Stochastic Processes and their Applications 91 (2001) 277–308
stochastic processes and their applications
www.elsevier.com/locate/spa

Locally adaptive fitting of semiparametric models to nonstationary time series

Rainer Dahlhaus[a,*], Michael H. Neumann[b]

[a] *Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany*

[b] *Humboldt-Universität zu Berlin, Sonderforschungsbereich 373, Spandauer Straße 1, D-10178 Berlin, Germany*

Received 24 June 1999; received in revised form 17 July 2000; accepted 19 July 2000

Abstract

We fit a class of semiparametric models to a nonstationary process. This class is parametrized by a mean function $\mu(\cdot)$ and a p-dimensional function $\theta(\cdot) = (\theta^{(1)}(\cdot), \ldots, \theta^{(p)}(\cdot))'$ that parametrizes the time-varying spectral density $f_{\theta(\cdot)}(\lambda)$. Whereas the mean function is estimated by a usual kernel estimator, each component of $\theta(\cdot)$ is estimated by a nonlinear wavelet method. According to a truncated wavelet series expansion of $\theta^{(i)}(\cdot)$, we define empirical versions of the corresponding wavelet coefficients by minimizing an empirical version of the Kullback–Leibler distance. In the main smoothing step, we perform nonlinear thresholding on these coefficients, which finally provides a locally adaptive estimator of $\theta^{(i)}(\cdot)$. This method is fully automatic and adapts to different smoothness classes. It is shown that usual rates of convergence in Besov smoothness classes are attained up to a logarithmic factor. © 2001 Elsevier Science B.V. All rights reserved.

MSC: primary 62M10; secondary 62F10

Keywords: Locally stationary processes; Nonlinear thresholding; Nonparametric curve estimation; Preperiodogram; Time series; Wavelet estimators

1. Introduction

Stationarity of the observations is crucial for the validity of proposed methods in the vast majority of papers in time-series analysis. However, this assumption is often violated in practical applications. In the present paper we develop methodology to fit semiparametric models which, in particular, allow for modeling the time-varying behavior of the process. In order to present a rigorous asymptotic theory, we suppose that the observations $X_1, \ldots, X_T$ stem from a locally stationary process as defined in Dahlhaus (1997) – see Definition 2.1 below. The main idea consists of a rescaling

* Corresponding author. Fax: +49-6221-545331.
E-mail addresses: dahlhaus@statlab.uni-heidelberg.de (R. Dahlhaus), neumann@wiwi.hu-berlin.de (M.H. Neumann).

0304-4149/01/$ - see front matter
PII: S0304-4149(00)00060-0

of the time axis to the interval [0,1]. By imposing smoothness conditions on certain parameter functions in this rescaled time we can develop an asymptotic theory for estimates of these parameter functions.

The model to be fitted is characterized by the mean function $\mu(u)$ and a p-dimensional parameter function $\theta(u)$, $u \in [0,1]$, which defines the time-varying spectral density $f_{\theta(u)}(\lambda)$. An example is the time-varying ARMA process

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

where $a_0(\cdot) \equiv b_0(\cdot) \equiv 1$, $\varepsilon_{t,T} \sim \mathcal{N}(0, \sigma(t/T)^2)$, $\theta(u) = (a_1(u), \ldots, a_r(u), b_1(u), \ldots, b_s(u), \sigma^2(u))$, and

$$f_{\theta(u)}(\lambda) = \frac{\sigma^2(u)\,|\sum_{j=0}^{s} b_j(u)\exp(\mathrm{i}\lambda j)|^2}{2\pi\,|\sum_{j=0}^{r} a_j(u)\exp(\mathrm{i}\lambda j)|^2}$$

(cf. Dahlhaus 1996a). Consistent estimation of these functions is possible under appropriate restrictions on the class of possible models. Accordingly, we impose smoothness conditions on the functions μ and θ.

The estimation of μ does not cause any substantial problems and can be carried out by "direct methods". It may be done, as proposed in the present article, by a usual kernel estimator. In contrast, the estimation of θ is much more involved. The reason is that we do not observe $\theta(\cdot)$ "directly" (as in nonparametric regression where we observe $\theta(\cdot)$ plus some noise). Instead all characteristics of the process (such as $f_{\theta(u)}(\lambda)$) usually depend on the parameter curves in a highly nonlinear way. We therefore suggest in Section 3 a minimum distance method for the estimation of $\theta(\cdot)$ which is based on a distance between the time-varying spectral density and some nonparametric estimate of it. This raises the problem of finding a suitable nonparametric estimate. One may certainly use usual periodograms on small segments, as proposed in Dahlhaus (1997) and von Sachs and Schneider (1996). However, such local periodograms contain an additional parameter, the segment length N, which acts like a smoothing parameter in time direction (see Section 2). This means that beside the major smoothing step of nonlinear wavelet thresholding (which we aim at) we have an additional nonadaptive smoothing step at this preliminary stage. A possible alternative is given by the so-called preperiodogram proposed in Neumann and von Sachs (1997) – see also Dahlhaus (2000). Motivated by the convergence result (2.4), this preperiodogram has the form

$$J_T^{\tilde{\mu}}(u,\lambda) = \frac{1}{2\pi} \sum_{\substack{k:k\in\mathbb{Z},1\leqslant[uT+0.5-k/2]\leqslant T \\ \text{and } 1\leqslant[uT+0.5+k/2]\leqslant T}} \left(X_{[uT+0.5-k/2],T} - \tilde{\mu}\left(\frac{[uT+0.5-k/2]}{T}\right)\right) \times \left(X_{[uT+0.5+k/2],T} - \tilde{\mu}\left(\frac{[uT+0.5+k/2]}{T}\right)\right)\exp(-\mathrm{i}k\lambda), \tag{1.1}$$

where $\tilde{\mu}(\cdot)$ is some estimator of the mean function and $[x]$ is the largest integer smaller or equal to x. The preperiodogram has a similar structure as the Wigner–Ville spectrum (cf. Martin and Flandrin, 1985). It may be regarded as a raw estimate of the spectral density at time u and frequency λ. However, in order to obtain a consistent estimator smoothing in time and frequency is necessary. The obvious advantage of this definition is that it does not contain any implicit smoothing, neither in frequency nor in time. Consequently, the decision about the degree of smoothing in each of these directions is left to the major smoothing step itself.

An estimate for the parameter curves $\theta(\cdot)$ can now be obtained by a minimum distance fit of $f_{\theta(\cdot)}(\lambda)$ to $J_T^{\tilde{\mu}}(\cdot,\lambda)$ together with some smoothing. Dahlhaus (2000) has used this strategy in the finite-dimensional case, i.e. in the case where the parameter curves themselves are parameterized. In this paper we show how this approach together with a truncated wavelet series approximation can be used in the case of strong inhomogeneous smoothness of $\theta(\cdot)$ – for example, for piecewise smooth functions with jumps. In order to adapt the degree of smoothing to the local smoothness characteristics of the curves we apply a nonlinear regularization method based on a truncated wavelet series approximation of the parameter functions $\theta^{(i)}$, that is

$$\theta^{(i)}(u) = \sum_k \alpha_{l,k;i}\phi_{l,k}(u) + \sum_{l\leqslant j<j^*}\sum_k \beta_{j,k;i}\psi_{j,k}(u).$$

Here $\{\phi_{l,k}\}_k \cup \{\psi_{j,k}\}_{j\geqslant l;k}$ forms an orthonormal wavelet basis of $L_2[0,1]$. As a starting point, we obtain empirical versions, $\tilde{\alpha}_{l,k;i}$ and $\tilde{\beta}_{j,k;i}$, of the coefficients $\alpha_{l,k;i}$ and $\beta_{j,k;i}$, respectively, by minimizing some empirical distance measure. In order to get a locally (in u) (near-)optimal degree of smoothing, we intend to apply nonlinear thresholding of these empirical coefficients. This method was introduced into statistics by Donoho and Johnstone in a series of papers (see, for example, Donoho and Johnstone, 1998). Nonlinear thresholding is roughly the same as a pre-test estimator, where the significance of each individual coefficient is separately tested in a previous step.

