Resenhas IME-USP 2000, Vol. 4, No. 4, 457 - 477.

Likelihood Methods for Nonstationary Time Series and Random Fields

R. Dahlhaus and M. Sahm

Abstract: In this article we discuss a generalization of the Whittle likelihood approximation from stationary processes to locally stationary processes and random fields. This yields a local likelihood for these processes which can be used as a starting point for both parametric and semiparametric estimation procedures. For parametric inference, we show the asymptotic normality of the corresponding estimates. In particular the bias reducing effect of using a data taper is discussed, which is essential in the random field scenario.

Key words: Locally stationarity processes, random fields, likelihood approximation, local likelihood, bias reduction, data taper, edge effects.

1. Introduction

In this paper we derive a likelihood approximation to the Gaussian likelihood for locally stationary processes and random fields and show how this likelihood can be used in parametric and nonparametric statistical inference. As a framework for our considerations we use the notion of local stationarity as given in Dahlhaus (1996a) for time series and in Dahlhaus and Sahm (2000) for random fields. This notion is a rigorous framework for an asymptotic theory for processes that are locally close to a stationary process or a stationary random field and whose characteristics such as covariance functions, spectral densities and parameters change slowly over time. Such processes have been considered by numerous authors such as Priestley (1965), Neumann and v. Sachs (1997) or Mallat et al. (1998). In Section 2 a likelihood approximation for locally stationary time series is given which is a weighted average over local likelihoods and a generalization of the well known Whittle likelihood for stationary time series. For parametric models we state the consistency and asymptotic normality of the corresponding estimates. This is achieved by considering the time series with time rescaled to the unit interval [0, 1]. Section 3 provides a brief overview on non- and semiparametric methods based on the local likelihood derived in Section 2 such as kernel estimators, local polynomials and wavelet estimators. In Section 4 the notion of local stationarity as well as the Whittle type likelihood approximation are generalized to the random field case. We present asymptotic results and discuss the problem of overcoming the bias arising in fields of dimension $d > 1$. Section 5 gives an insight into the underlying idea of the Whittle likelihood for nonstationary processes. Here the approximation of Toeplitz matrices is generalized to nonstationary processes.

2. A Likelihood Representation for Locally Stationary Processes

We start with the definition of a locally stationary process. It is given in the form of a time varying spectral representation. The equivalent form of a time varying $MA(\infty)$-representation is discussed below.

In comparison to Dahlhaus (1996 a,b; 2000) we have made a small change: The observation $X_{t,T}$ is assigned to the rescaled time $u = (t - 1/2)/T$ leading to a definition which coincides with the random field case treated in section 4. All asymptotic results stay the same with this modified notation.

Definition 2.1 *A sequence of stochastic processes $X_{t,T}(t = 1, \ldots, T)$ is called locally stationary with transfer function A^0 and trend μ if there exists a representation*

$$X_{t,T} = \mu\left(\frac{t - 1/2}{T}\right) + \int_{-\pi}^{\pi} \exp(i\lambda t) A^0_{t,T}(\lambda) d\xi(\lambda) \tag{1}$$

where

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

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

where cum $\{\ldots\}$ denotes the cumulant of kth order, $h_1 = 0$, $h_2(\lambda) = 1$, $|h_k(\lambda_1, \ldots, \lambda_{k-1})| \leq const_k$ 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

$$\sup_{t,\lambda} |A^0_{t,T}(\lambda) - A(\frac{t - 1/2}{T}, \lambda)| \leq KT^{-1} \tag{2}$$

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

$f(u, \lambda) := A(u, \lambda)\overline{A(u, \lambda)}$ is the time varying spectral density of the process. Under suitable regularity conditions it is uniquely determined (cf. Dahlhaus, 1996a, Theorem 2.2) We denote by

$$c(u, k) := \int_{-\pi}^{\pi} f(u, \lambda) \exp(i\lambda k) d\lambda \tag{3}$$

the local covariance of lag k at time u. If $A(u, \lambda)$ is uniformly Lipschitz continuous in u we have

$$\mathrm{cov}(X_{[uT],T}, X_{[uT]\pm k,T}) = c(u, k) + O(T^{-1}), \tag{4}$$

and in view of the definition of the preperiodogram (see (9) below)

$$\mathrm{cov}(X_{[uT+1+k/2],T}, X_{[uT+1-k/2],T}) = c(u, k) + O(T^{-1}), \tag{5}$$

uniformly in u (under additional regularity assumptions also uniformly in k).

Remark 2.2 (time varying MA(∞)-representations) There exists a close connection between the above spectral representation and time varying MA-representations. Let

$$a_{t,T,k} := \int_{-\pi}^{\pi} A_{t,T}^{o}(\lambda)\exp(i\lambda k)d\lambda,$$
$$a_k(u) := \int_{-\pi}^{\pi} A(u,\lambda)\exp(i\lambda k)d\lambda$$

and

$$\varepsilon_t := \int_{-\pi}^{\pi} \exp(i\lambda t)d\xi(\lambda).$$

Then $E\varepsilon_t = 0$ and $E\varepsilon_s\varepsilon_t = 2\pi\delta_{st}$, i.e. the ε_t are uncorrelated. Since

$$A_{t,T}^{o}(\lambda) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} a_{t,T,k}\exp(-i\lambda k)$$

and

$$A(u,\lambda) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} a_k(u)\exp(-i\lambda k)$$

we obtain

$$X_{t,T} = \mu(\frac{t-1/2}{T}) + \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} a_{t,T,k}\varepsilon_{t-k}. \tag{6}$$

Condition (2) implies

$$\sup_{t,k} |a_{t,T,k} - a_k(\frac{t-1/2}{T})| = O(T^{-1}).$$

If we start conversely with an infinite MA-representation (6) where the coefficients fulfill

$$\sup_t \sum_{k=-\infty}^{\infty} |a_{t,T,k} - a_k(\frac{t-1/2}{T})| = O(T^{-1}) \tag{7}$$

then it can be shown in the same way that a representation (1) exists and (2) is fulfilled. Note that heteroscedastic ε_t and ε_t with dependent components can be included by choosing other $a_{t,T,k}$ in (6). The complicated construction with different functions $A_{t,T}^{o}(\lambda)$ and $A(\frac{t-1/2}{T},\lambda)$ ($a_{t,T,k}$ and $a_k(\frac{t-1/2}{T})$ respectively) is necessary since we need on the one hand a certain smoothness in time direction (guaranteed by the functions $A(u,\lambda)$ and $a_k(u)$) and on the other hand a class which is rich enough to cover interesting applications. For example, the time varying AR(1)-process $X_{t,T} = \phi(\frac{t-1/2}{T})X_{t-1,T} + \varepsilon_t$ does not have a solution of the form $X_{t,T} = \sum_{k=0}^{\infty} a_k(\frac{t-1/2}{T})\varepsilon_{t-k}$ but only of the form $X_{t,T} = \sum_{k=0}^{\infty} a_{t,T,k}\varepsilon_{t-k}$ with (7) where $a_k(u) = \phi(u)^k$.

As mentioned above the time parameter $u = (t-1/2)/T$ in $\mu(u)$ and $A(u,\lambda)$ is rescaled for a meaningful asymptotic theory leading to the above triangular array $X_{t,T}$. This is the same approach as in nonparametric regression. The classical asymptotics for stationary sequences are contained as a special case (if μ and A do not depend on t). A detailed discussion of this definition and a comparison to Priestley's approach can be found in Dahlhaus (1996b). Another definition of local stationarity has recently been given by Mallat, Papanicolaou and Zhang (1998). We remark that the methods presented in this paper do not depend on the special definition of local stationarity.

