_{\theta_0}^{-1}\,d\lambda+(\mu_{\theta_0}(u)-\mu(u))^2\nabla f_{\theta_0}^{-1}(u,0)\right.$$

$$\left.+2(\mu_{\theta_0}(u)-\mu(u))(\nabla\mu_{\theta_0}(u))f_{\theta_0}^{-1}(u,0)\right\}du.$$

It follows from (2.1) and Lemma A.5 that $E\,\mathscr{L}_T(\theta_0)$ converges to the same expression. Therefore, we have with Lemma A.5

$$\sqrt{T}E\nabla\mathscr{L}_T(\theta_0)=o(1).$$

(Note, that this is the only part of the paper where we need the stronger rate of Lemma A.5 and Assumption 2.1(v). If the model is correctly specified then $E\nabla\mathscr{L}_T(\theta_0)=0$ and we can omit this condition). Furthermore, we get from (2.2)

$$T\operatorname{cov}(\nabla\mathscr{L}_T(\theta_0)_i,\nabla\mathscr{L}_T(\theta_0)_j)$$

$$=\frac{1}{2T}\operatorname{tr}\left\{\sum_{\theta_0}^{-1}C_{\theta_0}^{(i)}\sum_{\theta_0}^{-1}\sum\sum_{\theta_0}^{-1}C_{\theta_0}^{(j)}\sum_{\theta_0}^{-1}\sum\right\}$$

$$+\frac{1}{T}\left[(\nabla_i\underline{\mu}_{\theta_0})'+(\underline{\mu}-\underline{\mu}_{\theta_0})'\sum_{\theta_0}^{-1}C_{\theta_0}^{(i)}\right]\sum_{\theta_0}^{-1}\sum\sum_{\theta_0}^{-1}\left[(\nabla_j\underline{\mu}_{\theta_0})+C_{\theta_0}^{(j)}\sum_{\theta_0}^{-1}(\underline{\mu}-\underline{\mu}_{\theta_0})\right].$$

Lemma A.5 implies that this tends to V_{ij}.

To study the higher-order cumulants we see from (2.1) that $\nabla \mathscr{L}_T(\theta_0)_i$ can be written as

$$\nabla \mathscr{L}_T(\theta_0)_i = -\frac{1}{2T} \underline{Y}' A_i \underline{Y} - \frac{1}{T} \underline{v}_i' B \underline{Y} + \text{const.}$$

where $E\underline{Y} = 0$. The cumulants of order $\geqslant 3$ of the $1/T\underline{v}_i' B\underline{Y}$-terms are zero, while the mixed cumulants of the $1/(2T)\underline{Y}' A_i \underline{Y}$ and $1/T\underline{v}_i' B \underline{Y}$-terms are nonzero if and only if there are exactly two $1/T\underline{v}_i' B\underline{Y}$-terms involved (this follows from the product theorem for cumulants, cf. Brillinger, 1975, Theorem 2.3.2, $E\underline{Y} = 0$, and the normality of $\underline{Y}$).

Therefore we obtain with the product term for cumulants

$$\begin{aligned} &T^{\ell/2}\operatorname{cum}(\nabla \mathscr{L}_T(\theta_0)_{i_1}, \ldots, \nabla \mathscr{L}_T(\theta_0)_{i_\ell}) \\ &= \frac{1}{2} T^{-\ell/2}(-1)^\ell \frac{1}{\ell} \sum_{\substack{(j_1,\ldots,j_\ell) \\ \text{permutation of} \\ (i_1,\ldots,i_\ell)}} \operatorname{tr}\left[\prod_{k=1}^{\ell} \left\{ \Sigma_{\theta_0}^{-1} C_{\theta_0}^{(j_k)} \Sigma_{\theta_0}^{-1} \Sigma \right\} \right] \\ &\quad + \frac{1}{2} T^{-\ell/2}(-1)^\ell \sum_{\substack{(j_1,\ldots,j_\ell) \\ \text{permutation of} \\ (i_1,\ldots,i_\ell)}} \underline{v}_{j_1}' B \Sigma \left(\prod_{k=2}^{\ell} \left\{ \Sigma_{\theta_0}^{-1} C_{\theta_0}^{(j_k)} \Sigma_{\theta_0}^{-1} \Sigma \right\} \right) B \underline{v}_{j_\ell}'. \end{aligned}$$

Lemma A.5 implies that all terms are of order $O(T^{-\ell/2+1})$. Therefore, the theorem is proved.