To carry out this plan, we need knowledge about the stochastic behavior of $\tilde{\beta}_{j,k;i} - \beta_{j,k;i}$. It will turn out that $\tilde{\beta}_{j,k;i}$ can be approximated by a certain quadratic form in $(X_{1,T} - \mathbb{E}X_{1,T}),\ldots,(X_{T,T} - \mathbb{E}X_{T,T})$. On the basis of cumulant methods and some simple truncation technique, we derive a general result on the asymptotic normality of such quadratic forms in terms of probabilities of large deviations. This result is used to derive estimates for the tail behavior of the empirical coefficients $\tilde{\beta}_{j,k;i}$. This allows us to act in the same way as in the case of normally distributed wavelet coefficients. Hence, we can transfer methodology developed in several papers by Donoho and Johnstone to our particular estimation problem. The empirical wavelet coefficients are treated by nonlinear rules, which finally provide near-optimal estimators of $\theta^{(i)}$ after applying the inverse wavelet transform. If the fitted model is an AR(p)-model with time varying coefficients the estimate is very similar to the one obtained in Dahlhaus et al. (1999). This paper also contains two examples with simulations which demonstrate the behavior of the estimate.

The paper is organized as follows. In Section 2 we recall the definition of locally stationary processes and discuss two versions of local periodograms. In Section 3 we describe the construction of empirical versions of the wavelet coefficients that correspond to a wavelet expansion of the parameter curves $\theta^{(i)}(\cdot)$. Nonlinear thresholding of these coefficients, which finally leads to a locally adaptive estimate for $\theta^{(i)}$, is described in Section 4. In order to preserve a readable structure of this paper, most of the technicalities are deferred to the appendix. Part I of this appendix contains some technical results on asymptotic normality of quadratic forms while Part II contains the proofs of the assertions.

2. Some tools for a local analysis of nonstationary processes

Assume we observe a stretch $X_1,\dots,X_T$ of a possibly nonstationary time series and we intend to fit a certain semiparametric model to this process. As always in nonparametric curve estimation, desirable properties like consistency can only be derived under certain restrictions on the complexity of the object of interest. An appropriate framework, which allows for a rigorous asymptotic treatment of nonstationary time series, is the following model for locally stationary processes (cf. Dahlhaus, 1997). We cite this definition, which generalizes the Cramér representation of a stationary stochastic process.

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

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} A^o_{t,T}(\omega)\exp(\mathrm{i}\omega t)\,\mathrm{d}\xi(\omega), \tag{2.1}$$

where

(i) $\xi(\omega)$ is a stochastic process on $[-\pi,\pi]$ with $\overline{\xi(\omega)} = \xi(-\omega)$, $\mathbb{E}\xi(\omega) = 0$ and orthonormal increments, i.e.,

$$\mathrm{cum}\{\mathrm{d}\xi(\omega_1),\dots,\mathrm{d}\xi(\omega_k)\} = \eta\left(\sum_{j=1}^{k}\omega_j\right) h_k(\omega_1,\dots,\omega_{k-1})\,\mathrm{d}\omega_1,\dots\mathrm{d}\omega_k,$$

where cum$\{\dots\}$ denotes the cumulant of order k, $|h_k(\omega_1,\dots,\omega_{k-1})| \leqslant \mathrm{const.}_k$ for all k (with $h_1 = 0, h_2(\omega) = 1$) and $\eta(\omega) = \sum_{j=-\infty}^{\infty}\delta(\omega + 2\pi j)$ is the period 2π extension of the Dirac delta function (Dirac comb), and where

(ii) there exists a positive constant K and a smooth function $A(u,\omega)$ on $[0,1]\times[-\pi,\pi]$ which is 2π-periodic in ω, with $A(u,-\omega) = \overline{A(u,\omega)}$, such that for all T,

$$\sup_{t,\omega}|A^o_{t,T}(\omega) - A(t/T,\omega)| \leqslant KT^{-1}. \tag{2.2}$$

$A(u,\omega)$ and $\mu(u)$ are assumed to be continuous in u.

The smoothness of A and μ in u restricts the departure from stationarity and ensures the locally stationary behavior of the process. A detailed motivation and discussion of the model is contained in Dahlhaus (1996a). It allows also for a reasonable definition of a time-varying spectral density.

Definition 2.2. As the *time-varying spectral density* of $\{X_{t,T}\}$ given by (2.1) we define, for $u \in (0,1)$,

$$f(u,\omega) := |A(u,\omega)|^2. \tag{2.3}$$

By Dahlhaus (1996a, Theorem 2.2) if $A(u,\omega)$ is uniformly Lipschitz in u and ω with index $\alpha > 1/2$, then, for all $u \in (0,1)$,

$$\int_{-\pi}^{\pi} \left| f(u,\lambda) - \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \operatorname{cov}(X_{[uT+0.5-k/2],T}, X_{[uT+0.5+k/2],T}) \exp(-\mathrm{i}\lambda k) \right|^2 \mathrm{d}\lambda = \mathrm{o}(1) \tag{2.4}$$

which establishes the link between the Definitions 2.1 and 2.2.

2.1. Preperiodogram versus periodogram on small segments

In this section we assume for simplicity that $\mu \equiv 0$. The properties of the preperiodogram J_T in relation to the ordinary periodogram I_T can be best understood by the following relation (cf. Dahlhaus 2000):

$$\begin{aligned} I_T(\lambda) &= \frac{1}{2\pi T} \left| \sum_{t=1}^{T} X_{t,T} \exp(-\mathrm{i}\lambda t) \right|^2 \\ &= \frac{1}{2\pi} \sum_{k=-(T-1)}^{T-1} \left(\frac{1}{T} \sum_{t=1}^{T-|k|} X_{t,T} X_{t+|k|,T} \right) \exp(-\mathrm{i}\lambda k) \\ &= \frac{1}{T} \sum_{t=1}^{T} \frac{1}{2\pi} \sum_{\substack{k: k\in\mathbb{Z}, 1 \leqslant [t+0.5-k/2] \leqslant T \\ \text{and } 1 \leqslant [t+0.5+k/2] \leqslant T}} X_{[t+0.5-k/2],T} X_{[t+0.5+k/2],T} \exp(-\mathrm{i}\lambda k) \\ &= \frac{1}{T} \sum_{t=1}^{T} J_T\left(\frac{t}{T}, \lambda\right). \end{aligned} \tag{2.5}$$

Thus, the periodogram over the whole stretch of data is the average of the preperiodogram over time. While the periodogram is the Fourier transform of the covariance estimator of lag k over the whole segment the preperiodogram $J_T(t/T, \lambda)$ just uses the single product $X_{[t+0.5-k/2],T} X_{[t+0.5+k/2],T}$ as a kind of “local estimator” of the covariance of lag k at time t (note that $[t+0.5+k/2] - [t+0.5-k/2] = k$).