Examples of locally stationary processes can be found in Dahlhaus (1996a). We just remark that for example ARMA-models with time varying coefficients are locally stationary, that is $X_{t,T}$ defined by the difference equations

$$\sum_{j=0}^{p} \phi_j\left(\frac{t-1/2}{T}\right) X_{t-j,T} = \sum_{j=0}^{q} \psi_j\left(\frac{t-1/2}{T}\right) \sigma\left(\frac{t-j-1/2}{T}\right) \varepsilon_{t-j}$$

with $\phi_o(u) \equiv \psi_o(u) \equiv 1$ and ε_t iid with mean zero and variance 1 where the roots of $\sum_{j=0}^{p} \phi_j(u) z^j$ lie outside the unit circle and are assumed to be uniformly bounded away from the unit circle. The time varying spectral density is

$$f(u,\lambda) = \frac{\sigma^2(u)}{2\pi} \frac{|\sum_{j=0}^{q} \psi_j(u) \exp(i\lambda j)|^2}{|\sum_{j=0}^{p} \phi_j(u) \exp(i\lambda j)|^2}. \tag{8}$$

(cf. Dahlhaus, 1996a, Theorem 2.3 and the discussion thereafter).

The topic of this paper is statistical inference for parameter curves describing local stationarity such as the curves $\phi_j(u)$ and $\psi_j(u)$ from the above example. This may either be treated as a nonparametric problem or as a parametric problem where the $\phi_j(u)$ and $\psi_j(u)$ are modelled e.g. as polynomials in time with the coeffficients being the parameters. For the estimation we will use a generalization of the Whittle likelihood (Whittle 1953, 1954) i.e. an approximation to the Gaussian likelihood based on some distance measure in the frequency domain. For this reason we start our discussion with some considerations on spectral estimation.

A statistic which plays a fundamental role is the preperiodogram. It was introduced by Neumann and von Sachs (1997) as a starting point for a wavelet estimate of the time varying spectral density. Here we use a modified form which in particular uses a data taper. For $u \in [0,1]$ let

$$J_T^{(h)}(u,\lambda) := \frac{1}{2\pi} h_T(u)^{-2} \sum_{k \atop 1 \le [uT+1 \pm k/2] \le T} X^{(h)}_{[uT+1-k/2],T} X^{(h)}_{[uT+1+k/2],T} \exp(-i\lambda k \tag{9}$$

where $[x]$ denotes the largest integer less or equal to x. $X_{t,T}^{(h)} := h_T(\frac{t-1/2}{T}) X_{t,T}$ are the tapered data where $h_T(u)$ is some rescaled data taper with

$$\frac{1}{T} \sum_{t=1}^{T} h_T\left(\frac{t-1/2}{T}\right)^2 = 1,$$

i.e.

$$h_T(u) = \Big(\frac{1}{T}\sum_{t=1}^{T} h_\rho\big(\frac{t-1/2}{T}\big)^2\Big)^{-1/2} h_\rho(u) \tag{10}$$

where ρ is the precentage of tapered data. For example $h_\rho(u)$ may be the cosine taper

$$h_\rho(u) = \begin{cases} \frac{1}{2}(1-\cos\frac{2\pi u}{\rho}), & \text{if } 0 \le u \le \rho/2 \\ 1, & \text{if } \rho/2 \le u \le 1/2. \\ h_\rho(1-u), & \text{if } 1/2 \le u \le 1 \end{cases} \tag{11}$$

If $\rho = 0$, we have the classical non-tapered case.

There exists a nice relation between the tapered preperiodogram and the tapered ordinary periodogram:

$$\begin{aligned} I_T^{(h)}(\lambda) &= \frac{1}{2\pi T}\Big|\sum_{r=1}^{T} X_{r,T}^{(h)}\exp(-i\lambda r)\Big|^2 \\ &= \frac{1}{2\pi}\sum_{k=-(T-1)}^{T-1}\left(\frac{1}{T}\sum_{t=1}^{T-|k|} X_{t,T}^{(h)} X_{t+|k|,T}^{(h)}\right)\exp(-i\lambda k) \\ &= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2\pi}\sum_{k \atop 1\le[t+1/2\pm k/2]\le T} X_{[t+1/2+k/2],T}^{(h)} X_{[t+1/2-k/2],T}^{(h)}\exp(-i\lambda k) \\ &= \frac{1}{T}\sum_{t=1}^{T} h_T\Big(\frac{t-1/2}{T}\Big)^2 J_T^{(h)}\Big(\frac{t-1/2}{T},\lambda\Big), \end{aligned} \tag{12}$$

i.e. the periodogram is the weighted average of the preperiodogram over time. (12) means that the periodogram $I_T^{(h)}(\lambda)$ is the Fourier transform of the covariance estimator of lag k over the whole segment while the preperiodogram $J_T^{(h)}(\frac{t-1/2}{T},\lambda)$ just uses the pair $X_{[t+1/2+k/2]}^{(h)} X_{[t+1/2-k/2]}^{(h)}$ as a kind of "local estimator" of the covariance of lag k at time t (note that $[t+1/2+k/2]-[t+1/2-k/2]=k$). For this reason Neumann and von Sachs also called $J_T(\frac{t}{T},\lambda)$ the localized periodogram.

A classical kernel estimator of the spectral density of a stationary process at some frequency λ_0 therefore can be regarded as a weighted average of the preperiodogram over all time points and over the frequencies in the neighbourhood of λ_0. It is therefore plausible that averaging the preperiodogram about some frequency λ_0 and about some time-point t_0 gives an estimate of the time-varying spectrum $f(\frac{t_0-1/2}{T},\lambda)$.

(5) shows that $J_T^{(h)}(u,\lambda)$ is an asymptotically unbiased estimate of the time varying spectral density $f(u,\lambda)$. Similar to the ordinary periodogram it is, however, not consistent. The large variability is for example reflected by the relation

$$\int_{-\pi}^{\pi} J_T^{(h)}\big(\frac{t-1/2}{T},\lambda\big)d\lambda = \frac{2\pi}{T}\sum_{s=1}^{T} J_T^{(h)}\big(\frac{t-1/2}{T},\lambda_s\big) = X_t^2.$$

This means that smoothing of the preperiodogram in time and frequency direction is essential to make a reasonable estimate out of it. The advantage over (say) a classical periodogram on some small time segment is that it does not contain any implicit smoothing which makes it a valuable raw estimate for adaptive smoothing techniques such as wavelet estimates of the time varying spectral density or likelihood oriented methods as discussed in this paper.

In the stationary situation data tapers are known to reduce the bias of the ordinary periodogram $I_T(\lambda)$ due to spectral leakage. The same holds for all estimates which can be written as functionals of the periodogram such as empirical covariances, Yule-Walker estimates and Whittle estimates (cf. Dahlhaus, 1988; Dahlhaus and Künsch, 1987). The percentage ρ is usually chosen to be small (say between 0.1 and 0.3) and $\rho = \rho_T \to 0$ as $T \to \infty$ is a realistic assumption leading to asymptotically efficient estimates. In the present nonstationary situation the data taper leads to a downweighting of the observations at the edges and in addition to a downweighting of the preperiodograms at the edges (as in (12) or in the Whittle function (14) below), which is plausible in view of the small number of summands over k in the preperiodogram $J_T^{(h)}(u, \lambda)$ for u close to 0 or 1.

We now use the preperiodogram to construct an approximation to the Gaussian likelihood of a locally stationary process. For stationary processes Whittle (1953, 1954) had introduced the likelihood

$$\frac{1}{4\pi} \int_{-\pi}^{\pi} \left\{ \log 4\pi^2 f_\theta(\lambda) + \frac{I_T(\lambda)}{f_\theta(\lambda)} \right\} d\lambda \tag{13}$$

which is an approximation to $-\frac{1}{T} \log$ Gaussian likelihood. Dahlhaus (1987) had suggested to use the Whittle likelihood with the tapered periodogram instead.