Remark 2.5. We may also regard $\hat{\theta}_T$ as an estimate of $\theta_T := \arg\min L_T(\theta)$ where $L_T(\theta) = E\mathscr{L}_T(\theta)$. Then we have the same central limit theorem for $\sqrt{T}(\hat{\theta}_T - \theta_T)$. The proof is completely analogous to the above proof. However, we do not need Assumption 2.1(*v*).

We now discuss several special cases.

Remark 2.6. (correctly specified models). If the model is correctly specified, i.e., $f = f_{\theta_0}$ and $\mu = \mu_{\theta_0}$ then we have $V = \Gamma$ with

$$\Gamma = \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} (\nabla \log f_{\theta_0})(\nabla \log f_{\theta_0})' \, d\lambda \, du + \frac{1}{2\pi} \int_0^1 (\nabla \mu_{\theta_0}(u))(\nabla \mu_{\theta_0}(u))' f_{\theta_0}^{-1}(u,0) \, du.$$

In this situation the estimate $\hat{\theta}_T$ is asymptotically efficient. This is proved in Section 4. If the model is correct and in addition the parameters separate, i.e. if $\theta = (\kappa, \tau, \nu)$ with $f_\theta(u,\lambda) = \sigma_\kappa^2(u)/2\pi \, h_\tau(u,\lambda)$, $\mu_\theta(u) = \mu_\nu(u)$ and

$$\int_{-\pi}^{\pi} \log f_\theta(u,\lambda) \, d\lambda = 2\pi \log \frac{\sigma_\kappa^2(u)}{2\pi} \tag{2.6}$$

then $V = \Gamma$ and Γ is diagonal with

$$\Gamma = \begin{pmatrix} \Gamma_{\kappa\kappa} & 0 & 0 \\ 0 & \Gamma_{\tau\tau} & 0 \\ 0 & 0 & \Gamma_{\nu\nu} \end{pmatrix}$$

where

$$\Gamma_{\kappa\kappa} = \frac{1}{2}\int_0^1 (\nabla_\kappa \log \sigma^2_{\kappa_0}(u))(\nabla_\kappa \log \sigma^2_{\kappa_0}(u))' \, du,$$

$$\Gamma_{\tau\tau} = \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} (\nabla_\tau \log f_{\tau_0})(\nabla_\tau \log f_{\tau_0})' \, d\lambda \, du,$$

and

$$\Gamma_{\nu\nu} = \frac{1}{2\pi}\int_0^1 (\nabla_\nu \mu_{\nu_0}(u))(\nabla_\nu \mu_{\nu_0}(u))' \, du,$$

i.e., the estimates are asymptotically independent. $\Gamma_{\kappa\tau} = 0$ follows from (2.6) because

$$\int_{-\pi}^{\pi} f_\theta \nabla_\tau f_\theta^{-1} \, d\lambda = 0$$

and therefore

$$\int_{-\pi}^{\pi} (\nabla_\kappa \log f_\theta)(\nabla_\tau \log f_\theta)' \, d\lambda = -\int_{-\pi}^{\pi} \nabla_\kappa f_\theta \nabla_\tau f_\theta^{-1} \, d\lambda$$

$$= \int_{-\pi}^{\pi} f_\theta \nabla_\kappa \nabla_\tau f_\theta^{-1} \, d\lambda = (\sigma_\kappa^2 \nabla_\kappa \sigma_\kappa{}^2) \int_{-\pi}^{\pi} f_\theta \nabla_\tau f_\theta^{-1} \, d\lambda = 0.$$

(2.6) is e.g., fulfilled for time varying ARMA-models (Kolmogorov's formula–cf. Brockwell and Davis, 1987, Theorem 5.8.1).