A classical kernel estimator of the spectral density of a stationary process at some frequency λ_0 therefore can be regarded as an average of the preperiodogram over the frequencies in the neighborhood of λ_0 and all time points which is justified if the process is stationary. It is therefore plausible that averaging the preperiodogram around some frequency λ_0 and some time point u_0 gives a consistent estimate of the time-varying spectrum $f(u_0, \lambda)$.

For a locally stationary process the preperiodogram is asymptotically unbiased but has a diverging variance as $T \to \infty$. Thus, smoothing in time and frequency is essential to make a consistent estimate out of it. Beside the two-dimensional kernel estimate mentioned above, we may apply a local polynomial fit in time and frequency or even a nonlinear wavelet method in both directions. The latter approach has been studied in Neumann and von Sachs (1997) where it was shown that the resulting estimate has the optimal rate of convergence in anisotropic function classes up to a logarithmic factor.

A possible alternative for locally stationary processes seems to be to start with an ordinary periodogram over small segments. This has been proposed by Dahlhaus (1997) for the purpose of fitting parametric models to nonstationary processes and in von Sachs and Schneider (1996) as a starting point for a wavelet estimator of the time-varying spectral density. In the nontapered case, such a local periodogram has the form

$$I_N(u,\lambda) = \frac{1}{2\pi N}\left|\sum_{k=1}^{N} X_{[uT-N/2+k],T}\exp(-\mathrm{i}\lambda k)\right|^2.$$

Note, that the parameter $N = N(T)$, which is usually assumed to obey $N \to \infty$ and $N/T \to 0$ as $T \to \infty$, acts in two ways: First, as (2.5) shows it delivers a cut-off point, from which on covariances of higher lags are excluded. Hence a small value of N introduces some bias. Second, $I_N(u,\lambda)$ already contains some implicit smoothing in time: it is equivalent to a kernel estimate based on a modified preperiodogram (with N instead of T) around u with bandwidth N/T (with a rectangular kernel – with tapering one gets a smooth kernel!).

The use of $I_N(u,\lambda)$ as a starting point is reasonable as long as the degree of smoothing in time in the following smoothing step exceeds the degree of smoothing implicitly contained in the definition of $I_N(u,\lambda)$ (e.g. if a kernel with bandwidth $b \gg N/T$ is used) or if the smoothing in time direction is purely controlled by the parameter N and in addition only some smoothing in frequency direction is applied (e.g. a kernel estimate as in Dahlhaus, 1996b). However, since $I_N(u,\lambda)$ is nearly the same as our preperiodogram $J_T(t/T,\lambda)$ smoothed with a rectangular kernel we cannot make full use of smoothness of higher order of $f(u,\lambda)$ in time direction. Moreover, problems clearly arise if an adaptive method is used in the second step (as in the present paper). For that reason we prefer the preperiodogram where we have full control over the smoothing in the second step. Below we use the preperiodogram in a minimum distance functional to obtain estimates of the parameter curves $\theta(\cdot)$.

From here on, we consider the general case of an unknown mean function μ. A particular estimate for μ is given by the kernel estimator introduced in Gasser and Müller (1979),

$$\tilde{\mu}(u) = \sum_{t=1}^{T}\left[\int_{(t-1)/T}^{t/T} \frac{1}{b}K\left(\frac{u-v}{b}\right)\,\mathrm{d}v\right] X_{t,T}. \tag{2.6}$$

Here b denotes the bandwidth and K is an ordinary kernel with support $[-1,1]$, if $0 \leqslant u-b < u+b \leqslant 1$, and a boundary kernel otherwise. We can of course also use other nonparametric estimates here, as for example local polynomial estimates. From now on we use the preperiodogram $J_T^{\tilde{\mu}}(u,\lambda)$ with mean corrected data as defined in (1.1).

2.2. Some properties of the preperiodogram

To reduce the burden of technicalities in the following sections, we investigate first some stochastic properties of $J_T^{\tilde{\mu}}(u,\lambda)$. In what follows, we have to deal with quantities of the form

$$F(J_T^{\tilde{\mu}}) = \int_0^1 q(u) \int_{-\pi}^{\pi} p(u,\lambda) J_T^{\tilde{\mu}}(u,\lambda)\, \mathrm{d}\lambda\, \mathrm{d}u,$$

where $q(u)$ may depend on T. $F(J_T^{\tilde{\mu}})$ describes the asymptotic behavior of the wavelet coefficients, cf. (B.7) below. $q(u)$ typically plays the role of a wavelet while $p(u,\lambda)$ is some function of the spectral density. The result stated in Lemma 2.1 below is of interest beyond its particular application in this paper. For example, if q is a kernel function and $p(u,\lambda) = \exp(\mathrm{i}\lambda k)$ then $F(J_T^{\tilde{\mu}})$ is an estimate of the local covariance function of lag k. We will show below that $F(J_T^{\tilde{\mu}})$ is asymptotically equivalent to $F(J_T^{\mu})$ and that these quantities are asymptotically normally distributed in terms of probabilities of large deviations. Before we investigate $F(J_T^{\tilde{\mu}})$ and $F(J_T^{\mu})$, we introduce a convenient notion, which is slightly stronger than the usual $\mathrm{O_P}$.

Definition 2.3. We write

$$Z_T = \tilde{O}(\eta_T)$$

if for each $\lambda < \infty$ there exists a $C = C(\lambda) < \infty$ such that

$$P(|Z_T| > C\eta_T) \leqslant CT^{-\lambda}.$$

The statement $Z_T = \tilde{\mathrm{O}}(\eta_T)$ describes the fact that Z_T is $\mathrm{O}(\eta_T)$ with a probability exceeding $1 - \mathrm{O}(T^{-\lambda})$. Here and in the following we use the convention that λ in the exponent of T denotes an arbitrarily large and δ an arbitrarily small coefficient.

To derive some useful stochastic properties of $F(J_T^{\tilde{\mu}})$ we use the following assumptions:

(A1) $\{X_{t,T}\}$ is a locally stationary process and $A(u,\lambda)$ is Lipschitz continuous in u.
(A2) (i) Assume that for all $L < \infty$ there exists a $K_L < \infty$ such that

$$\mathbb{E}|X_{t,T}|^L \leqslant K_L$$

(ii) $\{X_{t,T}\}$ is α-mixing uniformly in T with mixing coefficients

$$\alpha(s) \leqslant C_1 \exp(-C_2|s|).$$

(A3) μ is r times differentiable with $|\mu^{(r)}(x)| \leqslant C$ and the kernel K is of order r.