If we use $I_T^{(h)}(\lambda)$ instead of $I_T(\lambda)$ in this likelihood and replace the model spectral density $f_\theta(\lambda)$ by the time-varying spectral density $f_\theta(u, \lambda)$ of a parametric nonstationary model, we obtain

$$\begin{aligned}\mathcal{L}_T(\theta) \quad = \quad & \frac{1}{4\pi} \frac{1}{T} \sum_{t=1}^{T} h_T\left(\frac{t-1/2}{T}\right)^2 \\ & \int_{-\pi}^{\pi} \left\{ \log \left[4\pi^2 f_\theta \left(\frac{t-1/2}{T}, \lambda \right) \right] + \frac{J_T^{(h)}(\frac{t-1/2}{T}, \lambda)}{f_\theta(\frac{t-1/2}{T}, \lambda)} \right\} d\lambda \end{aligned} \tag{14}$$

as a generalization of the Whittle likelihood to nonstationary processes. If the model is stationary, i.e. $f_\theta(u, \lambda) = f_\theta(\lambda)$ then the above likelihood is identical to the classical Whittle likelihood, i.e. we have a true generalization to nonstationary processes.

In Dahlhaus (2000) the asymptotic properties of this likelihood and the corresponding estimate have been investigated in the nontapered case. The properties in the tapered case follow as a special case from the investigations for random fields in Dahlhaus and Sahm (2000) (see also Section 4 below). We briefly describe the results here. Let

$$\hat{\theta}_T := \underset{\theta \in \Theta}{\operatorname{argmin}} \mathcal{L}_T(\theta). \tag{15}$$

Below we state that $\hat{\theta}_T \xrightarrow{P} \theta_o$ where θ_o is the true parameter. If the model is misspecified the same holds with

$$\theta_0 := \underset{\theta \in \Theta}{\operatorname{argmin}} \mathcal{L}(\theta) \tag{16}$$

where

$$\mathcal{L}(\theta) := \frac{1}{4\pi} \int_0^1 h(u)^2 \int_{-\pi}^{\pi} \Big\{ \log[4\pi^2 f_\theta(u,\lambda)] + \frac{f(u,\lambda)}{f_\theta(u,\lambda)} \Big\} d\lambda du \tag{17}$$

with

$$h(u) = \lim_{T\to\infty} h_T(u)$$

is the limit of $\mathcal{L}_T(\theta)$. In the case where the model is correctly specified, i.e. $f(u,\lambda) = f_{\theta^*}(u,\lambda)$ with some $\theta^* \in \Theta$ one can show that $\theta_0 = \theta^*$.

We also define the exact Gaussian likelihood estimate by

$$\tilde{\theta}_T := \underset{\theta \in \Theta}{\operatorname{argmin}} \tilde{\mathcal{L}}_T(\theta). \tag{18}$$

where

$$\tilde{\mathcal{L}}_T(\theta) := \frac{1}{2}\log(2\pi) + \frac{1}{2T}\log\det \Sigma_\theta + \frac{1}{2T}\underline{X}'\Sigma_\theta^{-1}\underline{X} \tag{19}$$

is $-\frac{1}{T}\log$ Gaussian likelihood and Σ_θ is the variance covariance matrix of the model under consideration.

We set $\nabla_i = \frac{\partial}{\partial\theta_i}$ and $\nabla^2_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}$. The results hold under the following assumptions.

Assumption 2.3 *(i) We observe a realization $X_{1,T}, ..., X_{T,T}$ of a locally stationary Gaussian process with mean 0, transfer function A^o and covariance matrix Σ. We fit a class of locally stationary Gaussian processes with mean 0, transfer function A^o_θ and covariance matrix Σ_θ, $\theta \in \Theta \subset \mathbb{R}^p$, Θ compact.*

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

(iii) $A_\theta(u,\lambda)$ is differentiable with respect to θ, u and λ with uniformly continuous derivatives $\nabla^2_{ij}\frac{\partial^2}{\partial u^2}\frac{\partial}{\partial\lambda}A_\theta(u,\lambda)$.

(iv) $f_\theta(u,\lambda) = |A_\theta(u,\lambda)|^2$ and $f(u,\lambda) = |A(u,\lambda)|^2$ are bounded from below by some constant $C > 0$ uniformly in θ, u and λ.

(v) The data taper h_T is a normalized taper of proportion $\rho \in [0,1]$ (see (10)), where h_ρ is symmetric about $1/2$ and $h_\rho(u) = 1$ for $u \in [\rho/2, 1/2]$ and $h(u) = \omega(u/(2\rho))$ for $u \in [0, \rho/2]$; $\omega : [0,1] \to [0,1]$ is twice continously differentiable and strictly increasing with $\omega(0) = 0$, $\omega(1) = 1$ and $\omega'(0) = \omega'(1) = 0$. (Note that this also includes the nontapered case by letting $\rho = 0$.) The limit of h_T $(T \to \infty)$ will be denoted by h.

In Dahlhaus (2000) (nontapered case) and in Dahlhaus and Sahm (2000) (tapered case) we have proved the following result.

Theorem 2.4 *Suppose that Assumption 2.3 holds. Then we have*

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_h^{-1} V_h \Gamma_h^{-1}) \text{ and } \sqrt{T}(\tilde{\theta}_T - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_1^{-1} V_1 \Gamma_1^{-1})$$

with

$$(\Gamma_h)_{ij} = \frac{1}{4\pi} \int_0^1 h^2(u) \int_{-\pi}^{\pi} \left[(f - f_{\theta_0}) \nabla_{ij} f_{\theta_0}^{-1} + f_{\theta_0} (\nabla_i f_{\theta_0}^{-1}) f_{\theta_0} (\nabla_j f_{\theta_0}^{-1}) \right] d\lambda du ,$$

and

$$(V_h)_{ij} = \frac{1}{4\pi} \int_0^1 h^4(u) \int_{-\pi}^{\pi} f (\nabla_i f_{\theta_0}^{-1}) f (\nabla_j f_{\theta_0}^{-1})\, d\lambda du$$

where Γ_1, V_1 dentote the matrices Γ_h; V_h corresponding to the nontapered case $h(u) = \chi_{[0,1]}$.

Remark 2.5 (i) Theorem 2.4 contains the asymptotic distribution of the Whittle-estimate and the MLE in the stationary situation as a special case (if f and f_{θ_o} do not depend on u). Theorem 2.4 also gives the asymptotic distribution in the case where a stationary model is used with the classical Whittle-likelihood but the process is only locally stationary.
(ii) The matrices Γ and V from Theorem 2.4 simplify in several situations, in particular when the model is correctly specified (i.e. $f = f_{\theta_o}$), when a stationary model is fitted (f_θ does not depend on u), and when the parameters separate.
(iii) In the correctly specified case ($f = f_{\theta_o}$) we have $\Gamma_1 = V_1$. Furthermore V_1 is the limit of the Fisher information matrix, i.e. the MLE is efficient (for LAN see also Remark 3.3 in Dahlhaus, 2000). By using the Cauchy-Schwarz inequality it can be shown in the case $f = f_{\theta_0}$ that $\Gamma_h^{-1} V_h \Gamma_h^{-1} - V_1^{-1} \geq 0$, i.e. $\widehat{\theta}_T$ is less efficient than the MLE with equality if the data taper is asymptotically vanishing (which is a reasonable assumption).
(iv) It can be shown in the nontapered case (cf. Dahlhaus, 2000, Theorem 2.8) that $\mathcal{L}_T(\theta)$ is an approximation to $\tilde{\mathcal{L}}_T(\theta)$ thus confirming the heuristics of the beginning of this section. Furthermore $\hat{\theta}_T - \tilde{\theta}_T = O_p(T^{-1+\varepsilon})$ holds for each $\varepsilon > 0$ (Dahlhaus, 2000, Remark 3.4). Presumably this also holds for asymptotically vanishing tapers but not for non-vanishing tapers.
(v) In Dahlhaus (2000) the more general situation of a multivariate locally stationary process with time varying trend μ and a time varying trend model μ_θ has been considered.