Remark 2.7. (stationary models). If the model is stationary ($f_\theta(\lambda) = f_\theta(u,\lambda)$ and $m = \mu_\theta(u)$ do not depend on u) Theorem 2.4 gives the asymptotic distribution of the classical stationary MLE also in the situation where the true underlying process is not stationary. In this situation we have (if $\theta = (\tau, m)$)

$$\mathscr{L}(\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \left\{ \log 4\pi^2 f_\tau(\lambda) + \frac{\int_0^1 f(u,\lambda)\, d\lambda}{f_\tau(\lambda)} \right\} d\lambda$$

$$+ \frac{1}{4\pi} f_\tau(0)^{-1} \int_0^1 (m - \mu(u))^2 \, du,$$

i.e., the optimal parameter $\theta_0 = (\tau_0, m_0)$ is $m_0 = \int_0^1 \mu(u)\, du$ and that value of τ_0 such that $f_\tau(\lambda)$ approximates the time integrated spectrum $\int_0^1 f(u,\lambda)\, du$ best (an example is given in Section 3). If $\theta = (\sigma^2, \tau, m)$ with $f_\theta(\lambda) = \sigma^2/2\pi\, h_\tau(\lambda)$ and (2.6) holds then Γ again is diagonal. To see this note that $\nabla_{\sigma^2} \mathscr{L}(\theta) = 0$ implies

$$\int_0^1 \int_{-\pi}^{\pi} (f - f_{\theta_0}) h_{\tau_0}{}^{-1} \, d\lambda du + \int_0^1 (m - \mu(u))^2 \, du h_{\tau_0}^{-1}(0) = 0$$

which leads to $\Gamma_{\sigma^2\tau} = 0$. However, V will be in general not diagonal. If in addition the model is correct, then Theorem 2.4 is the classical central limit theorem for the MLE (Note, that in this situation the observed triangular array reduces to an

ordinary stationary process). This was e.g., proved in Brockwell and Davis (1987) or Azencott and Dacunha-Castelle (1986). Their result is also a special case of Theorem 2.4.

Remark 2.8 (model selection). Akaike (1974) has suggested to select a model by estimating the expected Kullback-Leibler information divergence between the true unknown process and the fitted model. As an estimate he suggested the AIC-criterion. We now derive the AIC-criterion in the present situation. The expected Kullback-Leibler information divergence is up to a constant (which is independent of the model) $E\mathscr{L}(\hat{\theta}_T)$ where the particular model is taken into account by the special form of $\mathscr{L}(\theta)$. A quadratic expansion of $\mathscr{L}(\theta)$ around θ_0 and $\mathscr{L}_T(\theta)$ around $\hat{\theta}_T$ gives

$$\mathscr{L}(\hat{\theta}_T) \approx \mathscr{L}(\theta_0) + \frac{1}{2}(\hat{\theta}_T - \theta_0)' \nabla^2 \mathscr{L}(\theta_0)(\hat{\theta}_T - \theta_0)$$

and

$$\mathscr{L}_I(\theta_0) \approx \mathscr{L}_I(\hat{\theta}_I) + \frac{1}{2}(\hat{\theta}_I - \theta_0)' \nabla^2 \mathscr{L}_I(\hat{\theta}_I)(\hat{\theta}_I - \theta_0).$$

Since $E\mathscr{L}_I(\theta_0) \approx \mathscr{L}(\theta_0), \nabla^2 \mathscr{L}(\theta_0) - \Gamma$ and $\nabla^2 \mathscr{L}_I(\hat{\theta}_I) \xrightarrow{p} \Gamma$ with Γ as in Theorem 2.4 we may now estimate $E\mathscr{L}(\hat{\theta}_T)$ by

$$\begin{aligned} &\mathscr{L}_T(\hat{\theta}_T) + E(\hat{\theta}_T - \theta_0)\Gamma(\hat{\theta}_T - \theta_0) \\ &\approx \mathscr{L}_T(\hat{\theta}_T) + \frac{1}{T}\operatorname{tr}\{\Gamma^{-1}V\} \end{aligned} \tag{2.8}$$

with V and Γ as in Theorem 2.4. If the model is Gaussian and correctly specified ($f = f_{\theta_0}, \mu = \mu_{\theta_0}$), then $V = \Gamma$, leading to the nonstationary information criterion

$$\mathrm{CRIT}(p) = \mathscr{L}_T(\hat{\theta}_T) + \frac{p}{T} \quad (= \mathrm{AIC}(p)/2).$$