Remark 1. Assumptions (A2) and (A3) are used as follows. Lemma 2.1 below states asymptotic normality for a functional of the preperiodogram in terms of probabilities of large deviations. Such results are usually derived by means of cumulant techniques under the assumption that there exist constants $M < \infty$ and $\gamma < \infty$ such that

$$\sum_{t_2,\ldots,t_k} |\mathrm{cum}(X_{t_1,T},\ldots,X_{t_k,T})| \leqslant M^k (k!)^{1+\gamma}; \tag{2.7}$$

see for example Saulis and Statulevicius (1991, Lemmas 2.3 and 2.4). Since such a quantitative condition to hold simultaneously for all cumulants is unnecessarily restrictive we assume instead (A2) which yields an upper bound similar to (2.7) by a result of Statulevicius and Jakimavicius (1988) and a simple truncation argument (for details, see Lemma A.1 in the appendix).

Assumption (A3) leads to a certain decay of the bias of a nonparametric kernel estimator of μ which implies that the difference between $F(J_T^{\tilde{\mu}})$ and $F(J_T^{\mu})$ is asymptotically negligible.

Lemma 2.1. *Suppose that* (A1), (A2),

$$\left\{ \sum_k |k| \sup_{u\in[0,1]} \left| \int_{-\pi}^{\pi} p(u,\lambda)\exp(-\mathrm{i}k\lambda)\,\mathrm{d}\lambda \right| \right\} < \infty,$$

$$\|q\|_1 \|q\|_\infty = \mathrm{O}(1)$$

as well as

$$\|q\|_\infty = \mathrm{O}(T^{1/2-\delta})$$

for any $\delta > 0$, *are fulfilled. The asymptotic variance of* $F(J_T^{\mu})$ *is given by*

$$\sigma_T^2 = 2\pi T^{-1} \left\{ \int_0^1 \left[\int_{-\pi}^{\pi} 2|q(u)p(u,\lambda)f(u,\lambda)|^2\,\mathrm{d}\lambda \right.\right.$$
$$\left.\left. + \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} q(u)p(u,\lambda)\overline{q(u)p(u,-\mu)}f(u,\lambda)f(u,\mu)h_4(\lambda,-\lambda,\mu)\,\mathrm{d}\lambda\,\mathrm{d}\mu \right] \mathrm{d}u \right\}.$$

(i) *If* $\sigma_T \geqslant C_0 T^{-1/2}$ *for some* $C_0 > 0$, *then*

$$P(\pm[F(J_T^{\mu}) - \mathbb{E}F(J_T^{\mu})] \geqslant \sigma_T x) = (1-\Phi(x))(1+\mathrm{o}(1)) + \mathrm{O}(T^{-\lambda})$$

holds uniformly in $x \in \mathbb{R}$, *where* Φ *denotes the cumulative distribution function of a standard normal random variable.*

(ii) *If* $\sigma_T = \mathrm{O}(T^{-1/2})$, *then, for* $\bar{\sigma}_T = \max\{\sigma_T, C_0/\sqrt{T}\}$ *and arbitrary* $C_0 > 0$,

$$P(\pm[F(J_T^{\mu}) - \mathbb{E}F(J_T^{\mu})] \geqslant \bar{\sigma}_T x) \leqslant 2(1-\Phi(x))(1+\mathrm{o}(1)) + \mathrm{O}(T^{-\lambda})$$

holds uniformly in $x \in \mathbb{R}$.

(iii) $\mathbb{E}F(J_T^{\mu}) - F(f) = \mathrm{O}(\|q\|_\infty T^{-1})$.

(iv) *If additionally* (A3) *is satisfied, then*

$$F(J_T^{\tilde{\mu}}) - F(J_T^{\mu})$$
$$= \tilde{O}(\|q\|_2 T^{-1/2}\{[(Tb)^{-1/2} + b^r]\sqrt{\log T} + b^r T^{\delta-1/2}\|q\|_1\|q\|_\infty^2/\|q\|_2^3\}$$
$$+ \|q\|_1[(Tb)^{-1} + b^{2r}]),$$

where b *denotes the bandwidth of the estimator for* $\tilde{\mu}$.

Part (i) states asymptotic normality in terms of probabilities of large deviations in the case that there is a favorable relation between σ_T and our upper estimates for the cumulants. If such a relation is not guaranteed, we can still show that the quadratic forms "behave not worse" than a Gaussian random variable (see (ii)). Assertion (iii) provides an estimate for the bias, $\mathbb{E}F(J_T^{\mu}) - F(f)$. Typically, we have in this article that this bias is of order $\mathrm{o}(T^{-1/2})$, and therefore negligible. Finally, according to (iv), the effect of estimating μ is also of negligible order.

In the Gaussian case a slightly different large deviation result for the statistic J_T^{μ} has recently also been obtained by Zani (1999).

3. Fitting a semiparametric model

As the mean and the spectral density being important characteristics of a stationary time series, so the mean function and the time-varying spectral density are central quantities of a locally stationary process. In this article we study the fitting of a semiparametric model parametrized by the mean function $\mu(\cdot)$ and a parameter curve $\theta:[0,1] \to \Theta \subseteq \mathbb{R}^p$ that defines a time-varying spectral density $f_{\theta(u)}(\lambda)$. The special case of fitting an AR(p) process with time-varying coefficients has been considered by Dahlhaus et al. (1999). In contrast to that paper, we do not assume that the data generating process obeys the structure of the fitted model. Moreover, we develop the theory in a more general context.

If the goal of the analysis is the estimation of the time-varying spectrum, then one can use fully nonparametric estimates of the spectral density $f(u,\lambda)$. Based on the preperiodogram, Neumann and von Sachs (1997) developed a nonlinear wavelet estimator of the time-varying spectral density. On the other hand, there are some good reasons why a semiparametric estimate $f_{\hat{\theta}(u)}(\lambda)$ could be preferable over a fully nonparametric estimate. A successful estimate of a two-dimensional curve usually requires a considerable number of observations while a semiparametric estimate has good properties with much less observations provided that the model class describes the underlying process reasonably well. Furthermore, semiparametric models are a good tool for describing special features of the time-varying spectrum, such as the location of peaks in the spectrum over time. Another example is the time-varying version of Bloomfield's exponential model (cf. Bloomfield, 1973) which can be fitted by the methods of this paper.

Very often one is interested in time-varying models that are purely defined in the time domain, such as time-varying ARMA models. In this case the method of this paper via the spectrum may just be regarded as a technical tool for estimation.

In the present paper we intend to develop a nonparametric estimate of the parameter curve θ. There are two reasons to employ wavelet thresholding as the main smoothing step. It is well known that such estimators adapt to spatially inhomogeneous smoothness properties of a function; see, e.g., Donoho and Johnstone (1998) for minimax results in Besov classes, and Hall and Patil (1995) as well as Hall et al. (1996) who show that usual rates of convergence remain valid if the function to be estimated is smooth only in a piecewise sense. Another advantage is the simplicity of the estimation scheme: rather than aiming at the optimal compromise between variability of the estimate and bias due to smoothing; we have here an orthogonal series estimator based on empirical versions of coefficients that passed a simple significance test.

3.1. A wavelet expansion for the parameter function

Since the nonparametric estimation of μ is straightforward, we concentrate on the estimation of the parameter function $\theta(\cdot)=(\theta^{(1)}(\cdot),\ldots,\theta^{(p)}(\cdot))'$. First, we introduce an appropriate orthonormal basis of $L_2[0,1]$. Assume we have a scaling function ϕ and a so-called wavelet ψ such that $\{2^{l/2}\phi(2^l\cdot -k)\}_{k\in\mathbb{Z}}\cup\{2^{j/2}\psi(2^j\cdot -k)\}_{j\geqslant l;k\in\mathbb{Z}}$ forms an orthonormal basis of $L_2(\mathbb{R})$. The construction of such functions ϕ and ψ, which are compactly supported, is described in Daubechies (1988).