There are a number of estimates with similar properties which are based on a simplified or modified form of the likelihood:
(i) The first simplification results from the observation that the first summand of the likelihood is often free from the parameters describing autocorrelations. Suppose for example that the process has a one-sided MA(∞)-representation

$$X_{t,T} = \mu(\frac{t-1/2}{T}) + \frac{1}{2\pi}\sum_{k=0}^{\infty} a_{t,T,k}\varepsilon_{t-k}$$

with $a_{t,T,k}$ as in Remark 2.2 and $E\varepsilon_t = 0$, $E\varepsilon_s\varepsilon_t = 2\pi\delta_{st}$. Let

$$\sigma^2(u) = \frac{1}{2\pi}a_0(u)a_0(u)$$

where $a_0(u) = \int_{-\pi}^{\pi} A(u,\lambda)d\lambda$. $\sigma^2((t-1/2)/T)$ is up to an $O(T^{-1})$-error (due to the approximation (2) or (7)) the variance of the one-step prediction error at time t and the variance of the innovations in a standardized MA(∞)-representation. It follows from Kolmogorov's formula (cf. Brockwell and Davis, 1987, Theorem 5.8.1) that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log[(2\pi)f(u,\lambda)]d\lambda = \log\sigma^2(u)$$

leading to a simplified version of $\mathcal{L}_T(\theta)$.
(ii) Another simplification results from the replacement of the integral in $\mathcal{L}_T(\theta)$ by $\frac{2\pi}{T}$ times the sum over the Fourier Frequencies. It can be shown that the resulting estimate has the same asymptotic properties as $\hat{\theta}_T$.

3. Nonparametric Local Likelihood Estimation

In this section we demonstrate how the above likelihood can be used for the construction of various nonparametric estimates of parameter curves $\theta(u)$ for locally stationary curves.

The likelihood derived in (14) is of the form

$$\mathcal{L}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} h_T\Big(\frac{t-1/2}{T}\Big)^2 \ell_T(\theta, \frac{t-1/2}{T}) \tag{20}$$

with

$$\ell_T(\theta,u) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\{\log\left[4\pi^2 f_\theta(u,\lambda)\right] + \frac{J_T^{(h)}(u,\lambda)}{f_\theta(u,\lambda)}\}d\lambda, \tag{21}$$

i.e. $\mathcal{L}_T(\theta)$ has (up to the weights $h_T((t-1/2)/T)^2$) a similar form as the negative log-likelihood function of iid observations where $\ell_T(\theta, \frac{t-1/2}{T})$ is the negative log-likelihood at time point t. In the present dependent situation $\ell_T(\theta, \frac{t-1/2}{T})$ may still be regarded as the negative log-likelihood at time point t which now in addition contains the full information on the dependence (correlation) structure of $X_{t,T}$ with all the other variables.

To illustrate this we give two examples:

1. Suppose we have independent observation with a time varying variance, i.e. our model is

$$X_{t,T} = \sigma(\frac{t-1/2}{T})\varepsilon_t, \qquad \varepsilon_t \text{ iid } \mathcal{N}(0,1),$$

with $\sigma(u) = \sigma_\theta(u)$. This process is locally stationary. It is easy to show that in this case

$$\ell_T(\theta, \frac{t-1/2}{T}) = \frac{1}{2}\log 2\pi\sigma_\theta^2(\frac{t-1/2}{T}) + \frac{1}{2\sigma_\theta^2(\frac{t-1/2}{T})}X_{t,T}^2 ,$$

i.e. in the nontapered case $\mathcal{L}_T(\theta)$ is exactly the negative Gaussian log-likelihood. We remark that this example can be extended to the general model of nonparametric regression with heteroscedastic errors

$$X_{t,T} = m(\frac{t}{T}) + \sigma(\frac{t}{T})\varepsilon_t, \quad \varepsilon_t \text{ iid } \mathcal{N}(0,1)$$

We then have to modify the definition of $\mathcal{L}_T(\theta)$ and $\ell_T(\theta, \frac{t}{T})$ to include the time varying mean (as in Dahlhaus, 2000, (2.6)) and obtain

$$\ell_T(\theta, \frac{t-1/2}{T}) = \frac{1}{2}\log 2\pi\sigma_\theta^2(\frac{t-1/2}{T}) + \frac{1}{2\sigma_\theta^2(\frac{t-1/2}{T})}(X_{t,T} - m_\theta(\frac{t-1/2}{T}))^2$$

which again gives exactly the Gaussian log-likelihood.

2. Suppose

$$X_{t,T} = \phi(\frac{t-1/2}{T})X_{t-1,T} + \sigma(\frac{t-1/2}{T})\varepsilon_t, \qquad \varepsilon_t \text{ iid } \mathcal{N}(0,1),$$

with $\phi(u) = \phi_\theta(u)$, $\sigma(u) = \sigma_\theta(u)$. Then $X_{t,T}$ is locally stationary with time varying spectrum

$$f_\theta(u,\lambda) = \frac{\sigma_\theta^2(u)}{2\pi}|1 - \phi_\theta(u)e^{i\lambda}|^{-2}$$

leading to

$$\begin{aligned}\ell_T(\theta, \frac{t-1/2}{T}) &= \frac{1}{2}\log 2\pi\sigma_\theta^2(\frac{t-1/2}{T}) + \\ &+ \frac{1}{2\sigma_\theta^2(\frac{t-1/2}{T})}(X_{t,T} - \phi_\theta(\frac{t-1/2}{T})X_{t-1,T})^2 + r_t\end{aligned}$$

with

$$r_t = \phi(\frac{t-1/2}{T})^2(X_{t,T}^2 - X_{t-1,T}^2),$$

and $\sum_{t=1}^T r_t = O_p(1)$.

The fact that $\ell_T(\theta, \frac{t-1/2}{T})$ can be seen as the local likelihood of the process at time t opens the door for various nonparametric estimation methods.

Recall that several nonparametric estimation techniques can be written as the solution of a least squares problem, for example for the simple nonparametric regression problem

$$X_{t,T} = m(\frac{t-1/2}{T}) + \varepsilon_t$$

a) a kernel estimate can be written as

$$\hat{m}(u) = \underset{m}{\operatorname{argmin}} \frac{1}{b_T T} \sum_t K\left(\frac{u-(t-1/2)/T}{b_T}\right)\{X_{t,T} - m\}^2$$

where K is the kernel and b_T is some bandwidth;

b) a local polynomial fit can be written as

$$\hat{c}(u) = \underset{c}{\operatorname{argmin}} \frac{1}{b_T T} \sum_t K\left(\frac{u-(t-1/2)/T}{b_T}\right) \cdot$$

$$\cdot \left\{X_{t,T} - \sum_{j=0}^{d} c_j(\frac{t-1/2}{T} - u)^j\right\}^2;$$

where $c = (c_0, \ldots, c_d)'$ are the coefficients of the fitted polynomial at time u;

c) an orthogonal series estimator (e.g. wavelets) can be written as

$$\bar{\alpha} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{T} \sum_t \left\{X_{t,T} - \sum_{j=1}^{J} \alpha_j \psi_j(\frac{t-1/2}{T})\right\}^2$$

together with some shrinkage to obtain the final estimator $\hat{\alpha}$. Here the $\psi_j(\cdot)$ $(j = 1, \ldots, J)$ denote some orthonormal functions. J usually increases with T.