One problem with the above derivation is that we can only prove $E\mathscr{L}_T(\theta_0) - \mathscr{L}(\theta_0) = O(T^{-2/3}\ln^4 T)$ which is of a higher order than p/T. However, we may regard $\mathscr{L}_T(\hat{\theta}_T) + p/T$ as an estimate of $E\mathscr{L}_T(\hat{\theta}_T)$ (instead of $E\mathscr{L}(\hat{\theta}_T)$). Then, the same derivation holds and the above bias-problem does not occur.

Another problem is the estimation of $1/T \operatorname{tr}\{\Gamma^{-1}V\}$ in the case $f \neq f_{\theta_0}$ and/or $\mu \neq \mu_{\theta_0}$ (which clearly is the more realistic assumption in this situation). However, this problem has not even been solved satisfactorily in the much simpler stationary case (cf. the discussion in Findley and Wei, 1990, who term the criterion (2.8) Ideal-AIC).

Analogously to stationary models the model selection step is carried out as follows: One chooses for example as candidate models time varying ARMA-models with time varying means and as candidate models for all coefficient functions polynomials. The model selection step then consists of selecting the ARMA-order and all polynomial orders. For each fixed model (i.e., for all orders) one has to calculate $\hat{\theta}_T, \mathscr{L}_T(\hat{\theta}_T)$ and $\mathrm{AIC}(p) = 2\mathscr{L}_T(\hat{\theta}_T) + 2p/T$ where p is the total number of parameters.

Note, that $\mathscr{L}_T(\hat{\theta}_T)$ is usually not of the simple form $1/2(\log \hat{\sigma}_T^2(p)) + \text{const.}$ as in the stationary case. In the situation where the process can be written in a state space form $\mathscr{L}_T(\hat{\theta}_T)$ can be calculated with the prediction error decomposition and the Kalman filter (cp. Section 3).

The benefits of this model selection criterion are obvious: The (time-varying) dependence model and the trend model can be selected (and therefore also balanced) at the same time by a unified criterion. Furthermore, stationary models can be compared to nonstationary models by the same criterion (very often stationary models occur as submodels to nonstationary models – e.g., when AR-coefficients or the mean are modelled by polynomials a polynomial order of 0 means stationarity). Therefore, no additional tools are necessary to decide whether the model is stationary.

A computational problem may be the large number of candidate models. For example in the case of a time varying ARMA-model each of the coefficient functions may be modelled by a different function system (polynomials, cosines, etc.). In practice one would restrict to one function system (e.g., polynomials) or decide on the function system by using a plot of a nonparametric estimate of the coefficients.

3. STATE SPACE REPRESENTATIONS FOR TIME VARYING ARMA-MODELS

As in the stationary case the estimate $\hat{\theta}_T$ may be calculated by calculating $\mathscr{L}_T(\theta)$ with the prediction error decomposition and the Kalman filter and by using a suitable numerical optimization procedure (cf. Harvey, 1989, Section 3.4). This requires that the model can be written in a (time dependent) state space form.