Using Daubechies' wavelets Meyer (1991) constructed an orthonormal basis of $L_2[0,1]$, essentially by truncation of the above functions to the interval $[0,1]$ and a subsequent orthonormalization step. We use throughout this paper Meyer's basis $\{\phi_{l,k}\}_{k\in I_l^0}\cup\{\psi_{j,k}\}_{j\geqslant l;k\in I_j}$, where $\#I_j=2^j$ and $\#I_l^0=2^l+N$ for some integer N depending on the regularity of the wavelet basis. The functions $\phi_{l,k}$ and $\psi_{j,k}$ are equal to $2^{l/2}\phi(2^l\cdot -k)$ and $2^{j/2}\psi(2^j\cdot -k)$, respectively, as long as the supports of the latters lie entirely in $[0,1]$. Otherwise they are derived from certain boundary-modified versions of ϕ and ψ. More exactly, there exist both N left-sided functions $\phi^{[-N]},\dots,\phi^{[-1]}$ and $\psi^{[-N]},\dots,\psi^{[-1]}$ as well as N right-sided functions $\phi^{[1]},\dots,\phi^{[N]}$ and $\psi^{[1]},\dots,\psi^{[N]}$. Appropriate translations and dilations of these functions yield the members of the Meyer basis:

$$\phi_{l,1}(u)=2^{l/2}\phi^{[-N]}(2^lu-1),\dots,\phi_{l,N}(u)=2^{l/2}\phi^{[-1]}(2^lu-N),$$

$$\phi_{l,N+1}(u)=2^{l/2}\phi(2^lu-(N+1)),\dots,\phi_{l,2^l}(u)=2^{l/2}\phi(2^lu-2^l),$$

$$\phi_{l,2^l+1}(u)=2^{l/2}\phi^{[1]}(2^lu-(2^l+1)),\dots,\phi_{l,2^l+N}(u)=2^{l/2}\phi^{[N]}(2^lu-(2^l+N)),$$

and analogously

$$\psi_{j,1}(u)=2^{j/2}\psi^{[-N]}(2^ju-1),\dots,\psi_{j,2^j}(u)=2^{j/2}\psi^{[N]}(2^ju-2^j).$$

Accordingly, we can expand the function $\theta^{(i)}$ in an orthogonal series

$$\theta^{(i)}=\sum_{k\in I_l^0}\alpha_{l,k;i}\ \phi_{l,k}+\sum_{j\geqslant l}\sum_{k\in I_j}\beta_{j,k;i}\ \psi_{j,k}, \tag{3.1}$$

where $\alpha_{l,k;i}=\int\theta^{(i)}(u)\phi_{l,k}(u)\,\mathrm{d}u$, $\beta_{j,k;i}=\int\theta^{(i)}(u)\psi_{j,k}(u)\,\mathrm{d}u$ are the usual generalized Fourier coefficients, also called wavelet coefficients in this context. Note that we could equally well use the boundary-adjusted basis of Cohen et al. (1993) rather than Meyer's basis.

The starting point in our construction is an approximation of $\theta^{(i)}$ by a truncated wavelet series

$$\theta^{(i)}\approx\sum_{k\in I_l^0}\alpha_{l,k;i}\ \phi_{l,k}+\sum_{l\leqslant j<j^*}\sum_{k\in I_j}\beta_{j,k;i}\ \psi_{j,k}, \tag{3.2}$$

where the range of appropriate values of j^* is described in Theorem 3.1 below.

The principal problem in deriving reasonable empirical coefficients is that we have no direct "access" to the $\theta^{(i)}(\cdot)$ which prevents us from finding simple empirical coefficients. For example in nonparametric regression, where we usually observe the parameter curve $\theta(\cdot)$ plus some noise, we can obtain empirical coefficients by a simple Fourier transform of the observations with respect to the wavelet basis.

A naive approach to this problem would be to estimate $\theta(u)$ by a classical (stationary) method based on the observations on some small segment around u and to apply a Fourier transform to the estimate in order to obtain the empirical wavelet coefficients.

However, as for the periodogram on small segments any such method implicitly contains some smoothing on the selected segment and will therefore be in conflict with the main smoothing step. In particular, certain features of the curve (such as jumps) are already lost by this implicit smoothing and can hardly be recovered afterwards.

Our solution out of this dilemma is to define the empirical coefficients implicitly by a minimum distance method where we use some distance $D(f_\theta, J_T^{\tilde{\mu}})$ between $f_{\theta(u)}(\lambda)$ and the preperiodogram $J_T^{\tilde{\mu}}(u,\lambda)$. The use of the preperiodogram in this distance guarantees that no implicit presmoothing is hidden in this step.

The distance we use is

$$D(f_\theta, f) = \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} \left\{ \log f_{\theta(u)}(\lambda) + \frac{f(u,\lambda)}{f_{\theta(u)}(\lambda)} \right\} \mathrm{d}\lambda\, \mathrm{d}u, \tag{3.3}$$

which is up to a constant the asymptotic Kullback–Leibler information divergence in the case of a Gaussian process (see Theorem 3.4 in Dahlhaus, 1996a). Thus, we take as the empirical distance

$$D(f_\theta, J_T^{\tilde{\mu}}) = \frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} \left\{ \log f_{\theta(u)}(\lambda) + \frac{J_T^{\tilde{\mu}}(u,\lambda)}{f_{\theta(u)}(\lambda)} \right\} \mathrm{d}\lambda\, \mathrm{d}u. \tag{3.4}$$

Dahlhaus (2000, Theorem 3.5) has proved that this is approximately the Gaussian likelihood of a locally stationary process with spectrum $f_{\theta(u)}(\lambda)$. Thus, by using this distance the empirical wavelet coefficients obtained by minimizing this distance are quasi-maximum likelihood estimates. However, other distances are possible as well, under appropriate modifications of the assumptions.

In the following we include the case of model-misspecification, that is we do not assume that the true spectral density $f(u,\lambda)$ lies in the class $\{f_\theta, \theta \in \Theta\}$. An example is the situation where we fit a time-varying AR(1)–model but our process is no AR(1) at all. In the case of model misspecification our estimate will not converge to the true parameter curve (which does not exist) but to

$$\theta_0(u) = \arg\min_{\theta\in\Theta} \int_{-\pi}^{\pi} \left\{ \log f_\theta(\lambda) + \frac{f(u,\lambda)}{f_\theta(\lambda)} \right\} \mathrm{d}\lambda, \tag{3.5}$$

which gives the best approximation with respect to the distance chosen above.