Note that the $\{\ldots\}$-brackets always contain the negative log likelihood of the parameters up to some constants.

Suppose now we have a locally stationary model which is parametrized by one or several curves in time. By using the local likelihood we may define completely analogously to above

a) a kernel estimate by

$$\hat{\theta}(u) = \underset{\theta}{\operatorname{argmin}} \frac{1}{b_T T} \sum_{t=1}^{T} K\left(\frac{u-(t-1/2)/T}{b_T}\right) \ell_T\left(\theta, \frac{t-1/2}{T}\right);$$

b) a local polynomial fit by

$$\hat{c}(u) = \underset{c}{\operatorname{argmin}} \frac{1}{b_T T} \sum_{t=1}^{T} K\left(\frac{u-(t-1/2)/T}{b_T}\right) \cdot$$

$$\ell_T\left(\sum_{j=0}^{d} c_j\left(\frac{t-1/2}{T}-u\right)^j, \frac{t-1/2}{T}\right);$$

c) an orthogonal series estimator (e.g. wavelets) by

$$\bar{\alpha} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \ell_T\left(\sum_{j=1}^{J} \alpha_j \psi_j\left(\frac{t-1/2}{T}\right), \frac{t-1/2}{T}\right);$$

together with some shrinkage of $\bar{\alpha}$.

In case of several parameter curves (a vector of curves) θ, the c_j and the α_j are also vectors. In case of a multivariate process or a process with mean different from zero the definition of $\ell_T(\theta, \frac{t-1/2}{T})$ from Remark 2.10 of Dahlhaus (2000) has to be used. Furthermore, we conjecture that it is beneficial to use the weights $h_T((t-1/2)/T)^2$ as in (20).

It is obvious that the properties of these estimators have to be investigated in detail. In Dahlhaus and Neumann (2000) this has been done in the nontapered case for the wavelet estimate from c). It has been shown that the usual rates of convergence in Besov smoothness classes are attained up to a logarithmic factor by the estimator.

4. Local Likelihood Methods for Random Fields

In this section we consider a process on the d-dimensional grid $D_N = \{1, \ldots, N\}^d$, denoted by X_n, $n \in D_N$. Here and throughout this section, n is a multiindex of dimension d (we use n and N instead of t and T to distinguish the random field case from the time series situation). We will extend the notion of local stationarity to the random field situation and introduce the generalized Whittle likelihood. Again, this likelihood has a representation as an average over local likelihoods and can therefore be used as a starting point for various non- and semiparametric procedures as discussed in the last section. We will focus on the parametric situation and discuss the main difficulty and difference to the time series case: a potential bias of crucial order.

The random field case can be treated in quite the same way as the time series situation; however, due to the presence of potential bias more thorough analysis of the derivative of the likelihood is called for.

First we define locally stationary random fields in the same way as for the time series.

Definition 4.1 *A sequence of random fields $X_{n,N}$, $n \in D_N$ is called locally stationary with transfer function A^0 and trend μ if there exists a representation*

$$X_{n,N} = \mu\Big(\frac{n-1/2}{N}\Big) + \int_{(-\pi,\pi]^d} \exp(i < \lambda, n >)\, A^0_{n,N}(\lambda)\, d\xi(\lambda)$$

where $\xi(\lambda)$ is a process on $(-\pi,\pi]^d$ fulfilling (i) of Definition 2.1.

Furthermore there are a constant K and a periodic function $A : [0,1]^d \times (-\pi,\pi]^d \to \mathbb{C}$ such that $A(u,\lambda) = \overline{A(u,-\lambda)}$ and

$$\sup_{n,\lambda} \Big| A^0_{n,N}(\lambda) - A\Big(\frac{n-1/2}{N}, \lambda\Big) \Big| \le \frac{K}{N} \tag{22}$$

The quantity $f(u,\lambda) = |A(u,\lambda)|^2$ is called the varying spectral density of the field.

Note that in the time series case the approximation error of order N^{-1} corresponds to the number of observations N, while in the random field scenario with N^d observations this error of order N^{-1} is much larger relative to the normalizing constant N^{-d} in the likelihood (see (24)). However, important examples as AR- or MA-processes with smoothly varying coefficients, which have to be included in the framework of locally stationary random fields, require this definition. For example, a two-dimensional causal AR(1) process with varying coefficients α and β and constant variance σ^2:

$$X_{n_1,n_2} = \alpha(\tfrac{n_1-1/2}{N}, \tfrac{n_2-1/2}{N})\, X_{n_1-1,n_2} + \beta(\tfrac{n_1-1/2}{N}, \tfrac{n_2-1/2}{N})\, X_{n_1,n_2-1} + \sigma\, \varepsilon_{n_1,n_2}$$

(where $n_1, n_2 \in \{1,\ldots,N\}$ and α and β are smooth functions on the unit square) has a spectral representation

$$X_n = \int_{(-\pi,\pi]^2} A^0_{n,N}(\lambda)\, e^{i<\lambda,n>}\, d\xi(\lambda)$$

with

$$A^0_{n,N}(\lambda) = A\Big(\frac{n-1/2}{N}, \lambda\Big) + \frac{1}{N}\, B\Big(\frac{n-1/2}{N}, \lambda\Big) + O(N^{-2}) \tag{23}$$

where

$$A(u,\lambda) = A(u_1, u_2, \lambda_1, \lambda_2)$$

$$
\begin{aligned}
&= \frac{\sigma}{4\pi^2\,(1-e^{-i\lambda_1}\alpha(u_1,u_2)-e^{-i\lambda_2}\beta(u_1,u_2))} \\
B(u,\lambda) &= B(u_1,u_2,\lambda_1,\lambda_2) \\
&= -\frac{\sigma}{4\pi^2}\sum_{j,k\geq 0} e^{-i(j\lambda_1+k\lambda_2)}\,\alpha(u_1,u_2)^{j-1}\,\beta(u_1,u_2)^{k-1}\binom{j+k}{k} \\
&\quad \times\left[\binom{j}{2}\beta(u_1,u_2)\nabla_1\alpha(u_1,u_2)+\binom{k}{2}\alpha(u_1,u_2)\nabla_2\beta(u_1,u_2)\right. \\
&\quad \left.+\frac{jk}{2}(\beta(u_1,u_2)\nabla_2\alpha(u_1,u_2)+\alpha(u_1,u_2)\nabla_1\beta(u_1,u_2))\right].
\end{aligned}
$$

This can easily be calculated from the MA-representation of the process. Note that $A(u,\cdot)$ is the transfer function of a stationary $AR(1)$-process with parameters $\alpha(u)$ and $\beta(u)$ and that in general $B(u,\lambda)$ does not vanish. This structure of an existing second order approximation seems to be the typical situation and will be used later in Theorem 4.5.