Consider the following system of difference equations

$$\sum_{j=0}^{p} a_j^{\theta}\left(\frac{t}{T}\right)\left(X_{t-j,T} - \mu_{\theta}\left(\frac{t-j}{T}\right)\right) = \sum_{j=0}^{q} b_j^{\theta}\left(\frac{t}{T}\right)\sigma_{\theta}\left(\frac{t-j}{T}\right)\varepsilon_{t-j} \tag{3.1}$$

where $a_0^{\theta}(u) \equiv b_0^{\theta}(u) \equiv 1$ and the ε_t are independent random variables with mean zero and variance 1. In Dahlhaus (1995, Theorem 2.3 and the following remark) we have shown under certain regularity conditions that $X_{t,T}$ is locally stationary and fulfills Assumptions 2.1 (in particular we have to assume that $\sum_{j=0}^{p} a_j^{\theta}(u) z^j \neq 0$ for all $|z| \leqslant 1 + \varepsilon$ uniformly in θ and u for some $\varepsilon > 0$). The time varying spectral density of this process is

$$f_{\theta}(u,\lambda) = \frac{\sigma_{\theta}^2(u)}{2\pi} \frac{\left|\sum_{j=0}^{q} b_j^{\theta}(u)\exp(i\lambda j)\right|^2}{\left|\sum_{j=0}^{p} a_j^{\theta}(u)\exp(i\lambda j)\right|^2}.$$

$X_{t,T}$ can be written in state space form. Let

$$\alpha_{t,T} = \left(X_{t,T} - \mu\left(\frac{t}{T}\right), \ldots, X_{t-p+1,T} - \mu\left(\frac{t-p+1}{T}\right), \sigma\left(\frac{t}{T}\right)\varepsilon_t, \ldots \sigma\left(\frac{t-q+1}{T}\right)\varepsilon_{t-q+1} \right)'$$

$$T_{t,T} = \begin{pmatrix} -a_1\left(\frac{t}{T}\right) & \cdots & -a_p\left(\frac{t}{T}\right) & b_1\left(\frac{t}{T}\right) & \cdots & b_q\left(\frac{t}{T}\right) \\ 1 & 0 \cdots & 0 & & & \\ & \ddots \ddots & \vdots & & 0 & \\ 0 & 1 & 0 & & & \\ & & & 0 & 0 & \\ & 0 & & 1 & \ddots & \\ & & & & \cdot & \\ & & & 0 & 1 & 0 \end{pmatrix} \qquad R_{t,T} = \begin{pmatrix} \sigma\left(\frac{t}{T}\right) \\ 0 \\ \vdots \\ 0 \\ \sigma\left(\frac{t}{T}\right) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Then

$$\alpha_{t,T} = T_{t,T}\,\alpha_{t-1,T} + R_{t,T}\,\varepsilon_t$$

and

$$X_{t,T} = (1,0,\ldots,0)\,\alpha_{t,T} + \mu\left(\frac{t}{T}\right).$$

The starting values are the same as for a stationary process with parameters $a_1(0),\ldots,\ a_p(0), b_1(0),\ldots, b_q(0), \sigma^2(0)$.

Therefore, the likelihood function $\mathscr{L}_T(\theta)$ for time varying ARMA-models can be calculated (for fixed T!!) by using the prediction error decomposition with the Kalman filter and $\hat{\theta}_T$ may then be obtained by numerical optimization. Furthermore, $\mathscr{L}_T(\hat{\theta}_T)$ and the value of the AIC-criterion are obtained directly.

Since the system matrices are time-dependent the whole procedure is computer-intensive, in particular, if different models are calculated and compared. Therefore, good starting values in the optimization step are important. For time varying AR-models such starting values have been suggested in Dahlhaus (1993, Section 4). Practical aspects of this calculation method will be studied in a forthcoming work.

Sometimes a different state space model is used for stationary models which does *not* transfer to the time varying case. We demonstrate this for the case $p = q = 1$ and $\mu(u) \equiv 0$. Let

$$\alpha_{t,T} = (\alpha_{t,T}^{(1)}, \alpha_{t,T}^{(2)})',$$

$$T_{t,T} = \begin{pmatrix} -a_1\left(\frac{t}{T}\right) & 0 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad R_{t,T} = \begin{pmatrix} \sigma\left(\frac{t}{T}\right) \\ 0 \end{pmatrix}$$

with

$$\alpha_{t,T} = T_{t,T}\,\alpha_{t-1,T} + R_{t,T}\,\varepsilon_t.$$

Then

$$X_{t,T} = \left(1, b_1\left(\frac{t}{T}\right)\right)\alpha_{t,T}$$

is a solution of (3.1) if $a_1(u)$ and $b_1(u)$ are constant over time but not for general time varying functions.