We will use the following technical assumptions which are mainly conditions on the parametrization of $\{f_\theta,\ \theta \in \Theta\}$:

(A4) (i) f_θ is four times differentiable in θ,
(ii) $\inf_{u\in[0,1]} \inf_{\theta\in\partial\Theta} \|\theta_0(u) - \theta\| \geqslant C > 0$,
(iii) $\int_{-\pi}^{\pi} [\{\log f_\theta(\lambda) + f(u,\lambda)/f_\theta(\lambda)\} - \{\log f_{\theta_0(u)}(\lambda) + f(u,\lambda)/f_{\theta_0(u)}(\lambda)\}]\,\mathrm{d}\lambda \asymp \|\theta - \theta_0(u)\|^2$, where $r_1(x) \asymp r_2(x)$ means that there are positive constants C_1, C_2 such that $C_1 r_1(x) \leqslant r_2(x) \leqslant C_2 r_1(x)$ for all arguments x,
(iv) $\nabla^2 D(f_\theta, f)$ is Lipschitz continuous in θ,
(v) $\inf_{\theta\in\Theta} \lambda_{\min}(\nabla^2 \int_{-\pi}^{\pi} \{\log f_\theta(\lambda) + f(u,\lambda)/f_\theta(\lambda)\}\,\mathrm{d}\lambda) \geqslant M > 0$.

Condition (ii) means that the best parameter curve is in the interior of Θ for all u. Condition (iii) basically says that the optimization of an empirical version of the

Kullback–Leibler distance leads to a reasonable estimate of the parameter θ_0. Condition (v) will be used to derive a certain rate for the empirical estimate of θ_0, see e.g. (B.12) below. Finally, (i) and (iv) are some technical regularity conditions.

The empirical distance (3.4) will be used to get preliminary estimates of the unknown coefficients in expansion (3.2). Note that the right-hand side of (3.2) can be rewritten as

$$\sum_{k\in I_l^0} \alpha_{l,k;i}\phi_{l,k}(u) + \sum_{l\leqslant j<j^*}\sum_{k\in I_j} \beta_{j,k;i}\;\psi_{j,k}(u) = \sum_{k\in I_{j^*}^0} \alpha_{j^*,k;i}\,\phi_{j^*,k}(u). \tag{3.6}$$

The latter representation allows a more convenient derivation of the stochastic properties of the empirical coefficients since the calculations are then on a single scale (see, e.g., parts (i) and (ii) of the proof of Theorem 3.1). Moreover, on the basis of this representation, we can now replace a high-dimensional optimization problem by a certain number of separate lower-dimensional optimization problems; see the discussion below. Let $\alpha = ((\alpha_{j^*,1})',\dots,(\alpha_{j^*,2^{j^*}+N})')'$, where $\alpha_{j^*,k} = (\alpha_{j^*,k;1},\dots,\alpha_{j^*,k;p})'$. Define

$$f_\alpha(u,\lambda) = f_{\sum_k \alpha_{j^*,k}\phi_{j^*,k}(u)}(\lambda).$$

Now one could obtain an empirical version of α by minimization of $D(f_\alpha, J_T^{\tilde\mu})$. However, this would lead to a optimization problem in $(2^{j^*}+N)\cdot p$ variables, which can turn out to be very time consuming, even for moderate j^*. (Notice that we will assume that 2^{j^*} grows at some rate T^ρ, for some $\rho>0$, as $T\to\infty$.) To end up with a practicable method, we define empirical wavelet coefficients as the solution to a number of separate, low-dimensional optimization problems.

Suppose we want to define an appropriate empirical version of $\alpha_{j^*,k}$, where $\phi_{j^*,k}$ is not one of the boundary-corrected wavelets. Then we take, besides $\phi_{j^*,k}^{[0]} = \phi_{j^*,k}$, N left-sided boundary functions, $\phi_{j^*,k-N}^{[-N]},\dots,\phi_{j^*,k-1}^{[-1]}$, and N right-sided boundary functions, $\phi_{j^*,k+1}^{[1]},\dots,\phi_{j^*,k+N}^{[N]}$. Let $\Delta_k = \bigcup_{l=-N}^{N} \operatorname{supp}(\phi_{j^*,k+l}^{[l]})$. According to the construction of Meyer (1991), $\{\phi_{j^*,k+l}^{[l]}\}_{l=-N,\dots,N}$ is an orthonormal system on Δ_k.

Now, we set

$$\tilde\alpha_k = ((\tilde\alpha_{j^*,k-N}^{[-N]})',\dots,(\tilde\alpha_{j^*,k+N}^{[N]})')' = \arg\inf_\alpha D_T^{(k)}(\alpha), \quad k = N+1,\dots,2^{j^*}, \tag{3.7}$$

where

$$D_T^{(k)}(\alpha) = \frac{1}{4\pi}\int_{\Delta_k}\int_{-\pi}^{\pi}\left\{\log f_{\sum_l \alpha_l\phi_{j^*,k+l}^{[l]}(u)}(\lambda) + \frac{J_T^{\tilde\mu}(u,\lambda)}{f_{\sum_l \alpha_l\phi_{j^*,k+l}^{[l]}(u)}(\lambda)}\right\}\mathrm{d}\lambda\,\mathrm{d}u \tag{3.8}$$

and define

$$\tilde\alpha_{j^*,k} = \tilde\alpha_{j^*,k}^{[0]}. \tag{3.9}$$

Empirical coefficients corresponding to the left-sided boundary functions, $\phi_{j^*,l}^{[-N]},\dots,\phi_{j^*,N}^{[-1]}$, are obtained as $\tilde\alpha_{j^*,1}^{[-N]},\dots,\tilde\alpha_{j^*,N}^{[-1]}$, whereas their right-sided counterparts are taken as $\tilde\alpha_{j^*,2^{j^*}+1}^{[1]},\dots,\tilde\alpha_{j^*,2^{j^*}+N}^{[N]}$. The original problem of minimizing $D(f_{\sum \alpha_{j^*,k}\phi_{j^*,k}}, J_T^{\tilde\mu})$ w.r.t.

$\alpha_{j^*,1},\ldots,\alpha_{j^*,2^{j^*}+N}$ is replaced by a collection of independent lower-dimensional optimization problems (3.7). In total we have to solve $2^{j^*}-N$ independent optimization problems in $(2N+1)\cdot p$ variables.

Notice that $\tilde{\alpha}_k$ primarily estimates $\alpha_{k;\inf}$, where

$$\alpha_{k;\inf}=((\alpha^{[-N]}_{j^*,k-N;\inf})',\ldots,(\alpha^{[N]}_{j^*,k+N;\inf})')'=\arg\inf_{\alpha} D^{(k)}(\alpha) \tag{3.10}$$

and

$$D^{(k)}(\alpha)=\frac{1}{4\pi}\int_{\Delta_k}\int_{-\pi}^{\pi}\left\{\log f_{\sum_l \alpha_l\phi^{[l]}_{j^*,k+l}(u)}(\lambda)+\frac{f(u,\lambda)}{f_{\sum_l \alpha_l\phi^{[l]}_{j^*,k+l}(u)}(\lambda)}\right\}\mathrm{d}\lambda\,\mathrm{d}u. \tag{3.11}$$

To demonstrate the usefulness of our proposed method, we intend to show that our nonlinear wavelet estimator attains rates of convergence that are usually not obtained by conventional linear smoothing methods. Accordingly, we consider Besov classes as function classes which may contain functions with high spatial inhomogeneity in their smoothness properties. Furthermore, Besov spaces represent the most convenient scale of function spaces in the context of wavelet methods, since the corresponding norm is equivalent to a certain norm in the sequence space of coefficients of a sufficiently regular wavelet basis. For an introduction to the theory of Besov spaces $B^m_{p,q}$ see, e.g., Triebel (1990). Here $m\geqslant 1$ denotes the degree of smoothness and p,q $(1\leqslant p,q\leqslant\infty)$ specify the norm in which smoothness is measured. These classes contain traditional Hölder and L_2-Sobolev smoothness classes, by setting $p=q=\infty$ and 2, respectively. Moreover, they embed other interesting spaces like Sobolev spaces W^m_p, for which the inclusions $B^m_{p,p}\subseteq W^m_p\subseteq B^m_{p,2}$ in the case $1<p\leqslant 2$, and $B^m_{p,2}\subseteq W^m_p\subseteq B^m_{p,p}$ if $2\leqslant p<\infty$ hold true; see, e.g., Theorem 6.4.4 in Bergh and Löfström (1976).