The generalized Whittle likelihood we consider is almost the same as in the time series case with some modification due to the potential presence of a bias. It is defined by

$$
\begin{aligned}
\widetilde{\mathcal{L}}_N(\theta) &= \frac{1}{2\,(2\pi)^d}\frac{1}{N^d}\sum_{n\in D_N} h_N^2(\tfrac{n-1/2}{N})\ \times \\
&\int_{(-\pi,\pi]^d}\left\{\log\left[(2\pi)^{2d}f_\theta(\tfrac{n-1/2}{N},\lambda)\right]+\frac{\widetilde{J}_N^{(h)}(\frac{n-1/2}{N},\lambda)}{f_\theta(\frac{n-1/2}{N},\lambda)}\right\}d\lambda \qquad (24)
\end{aligned}
$$

where the tapered preperiodogram now is defined as

$$
\begin{aligned}
\widetilde{J}_N^{(h)}\Big(\frac{n-1/2}{N},\lambda\Big) &= \frac{h_N^{-2}(u)}{2(2\pi)^d}\left[\sum_r e^{i<\lambda,r>}\ X^{(h)}_{[n+\frac{r}{2}],N}X^{(h)}_{[n-\frac{r}{2}],N}\ +\right. \qquad (25) \\
&\qquad \left.\sum_r e^{i<\lambda,r>}\ X^{(h)}_{[n+\frac{r}{2}]^*,N}X^{(h)}_{[n-\frac{r}{2}]^*,N}\right]
\end{aligned}
$$

where $[n]^*$ denotes the smallest integer larger or equal to n and the sums are taken over all $r\in\mathbb{Z}$ such that the corresponding observations $X^{(h)}_{n,N}=h_N(\frac{n-1/2}{N})X_{n,N}$ exist; h_N is a normalized d-dimensional taper of proportion ρ, i.e.

$$
h_N(u) = \prod_{i=1}^d h_\rho(u_i)\left(\frac{1}{N^d}\sum_{n\in D_N}\prod_{i=1}^d h_\rho^2(\tfrac{n_i-1/2}{N})\right)^{-1}
$$

where h_ρ is a one-dimensional taper of proportion ρ.

For the modified preperiodogram we have the same relation as in (12), namely

$$\frac{1}{N^d} \sum_{n \in D_N} h_N \left(\tfrac{n-1/2}{N}\right)^2 \widetilde{J}_N^{(h)}\left(\tfrac{n-1/2}{N}, \lambda\right) \tag{26}$$

$$= \tfrac{1}{(2\pi N)^d} \left| \textstyle\sum_{n \in D_N} X_{n,N}^{(h)} \exp(-i < \lambda, n >) \right|^2 . \tag{27}$$

This means that the likelihood $\widetilde{\mathcal{L}}_N(\theta)$ is for stationary fields (where $f_\theta(u,\lambda)$ is independent of u) a generalization of the stationary Whittle likelihood with a tapered periodogram. The latter was introduced by Dahlhaus and Künsch (1987) to reduce the bias of the stationary Whittle estimate from order N^{-1} to the order N^{-2} leading to a central limit theorem for $\widehat{\theta}_N$ with rate $N^{-d/2}$ for dimensions $d = 1, 2, 3$. More precisely, the scorefunction in the stationary case has a bias of order N^{-1} which is reduced to $O(N^{-2})$ by using a data taper (see also Guyon (1995), Section 4.2). Unfortunately the situation is more complicated in the locally stationary setting. The scorefunction $\nabla\mathcal{L}_N(\theta_0)$ as in (14) with the preperiodogram as in (9) generalized to the random field case, has three potential sources of bias. The first one is the aforementioned bias of the preperiodogram close to the boundary of the field, a phenomenon which also appears in stationary random fields. This bias is of order N^{-1}, the proportion of boundary points of the field. By using a data taper, this bias is reduced to the order N^{-2} as in the stationary case (see Dahlhaus und Sahm, 2000).

The second source of bias is the 'skew' definition of the preperiodogram (9), which is oriented at the Wigner spectrum for nonstationary processes. The preperiodogram defined in (25) takes care of this skewness and reduces the bias due to the definition of the preperiodogram to the order N^{-2}. This is the reason why we use $\widetilde{\mathcal{L}}_N(\theta)$ instead of $\mathcal{L}_N(\theta)$ as in (14). From a technical point of view, it is even better to use the quadratic form $\frac{1}{N^d} X^T U X$ as the second part of the likelihood, where the matrix U is given in (31). This, however, destroys the nice representation of the likelihood as a sum over local likelihoods.

The third source of bias is the notion of local stationarity itself, since due to (22) the local spectral density characterizes the considered process only up to the order N^{-1}, which also results in a bias of order N^{-1}. We will address this problem again after the asymptotic results for the estimates based on (24) are presented. The assumptions needed are essentially the same as those in Section 2, we only assume a little more smoothness to achieve a more accurate bias control. A detailed proof of the following results can be found in Dahlhaus and Sahm (2000).

Assumption 4.2 *(i) We observe a realization $X_{n,N}$, $n \in D_N$ of a locally stationary Gaussian random fields with mean 0, transfer function A^0 and quickly decaying covariances:* $\mathrm{Cov}(X_{n,N}, X_{m,N}) = O((|n_i - m_i| + 1)^{-3})$ *for $i = 1, \ldots, d$. We fit a class of locally stationary Gaussian random fields with mean 0 and transfer function A_θ^0, $\theta \in \Theta \subset \mathbb{R}^p$, Θ compact.*

(ii) A_θ is differentiable with respect to θ, u and λ with uniformly continous derivatives $\nabla_{i,j,k} \frac{\partial^2}{\partial u^2} \frac{\partial^2}{\partial \lambda^2} A_\theta$; A is differentiable with respect to u and λ with uniformly continous derivatives $\frac{\partial^2}{\partial u^2} \frac{\partial^2}{\partial \lambda^2} A$.

(iii) $f = |A|^2$, $f_\theta = |A_\theta|^2$ are bounded from above and bounded away from 0 uniformly in θ, u and λ.

(iv) The taper h_N fulfils Assumption 2.3 (v) and is of proportion $\rho > 0$.

Theorem 4.3 *Let X_n, $n \in D_N$ be a locally stationary random field of dimension d and let assumptions (i)-(iv) hold. Furthermore f denotes the true spectral density of the process and we assume that*

$$\theta_0 = \arg\min_{\theta\in\Theta} \int_{[0,1]^d} h^2(u) \int_{(-\pi,\pi]^d} \Big[\log f_\theta(u,\lambda) + \frac{f(u,\lambda)}{f_\theta(u,\lambda)} \Big] \, d\lambda \, du \tag{28}$$

exists uniquely and lies in the interior of Θ.

The estimator $\widehat{\theta}_N := \arg\min_{\theta\in\Theta} \widetilde{\mathcal{L}}_N(\theta)$ has the following properties.

a) Consistency: $\widehat{\theta}_N - \theta_0 \xrightarrow{(p)} 0$.

b) Bias of the scorefunction: $\mathbb{E}\big(\nabla \widetilde{\mathcal{L}}_N(\theta_0) \big) = O(N^{-1})$.

c) Normal Law:

$$N^{d/2} \left(\widehat{\theta}_N - \theta_0 - \Gamma_h^{-1} \, \mathbb{E}\big(\nabla \widetilde{\mathcal{L}}_N(\theta_0) \big) \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_h^{-1} V_h \Gamma_h^{-1})$$

with

$$(V_h)_{ij} = \frac{1}{2(2\pi)^d} \int\limits_{[0,1]^d} h^4(u) \int\limits_{(-\pi,\pi]^d} f\,(\nabla_i f_{\theta_0}^{-1}) f\,(\nabla_j f_{\theta_0}^{-1}) \, d\lambda du \tag{29}$$

$$(\Gamma_h)_{ij} = \frac{1}{2(2\pi)^d} \int\limits_{[0,1]^d} h^2(u) \int\limits_{(-\pi,\pi]^d} \Big[(f - f_{\theta_0}) \nabla_{ij} f_{\theta_0}^{-1} + f_{\theta_0} (\nabla_i f_{\theta_0}^{-1}) \, f_{\theta_0} (\nabla_j f_{\theta_0}^{-1}) \Big] \, d\lambda du \,, \tag{30}$$

Remark 4.4 (i) The result also holds for the likelihood defined in (14) generalized to the random field case; for the important Theorem 4.5, however, the use of $\widetilde{\mathcal{L}}_N$ is essential.