4. EFFICIENCY

We now show that $\hat{\theta}_T$ is also efficient, if the fitted model is correct, i.e., we prove that it is locally asymptotically minimax (LAM) (cf. Millar, 1983, Definition 2.4). To verify the LAM-property for a sequence of estimators we have to check the local asymptotic normality (LAN) of the sequence of experiments. Then, any central sequence is LAM (cf. Strasser, 1985, Remark 83.12).

To show this, let $X_{1,T},\ldots,X_{T,T}$ be a locally stationary process with mean function μ_θ and transfer function $A_\theta^0, \theta \in \Theta \subset \mathbb{R}^p$. Let θ_0 be a fixed inner point of Θ and $\Gamma = \Gamma(\theta_0)$ as in Theorem 2.4 (note that Γ and V coincide if $f = f_{\theta_0}$, $\mu = \mu_{\theta_0}$). Suppose that $H := \mathbb{R}^p$ with inner product $\langle h, k\rangle = h'\Gamma k$, $P_{Th} = \mathcal{N}(\mu_{\theta+hT^{-1/2}}, \Sigma_T(A_{\theta+hT^{-1/2}}, A_{\theta+hT^{-1/2}}))$ and $P_h = \mathcal{N}(h, \Gamma^{-1})$.

THEOREM 4.1. Suppose that Assumption 2.1(i)–(iv) holds and the model is correct ($A = A_{\theta_0}, \mu = \mu_{\theta_0}$). Then we have under P_{T0} with $Z_T = -\sqrt{T}\,\Gamma^{-1}\nabla\mathcal{L}_T(\theta_0)$

$$\log\frac{dP_{Th}}{dP_{T0}} - \langle h, Z_T\rangle + \frac{1}{2}\langle h,h\rangle \xrightarrow{P} 0.$$

and

$$Z_T \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma^{-1}),$$

i.e., the sequence of experiments $(\mathbb{R}^T, \mathbb{B}^T, \{P_{Th} : h \in H\})$ is LAN with limit experiment $(\mathbb{R}^p, \mathbb{B}^p, \{P_h : h \in H\})$.

Proof. We have

$$\log \frac{dP_{T,h}}{P_{T,0}} = -T\{\mathscr{L}_T(\theta_0 + hT^{-1/2}) - \mathscr{L}_T(\theta_0)\}$$

$$= -\sqrt{T}h\nabla\mathscr{L}_T(\theta_0) - \frac{1}{2}h'\nabla^2\mathscr{L}_T(\bar{\theta}_T)h$$

where $|\bar{\theta}_T - \theta_0| \leqslant T^{-1/2}$. As in the proof of Theorem 2.4 we have

$$\nabla^2\mathscr{L}_T(\bar{\theta}_T) \xrightarrow{P} \Gamma$$

and

$$Z_T = -\sqrt{T}\Gamma^{-1}\nabla\mathscr{L}_T(\theta_0) \xrightarrow{\mathscr{L}} N(0,\Gamma^{-1})$$

which implies the result (note that Assumption 2.1(iv) is not needed since the model is correctly specified).

THEOREM 4.2. Suppose that Assumption 2.1 (i)–(iv) holds and the model is correct ($A = A_{\theta_0}, \mu = \mu_{\theta_0}$). Then we have under $P_{T,0}$

$$T^{1/2}(\hat{\theta}_T - \theta_0) - Z_T \xrightarrow{P} 0,$$

i.e., $\hat{\theta}_T$ *is* LAM.

Proof. As in the proof of Theorem 2.4 we get with a Taylor expansion

$$Z_{T,i} = -\sqrt{T}\{\Gamma^{-1}\nabla\mathscr{L}_T(\theta_0)\}_i = \{\Gamma^{-1}\nabla^2\mathscr{L}_T(\theta_T^{(i)})\sqrt{T}(\hat{\theta}_T - \theta_0)\}_i + o_p(1).$$

Since $\Gamma^{-1}\nabla^2\mathscr{L}_T(\theta_T^{(i)}) \xrightarrow{P} I_p$ and $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is stochastically bounded (Theorem 2.4) the result is proved.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic control*, AC-**19**, 716–722.