According to given smoothness classes $\mathscr{F}_i=\mathscr{F}_i(m_i,p_i,q_i,C_1,C_2)$, for $\theta_0^{(i)}$, we have to choose a wavelet basis that is actually able to exploit the underlying smoothness. In accordance with this, we choose compactly supported wavelet functions of regularity $r>\max\{m_1,\ldots,m_p\}$, that is

(A5) (i) ϕ and ψ are $C^r[0,1]$ and have compact support,
(ii) $\int\phi(t)\,\mathrm{d}t=1$, $\int\psi(t)t^k\,\mathrm{d}t=0$ for $0\leqslant k\leqslant r$.

For convenience, we define our function class by constraints on the sequences of wavelet coefficients. Fix any positive constants C_1, C_2. We will assume that $\theta_0^{(i)}$ lies in the following set of functions:

$$\mathscr{F}_i=\left\{f=\sum_k \alpha_{l,k}\phi_{l,k}+\sum_{j,k}\beta_{j,k}\psi_{j,k}\,\middle|\,\|\alpha_{l\cdot}\|_\infty\leqslant C_1,\|\beta_{\cdot\cdot}\|_{m_i,p_i,q_i}\leqslant C_2\right\},$$

where

$$\|\beta_{\cdot\cdot}\|_{m,p,q}=\left(\sum_{j\geqslant l}\left[2^{jsp}\sum_{k\in I_j}|\beta_{jk}|^p\right]^{q/p}\right)^{1/q},$$

$s=m+1/2-1/p$. For the parameters defining the class $\mathscr{F}_i$, we assume that $m_i\geqslant 1$, $1\leqslant p_i,q_i\leqslant\infty$, and $m_i>1/p_i$. The latter condition implies that each function in $\mathscr{F}_i$

is continuous. It is well known that the class $\mathscr{F}_i$ lies between the classes $B^{m_i}_{p_i,q_i}(c)$ and $B^{m_i}_{p_i,q_i}(C)$, for appropriate constants c and C; cf. Donoho and Johnstone (1998, Theorem 2).

The coefficients $\alpha^{[0]}_{j^*,k;\mathrm{inf}}$ defined above may be different from the $\alpha_{j^*,k}$, but the following lemma asserts that this difference is asymptotically negligible in smoothness classes we intend to consider.

Lemma 3.1. *Suppose that* (A4) *and* (A5) *are fulfilled and that* $\theta^{(i)}_0 \in \mathscr{F}_i$ *for all* $i = 1,\dots,p$. *Then*

$$\sum_{k=N+1}^{2^{j^*}} \|\alpha^{[0]}_{j^*,k,\mathrm{inf}} - \alpha_{j^*,k}\|^2 + \sum_{k=1}^{N} \|\alpha^{[-k]}_{j^*,N+1-k;\mathrm{inf}} - \alpha^{[-k]}_{j^*,N+1-k}\|^2$$
$$+\|\alpha^{[k]}_{j^*,2^{j^*}+k;\mathrm{inf}} - \alpha^{[k]}_{j^*,2^{j^*}+k}\|^2 = \mathrm{O}(2^{-2j^*s}),$$

where $s = \min\{s_1,\dots,s_p\}$ *and* $s_i = m_i + 1/2 - \max\{1/2, 1/p_i\}$.

The difference between $\alpha^{[0]}_{j^*,k,\mathrm{inf}}$ and $\alpha_{j^*,k}$ is indeed negligible, because an error of $\mathrm{O}(2^{-2j^*s})$ is incurred in any case by the truncation of the wavelet expansion of $\theta_0(u)$ at the scale j^*.

It will be shown in the proof of Theorem 3.1 that, with a probability exceeding $1-\mathrm{O}(T^{-\lambda})$, $\tilde{\alpha}_k$ and $\alpha_{k;\mathrm{inf}}$ are interior points of the set of admissible values. Hence,

$$\nabla D^{(k)}_T(\tilde{\alpha}_k) = \nabla D^{(k)}(\alpha_{k;\mathrm{inf}}) = 0.$$

This yields that

$$0 = \nabla D^{(k)}_T(\tilde{\alpha}_k) - \nabla D^{(k)}_T(\alpha_{k;\mathrm{inf}}) + \nabla D^{(k)}_T(\alpha_{k;\mathrm{inf}}) - \nabla D^{(k)}(\alpha_{k;\mathrm{inf}})$$
$$= \nabla^2 D^{(k)}(\alpha_{k;\mathrm{inf}})(\tilde{\alpha}_k - \alpha_{k;\mathrm{inf}}) + \nabla D^{(k)}_T(\alpha_{k;\mathrm{inf}}) - \nabla D^{(k)}(\alpha_{k;\mathrm{inf}}) + R_{k,T},$$

where $R_{k,T} = \nabla D^{(k)}_T(\tilde{\alpha}_k) - \nabla D^{(k)}_T(\alpha_{k;\mathrm{inf}}) - \nabla^2 D^{(k)}(\alpha_{k;\mathrm{inf}})(\tilde{\alpha}_k - (\alpha_{k;\mathrm{inf}})$. It will be shown further that $R_{k,T}$ can be asymptotically neglected, which leads to the following explicit approximation of $\tilde{\alpha}_k$:

$$\tilde{\alpha}_k \approx \alpha_{k;\mathrm{inf}} - [\nabla^2 D^{(k)}(\alpha_{k;\mathrm{inf}})]^{-1}(\nabla D^{(k)}_T(\alpha_{k;\mathrm{inf}}) - \nabla D^{(k)}(\alpha_{k;\mathrm{inf}})). \tag{3.12}$$

This means that $\tilde{\alpha}_{j^*,k}$ can be approximated to first order by a weighted integral of the preperiodogram.