(ii) If the taper proportion ρ is chosen to be of order $N^{-1/5}$ and the model is correctly specified, the asymptotic variance simplifies to V_1^{-1}, where V_1 is obtained by inserting the constant funtion $\chi_{[0,1]}$ for h.

(iii) The bias of the scorefunction of order N^{-1} should give rise to a bias of the estimator $\widehat{\theta_N}$ of the same order, so that the normal law would read $N^{d/2}(\widehat{\theta}_N - \mathbb{E}\widehat{\theta}_N) \to \mathcal{N}(0, \Gamma_h^{-1} V_h \Gamma_h^{-1})$. Due to the restrictive moment assumptions it should be possible to prove this using properties along the lines of Lemma 4.1 in Dahlhaus and Giraitis (1998). However, this slight improvement of the result does not seem to justify the rather large technical effort.

The problem of Theorem 4.3 is the bias. Since the bias is of the same or larger order than the normalizing constant in the normal law, Theorem 4.3 is not useful for practical application in dimensions $d > 1$. However, the only source of bias of order N^{-1} is the approximation (22), the remaining bias is of order N^{-2}. This means, if there is a second order approximation (as for example in (23)) available, this part of the bias can be calculated (if the model is correctly specified) and we obtain an applicable normal law for the important dimensions $d = 2, 3$.

Theorem 4.5 *Let X_n, $n \in D_N$ be a locally stationary random field in dimension $d \leq 3$ with mean 0 and transfer function A^0 and Assupmtions 4.2 be fulfilled.*
(1) If $A_n^{0,N}(\lambda) = A(\frac{n-1/2}{N}, \lambda)$ for all $n \in D_N$ then

$$N^{d/2} \,(\widehat{\theta}_N \,-\, \theta_0) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}(0, \Gamma_h^{-1}\, V_h\, \Gamma_h^{-1})$$

where the matrices Γ_h and V_h are given in (29) and (30).

(2) If $A_n^{0,N}(\lambda) = A(\frac{n-1/2}{N}, \lambda) + \frac{1}{N} B(\frac{n-1/2}{N}, \lambda) + O(N^{-2})$, where B fulfils Assumption (ii), then

$$N^{d/2} \,(\widehat{\theta}_N - \theta_0 - \frac{1}{N}\Gamma_h^{-1} M(\theta_0)) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}(0, \Gamma_h^{-1}\, V_h\, \Gamma_h^{-1})$$

where

$$M(\theta) \;=\; \frac{1}{(2\pi)^d} \int\limits_{[0,1]^d} h^2(u) \int\limits_{(-\pi,\pi]^d} A(u,\lambda)\, \overline{B(u,\lambda)}\, \nabla_\theta f_\theta^{-1}(u,\lambda)\, d\lambda\, du\,.$$

Remark 4.6 (i) If the model is correctly specified, the quantity Γ_h^{-1} can be consistently estimated by $\Gamma_h^{-1}(\widehat{\theta}_N)$, obtained by substituting $\widehat{\theta}_N$ for θ_0 in (30); $N^{d/2} - 1\, M(\theta_0)$ can be consistently estimated using

$$\widehat{M}(\widehat{\theta}_N) \;=\; \frac{1}{(2\pi)^d} \int\limits_{[0,1]^d} h^2(u) \int\limits_{(-\pi,\pi]^d} A_{\widehat{\theta}_N}(u,\lambda)\, \overline{B_{\widehat{\theta}_N}(u,\lambda)}\, \nabla_\theta f_{\widehat{\theta}_N}^{-1}(u,\lambda)\, d\lambda\, du\,.$$

If moreover the taper proportion ρ is of order $N^{-1/5}$, one can use the corrected estimator $\theta_N^* = \widehat{\theta}_N - N^{-1} V_1^{-1}(\widehat{\theta}_N)\, M(\widehat{\theta}_N, \widehat{\theta}_N)$ and Theorem 4.5 simplifies to

$$N^{d/2}\,(\theta_N^* - \theta_0) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}(0, V_1^{-1})\,.$$

(ii) Theorem 4.5 yields the desired normal law for the important dimensions $d = 2, 3$. For higher dimensions the asymptotics generally fails due to additional bias of order N^{-2}. Here an alternative would be to consider the scorefunction

$$S_N(\theta) \;=\; \frac{1}{N^d}\,\mathrm{tr}\,\Big(U_N(\nabla f_\theta^{-1})\,(X\,X^T - \Sigma_\theta)\Big)$$

($U_N(f_\theta^{-1})$ as defined in (32)) and the estimate $\widetilde{\theta}$ defined by $S_N(\widetilde{\theta}) = 0$. This scorefunction is clearly inspired by the Whittle likelihood, but it is by definition unbiased. For estimates based on this scorefunction, one easily obtains a normal law in any dimension d (analogously to the proof for the Whittle likelihood). However, this procedure requires knowledge of the exact covariance matrix of the process, which can be very difficult to obtain to the desired precision; in the worst case, the covariance would have to be simulated, leading to some time-consuming annealing scheme. Another possibility is to apply a Newton step starting from $\widehat{\theta}_N$, i.e. to use

$$\theta_N^* \;=\; \widehat{\theta}_N - \nabla S_N(\widehat{\theta}_N)\,S_N(\widehat{\theta}_N)$$

as an estimator. This estimator has the required normal limit for $d \leq 3$, for d larger one needs multiple iterations of the Newton step.

5. Generalized Toeplitz Matrices

In this section we show how the local likelihood is derived by a certain approximation of the inverse of the covariance matrix together with a generalized version of the Szegö identity. For simplicity we restrict ourselves again to the time series case $d = 1$. Similar results also hold for the random field case (cf. Dahlhaus and Sahm, 2000, Section 4).

Note that in the stationary case the covariance matrix Σ is equal to the Toeplitz matrix

$$B_T(f) = \Big\{\int_{-\pi}^{\pi} \exp(i\lambda(r-s))f(\lambda)d\lambda\Big\}_{r,s=1,\ldots,T}$$

where $f(\lambda)$ is the spectral density of the process. It is well known (cf. Grenander and Szegö, 1958) that $B_T(f)$ can be approximated by $B_T(\frac{1}{4\pi^2}f^{-1})$ which leads together with the Szegö identity

$$\frac{1}{T}\log\det\Sigma = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log[2\pi f(\lambda)]d\lambda + O(T^{-1})$$

in the case of $f = f_\theta$ to the Whittle likelihood (13) as an approximation of $-\frac{1}{T}\log$ Gaussian likelihood.