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

Brillinger, D. R. (1981). Time Series: Data Analysis and Theory. Holden Day, San Francisco.

Brockwell, P. and Davis, R. A. (1987). Time Series: Theory and Methods. Springer Verlag, New York.

Dahlhaus, R. (1993). Fitting time series models to nonstationary process. Beiträge zur Statistik 4, Universität Heidelberg.

Dahlhaus, R. (1995). On the Kullback-Leibler information divergence of locally stationary processes. *Stoch. Proc. Appl.*, to appear

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

Findley, D. and Wei, C. Z. (1989). Beyond Chi-Square: Likelihood Ratio Procedures for Comparing Non-Nested, Possibly Incorrect Regressors. Bureau of the Census, Washington.

Graybill, F. A. (1983). Matrices with Application in Statistics, 2nd ed. Wadsworth. Belmont, California.

Harvey, A. C. (1989). Forecasting, structural time series models and the Kalman filter. Cambridge University Press.

Millar, P. W. (1983). The minimax principle in asymptotic statistical theory. In: Ecole d' Eté de Probabilités de Saint-Flour XI-1981 (Ed. P. L. Hennequin). Springer Lecture Notes in Mathematics **976**.

Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. *J. Roy. Statist. Soc. Ser.* B, **27**, 204–237.
Priestley, M. B. (1981). Spectral Analysis and Time Series, vol.2. Academic Press, London.
Strasser, H. (1985). Mathematical Theory of Statistics. W. de Gruyter, Berlin-New York.
Walker, A. M. (1964). Asymptotic properties of least-squares estimates of parameters of the spectrum of a stationary non-deterministic time series. *J. Austr. Math. Soc.* **4**, 363–384.

APPENDIX

In this appendix we summarize some results on matrix norms and the trace behaviour of $\sum_T(A, A)$ proved in Dahlhaus (1995).

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

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

If A is a real no n negative symmetric matrix, i.e., $A = P'DP$ with $PP' = P'P = I$ and $D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$, where $\lambda_i \geqslant 0$, then we define $A^{1/2} = P'D^{1/2}P$, where $D^{1/2} = \operatorname{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\}$. 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

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

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

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

$$\frac{\partial}{\partial\theta}\log\det A = \operatorname{tr}\left\{A^{-1}\frac{\partial}{\partial\theta}A\right\}, \tag{k}$$

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

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

Assumption. A.2.

(i) Suppose $A: [0,1] \times \mathbb{R} \to \mathbb{C}$ is a 2π-periodic function with $A(u,\lambda) = \overline{A(u,-\lambda)}$ which is differentiable in u and λ with uniformly bounded derivative $(\partial/\partial u)(\partial/\partial\lambda)A$. $A_{t,T}^{\circ}: \mathbb{R} \to \mathbb{C}$ are 2π-periodic functions with

$$\sup_{t,\lambda}\left|A_{t,T}(\lambda) - A\left(\frac{t}{T},\lambda\right)\right| \leqslant KT^{-1}.$$

(ii) Suppose $\phi: [0,1] \times \mathbb{R} \to \mathbb{C}$ is a 2π-periodic function which is differentiable in u and λ with uniformly bounded derivative $(\partial/\partial u)(\partial/\partial\lambda)\phi$.

(iii) Suppose $\mu: [0,1] \to \mathbb{R}$ is differentiable with uniformly bounded derivative.

Remark. All results stated in this appendix are uniform in the sense that the upper bounds depend only on the bounds of the involved functions A, ϕ and μ and their derivatives and not on the particular values.

LEMMA. A.3 Suppose A and B fulfill Assumption A.2(i). Then we have with

$$C_1 = \sup_{u,\lambda}|A(u,\lambda)\,B(u,\lambda)| \text{ and } C_2 = \inf_{u,\lambda}|A(u,\lambda)|^2$$

$$\sup_{|x|=1}|x^*\textstyle\sum_T(A,B)\,x| \leqslant 2\pi C_1 + o(1),\ \inf_{|x|=1}|x^*\textstyle\sum_T(A,A)\,x| \geqslant 2\pi C_2 + o(1)$$

and

$$\|\textstyle\sum_T(A,A)\| \leqslant 2\pi C_1 + o(1), \|\textstyle\sum_T(A,A)^{-1}\| \leqslant (2\pi C_2 + o(1))^{-1}.$$

Proof. This was proved in Lemma 4.4 of Dahlhaus (1995).

LEMMA. A.4 Suppose A_θ fulfills Assumption 2.1(iii). Let $\sum_\theta = \sum_T(A_\theta, A_\theta)$. Then

$$\left\|\sum_\theta\right\|, \left\|\frac{\partial}{\partial\theta_j}\sum_\theta\right\|, \left\|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_\theta\right\| \quad \text{and} \quad \left\|\sum_\theta^{-1/2}\right\|$$

are all uniformly bounded by constants.

Proof. The bounds for $\|\sum_\theta\|$ and $\|\sum_\theta^{-1/2}\|$ immediately follow from Lemma A.3. $\partial/\partial\theta_j\sum_\theta$ may be no longer positive definite. However, it is symmetric. This implies

with $\hat{c}/\hat{c}\theta_j \sum_\theta = P'DP$ where P is orthonormal and D is diagonal

$$\left\| \frac{\hat{c}}{\hat{c}\theta_j} \sum_\theta \right\| = \sup_{|x|=1} (x^* D^2 x)^{1/2} = \sup_{|x|=1} |x^* D x| = \sup_{|x|=1} |x^* \frac{\hat{c}}{\hat{c}\theta_j} \sum_\theta x|$$

$$\leqslant \sup_{|x|=1} |x^* \left(\sum_T \left(\frac{\hat{c}}{\hat{c}\theta_j} A_\theta, A_\theta \right) + \sum_T \left(A_\theta, \frac{\hat{c}}{\hat{c}\theta_j} A_\theta \right) \right) x| \leqslant K.$$

The rest is proved analogously.

LEMMA. A.5 Let $k \in \mathbb{N}$, $A_\prime$, $B_\prime$, $C_\prime$ fulfill Assumption A.2(i) and μ_1, μ_2 fulfill Assumption A.2(iii). Let $\sum_\prime = \sum_T (A_\prime, B_\prime)$ and $\Gamma_\prime = \sum_T (C_\prime, C_\prime)$. Then we have

$$\frac{1}{T} \operatorname{tr} \left\{ \prod_{\prime=1}^{k} \Gamma_\prime^{-1} \sum_\prime \right\} \tag{i}$$

$$- \frac{1}{2\pi} \int_0^1 \int_{-\pi}^{\pi} \left\{ \prod_{\prime=1}^{k} \frac{A_\prime(u,\lambda) B_\prime(u,-\lambda)}{|C_\prime(u,\lambda)|^2} \right\} d\lambda \, du + O(T^{-1/2} \ln^{2k+2} T)$$

$$\frac{1}{T} \mu_{1T}' \left\{ \prod_{\prime=1}^{k-1} \Gamma_\prime^{-1} \sum_\prime \right\} \Gamma_k^{-1} \mu_{2T} \tag{ii}$$

$$= \frac{1}{2\pi} \int_0^1 \left\{ \prod_{\prime=1}^{k-1} \frac{A_\prime(u,0) B_\prime(u,0)}{|C_\prime(u,0)|^2} \right\} |C_k(u,0)|^{-2} \mu_1(u) \mu_2(u) \, du$$

$$+ O(T^{-1/2} \ln^{2k+2} T).$$

If the $C_\prime$ are in addition twice differentiable in u then the remainder terms are of order $O(T^{-2/3} \ln^{2k+2} T)$.

Proof. The assertion is a special case of Lemma 4.8 of Dahlhaus (1995).