In the locally stationary case we obtain for the covariance matrix Σ

$$\begin{aligned}\Sigma_{r,s} \;&=\; \mathrm{cov}(X_{r,T}, X_{s,T})\\ &=\; \int_{-\pi}^{\pi}\exp\{i\lambda(r-s)\}A\left(\frac{r-1/2}{T},\lambda\right)\overline{A\left(\frac{s-1/2}{T},\lambda\right)}d\lambda + O(T^{-1}).\end{aligned}$$

If A is sufficiently smooth in time this implies

$$\Sigma_{r,s} = \int_{-\pi}^{\pi} \exp(i\lambda(r-s)) f\left(\frac{r+s-1}{2T}, \lambda\right) d\lambda + O(T^{-1}).$$

The above approximation of Toeplitz matrices now suggests to use

$$\left\{ \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \exp\{i\lambda(r-s)\} f\left(\frac{r+s-1}{2T}, \lambda\right)^{-1} d\lambda \right\}_{r,s=1,\ldots,T} \tag{31}$$

as an approximation of Σ^{-1} in (19) in the nonstationary case. Since it leads to a slightly nicer criterion we use instead $U_T(\frac{1}{4\pi^2} f^{-1})$ where

$$U_T(\phi) = \left\{ \int_{-\pi}^{\pi} \exp\{i\lambda(r-s)\} \phi\left(\frac{1}{T}\left[\frac{r+s}{2}\right]^o - \frac{1}{2T}, \lambda\right) d\lambda \right\}_{r,s=1,\ldots,T} \tag{32}$$

and $[x]^o$ denotes the smallest integer larger or equal to x. Note that $U_T(\frac{1}{4\pi^2} f^{-1}) = B_T(\frac{1}{4\pi^2} f^{-1})$ is the classical Toeplitz-approximation if f is constant over time (stationary case). We now use this approximation, i.e.

$$\Sigma_\theta^{-1} \approx U_T\left(\frac{1}{4\pi^2} f_\theta^{-1}\right)$$

to approximate the second part of the likelihood in (19). We obtain with the substitution $[\frac{r+s}{2}]^o = t$ and $r - s = k$

$$\begin{aligned}
\underline{X}' U_T(f_\theta^{-1}) \underline{X} &= \sum_{r,s=1}^{T} X_{r,T} X_{s,T} \int_{-\pi}^{\pi} \exp\{i\lambda(r-s)\} f_\theta\left(\frac{1}{T}\left[\frac{r+s}{2}\right]^o - \frac{1}{2T}, \lambda\right)^{-1} d\lambda \\
&= \sum_{t=1}^{T} \int_{-\pi}^{\pi} f_\theta\left(\frac{t-1/2}{T}, \lambda\right)^{-1} \sum_k X_{[t+1/2+k/2],T} X_{[t+1/2-k/2],T} \exp(i\lambda k) d\lambda \\
&= 2\pi \sum_{t=1}^{T} \int_{-\pi}^{\pi} f_\theta\left(\frac{t-1/2}{T}, \lambda\right)^{-1} J_T\left(\frac{t-1/2}{T}, \lambda\right) d\lambda,
\end{aligned}$$

i.e. the second part of the local likelihood $\mathcal{L}_T(\theta)$ as defined in (19) (with $h(u) = \chi_{[0,1]}(u)$).The first part of the approximation follows from a generalization of Szegö's formula to the nonstationary case (see Proposition 2.5 in Dahlhaus, 2000)

$$\frac{1}{T} \log \det \Sigma_\theta = \frac{1}{2\pi} \int_0^1 \int_{-\pi}^{\pi} \log[2\pi f_\theta(u, \lambda)] d\lambda du + O(T^{-1+\varepsilon})$$

with $\varepsilon > 0$. That $U_T(\{4\pi^2 f\}^{-1})$ is a good approximation to Σ^{-1} is made precise in Proposition 2.4 of Dahlhaus (2000) where it is proved under suitable regularity conditions that

$$\frac{1}{T} \left\| \Sigma_T^{-1} - U_T(\{4\pi^2 f\}^{-1}) \right\|^2 = O(T^{-1+\varepsilon}) \tag{33}$$

and

$$\frac{1}{T} \left\| U_T(\phi)^{-1} - U_T(\{4\pi^2\phi\}^{-1}) \right\|^2 = O(T^{-1+\varepsilon}).$$

for $\varepsilon > 0$ There exist a large number of additional results for these generalized Toeplitz matrices, e.g. on norms and matrix products (cf. Dahlhaus, 2000, Appendix A).

The tapered likelihood $\mathcal{L}_T^{(h)}(\theta)$ arises as an approximation of the exact Gaussian likelihood $\tilde{\mathcal{L}}_T(\theta)$ if one uses $U_T^{(h)}(\{4\pi^{-2}f\}^{-1})$ instead of $U_T(\{4\pi^2 f\}^{-1})$ as an approximation of Σ^{-1} where

$$U_T^{(h)}(\phi) = I_T^{(h)} U_T(\phi) I_T^{(h)}$$

and

$$I_T^{(h)} = \mathrm{diag}\{h_T(\frac{1/2}{T}), \ldots, h_T\Big(\frac{T-1/2}{T}\Big)\}$$

with $h_T(u)$ as in (9). In the stationary case (where $U_T(\phi) = B_T(\phi)$) it has been proved in Dahlhaus (1990) that this is indeed a better approximation of Σ^{-1} if the taper is asymptotically vanishing. Heuristically, the classical approximation without taper is particularly bad at the edges of the matrix. This is improved by downweighting the edges with a taper. Thus the use of a data taper in the Whittle-likelihood means using an improved approximation of Σ^{-1} and not downweighting the observations at the edges of the observation domain (as it looks from a first view).

All considerations of this section also hold in the random field case (cf. Dahlhaus and Sahm, 2000).

6. References

Brockwell, P.J. and Davis, R.A. (1987). *Time Series*, Springer, New York.

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

Dahlhaus, R. (1990) Approximations for the inverse of Toeplitz matrices with applications to stationary processes. *Linear Alg. Appl.* 127, 27-40.

Dahlhaus, R. (1996a). On the Kullback-Leibler information divergence of locally stationary processes. *Stoch. Proc. Appl.* **62**, 139-168.

Dahlhaus, R. (1996b). Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In *Athens conference on Applied Probability and Time Series Vol. II* (P.M. Robinson and M. Rosenblatt, eds.), **145-159**, Lecture Notes in Statistics 115, Springer, New York.

Dahlhaus, R. (2000). A likelihood approximation for locally stationary processes. To appear in *Ann. Statist.*.

DAHLHAUS, R. AND GIRAITIS, L. (1998). On the optimal segment length for parameter estimates for locally stationary time series. *J. Time Ser. Anal.*, **19**, 629-655.

DAHLHAUS, R. AND KÜNSCH, H. (1987). Edge effects and efficient parameter estimation for stationary random fields. *Biometrika* **74**, 877-882.

DAHLHAUS, R. AND NEUMANN, M.H. (2000). Locally adaptive fitting of semiparametric models to nonstationary time series. To appear in *Stoch. Proc. Appl.*

DAHLHAUS, R.AND SAHM, M. (2000). Locally stationary random fields. Submitted for publication.

GRENANDER, U. AND SZEGÖ, G. (1958). *Toeplitz Forms and their Applications.* University of California Press, Berkeley.

GUYON, X. (1995). *Random Fields on a Network.* Springer, New York.

MALLAT, S. PAPANICOLAOU, G. AND ZHANG, Z. (1998). Adaptive covariance estimation of locally stationary processes. *Ann. Statist.* **26**, 1-47.

NEUMANN, M. AND VON SACHS, R. (1997). Wavelet thresholding in anisotropic function classes and adaptive estimation of evolutionary spectra. *Ann. Statist.*, **25**,38-76.

PRIESTLEY, (1965). Evolutionary spectra and non-stationary processes. *J. Roy. Statist. Soc. Ser. B* **27**, 201-237.

WHITTLE, P. (1953). Some recent contributions to the theory of stationary processes. Appendix to *A study in the analysis of stationary time series* by H. Wold, 2nd ed., 196-228. Almqvist and Wiksell, Uppsala.

WHITTLE, P. (1954). Estimation and information in stationary time series. *Ark. Math.* **2**, 423-434.

Rainer Dahlhaus and **Michael Sahm**
Inst. f. Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
D - 69120 Heidelberg
dahlhaus@statlab.uni-heidelberg.de
msahm@statlab.uni-heidelberg.de
Germany