Note that both $\{\phi_{l,1},\dots,\phi_{l,2^l+N},\psi_{l,1},\dots,\psi_{l,2^l},\dots,\psi_{j^*-1,1},\dots,\psi^{2^{j^*-1}}_{j^*-1,2^{j^*}-1}\}$ and $\{\phi_{j^*,1},\dots,\phi_{j^*,2^{j^*}+N}\}$ are orthonormal bases of the same space V_{j^*}. Hence, there exists an orthonormal $((2^{j^*}+N)\times(2^{j^*}+N))$-matrix Γ with

$$(\phi_{l,1},\dots,\phi_{l,2^l+N},\psi_{l,1},\dots,\psi_{l,2^l},\dots,\psi_{j^*-1,1},\dots,\psi_{j^*-1,2^{j^*}-1})' = \Gamma(\phi_{j^*,1},\dots,\phi_{j^*,2^{j^*}+N})'.$$

This implies

$$(\alpha_{j^*,1;i},\ldots,\alpha_{j^*,2^{j^*}+N;i})\begin{pmatrix}\phi_{j^*,1}\\ \vdots\\ \phi_{j^*,2^{j^*}+N}\end{pmatrix} = (\alpha_{j^*,1;i},\ldots,\alpha_{j^*,2^{j^*}+N;i})\Gamma'\begin{pmatrix}\phi_{l,1}\\ \vdots\\ \phi_{l,2^l+N}\\ \psi_{l,1}\\ \vdots\\ \psi_{j^*-1,2^{j^*}-1}\end{pmatrix}.$$

Hence, having any reasonable estimate of $(\alpha_{j^*,1;i},\ldots,\alpha_{j^*,2^{j^*}+N;i})$, we can readily define a reasonable estimate of the corresponding coefficients in the other basis. We define

$$\tilde{\beta}_{j,k;i} = \Gamma_{j,k}(\tilde{\alpha}_{j^*,1;i},\ldots,\tilde{\alpha}_{j^*,2^{j^*}+N;i})', \tag{3.13}$$

where $\Gamma_{j,k}$ is the appropriate row of the matrix Γ'. ($\tilde{\alpha}_{l,1;i},\ldots,\tilde{\alpha}_{l,2^l+N;i}$ are defined analogously.)

According to (3.12), the coefficients $\tilde{\beta}_{j,k;i}$ can be again approximated by a weighted integral over the preperiodogram. By Lemma A.3 one can show asymptotic normality of $\sqrt{T}(\tilde{\beta}_{j,k;i} - \beta_{j,k;i})$. Moreover, this asymptotic normality can be expressed in terms of probabilities of large deviations. However, an explicit expression of the asymptotic variance of $\tilde{\beta}_{j,k;i}$ is presumably quite involved. Hence, we do not try to specify it further. It is only important to know that the tails of $\sqrt{T}(\tilde{\beta}_{j,k;i} - \beta_{j,k;i})$ can be approximated by tails of a certain normal distribution. This is formalized in the following theorem.

Theorem 3.1. *Suppose that* (A1)–(A5) *are fulfilled and that* $\theta_0^{(i)} \in \mathscr{F}_i$ *for all* $i = 1,\ldots,p$. *Further, assume that* $[(Tb)^{-1}+b^{2r}]=\mathrm{O}(T^{-1/2})$. *We choose the cut-off point* j^* *such that* $2^{j^*} = \mathrm{O}(T^{1-\delta})$ *and* $2^{-j^*} = \mathrm{O}(T^{-2/3})$. *Then there exists a universal constant* $\kappa < \infty$ *such that*

$$P(\pm(\tilde{\beta}_{j,k;i} - \beta_{j,k;i}) \geqslant x\kappa/\sqrt{T}) \leqslant 2(1-\Phi(x))(1+\mathrm{o}(1)) + \mathrm{O}(T^{-\lambda})$$

holds uniformly in x *and* $(j,k) \in \mathscr{T}_T$.

Even if we do not explicitly know the constant κ, this result will prove to be a reasonable starting point for devising a locally adaptive smoothing strategy by nonlinear thresholding.

4. Locally adaptive estimation by wavelet thresholding

In this section we turn to the major regularization step of our method. Whereas most of the commonly used smoothers (kernel, spline) modify noisy data in a linear manner, we intend to apply nonlinear thresholding to the empirical wavelet coefficients. It is well known that traditional linear estimators are able to achieve optimal rates of convergence in settings where these issues are usually studied, i.e., as long as the underlying smoothness of the curve to be estimated is not too inhomogeneous.

However, they are not able to achieve optimal rates in cases where the degree of smoothness varies strongly over the domain; see, for example, Donoho and Johnstone (1998), Hall and Patil (1995), and Hall et al. (1996). To achieve optimality in the latter case, one has to apply different degrees of smoothing at different locations. This, of course, coincides with the natural idea of using kernel estimators with locally varying bandwidths in cases of functions with inhomogeneous smoothness properties.

To explain the need for nonlinear smoothing schemes on a more technical level, assume that empirical coefficients $\tilde{\beta}_{j,k}$ are given which are exactly normally distributed, that is

$$\tilde{\beta}_{j,k} \sim \mathcal{N}(\beta_{j,k}, \sigma_T^2), \quad (j,k) \in \mathcal{J}_T. \tag{4.1}$$

For a linear estimator, $c\tilde{\beta}_{j,k}$, it is easy to see that

$$\sum_k \mathbb{E}(c\tilde{\beta}_{j,k} - \beta_{j,k})^2 \geqslant \frac{1}{2} \min\left\{ \sum_k \beta_{j,k}^2, \sum_k \sigma_T^2 \right\}. \tag{4.2}$$

In contrast, for nonlinear estimators $\delta^{(.)}(\tilde{\beta}_{j,k}, \lambda_T)$ with $\lambda_T = \sigma_T\sqrt{2\log \#\mathcal{J}_T}$ introduced below, it can be shown that

$$\sum_k \mathbb{E}(\delta^{(.)}(\tilde{\beta}_{j,k}, \lambda_T) - \beta_{j,k})^2 \leqslant C \sum_k \min\{\beta_{j,k}^2, \lambda_T^2\} + O(T^{-1}). \tag{4.3}$$

If now the majority of the coefficients within the scale j are of smaller order of magnitude than σ_T while a few of them are pretty large, then it may well happen that

$$\inf_c \left\{ \sum_k \mathbb{E}(c\tilde{\beta}_{j,k} - \beta_{j,k})^2 \right\} \gg \sum_k \mathbb{E}(\delta^{(.)}(\tilde{\beta}_{j,k}, \lambda_T) - \beta_{j,k})^2. \tag{4.4}$$

This is just the case for certain functions from Besov classes $B_{p,q}^m(C)$ if $p < 2$; see Donoho and Johnstone (1998). Another, even more obvious case are piecewise smooth functions with jumps between the smooth parts. This case was studied by Hall and Patil (1995) and Hall et al. (1996). They showed that the coefficients assigned to basis functions supported on one of the smooth parts decay at the rate $2^{-j(m+1/2)}$, where m is the degree of smoothness. In contrast, they decay at the much slower rate $2^{-j/2}$ around the jumps. This is a typical scenario leading to (4.4). The same effect, although in a less drastic form, occurs with certain functions from the Besov scale.

Two frequently used rules to treat coefficients obeying (4.1) are

(1) hard thresholding

$$\delta^{(h)}(\tilde{\beta}_{j,k}, \lambda) = \tilde{\beta}_{j,k} I(|\tilde{\beta}_{j,k}| \geqslant \lambda)$$

and

(2) soft thresholding

$$\delta^{(s)}(\tilde{\beta}_{j,k}, \lambda) = (|\tilde{\beta}_{j,k}| - \lambda)_+ \operatorname{sgn}(\tilde{\beta}_{j,k}).$$

To simplify notation, we will use the symbol $\delta^{(.)}$ to denote either $\delta^{(h)}$ or $\delta^{(s)}$.

An exceptionally simple all-purpose rule for the tuning of this function propagated in a series of papers by Donoho and Johnstone is given as

$$\lambda_T = \sigma_T \sqrt{2\log(\#\mathscr{J}_T)}, \tag{4.5}$$

where $\{\tilde{\beta}_{j,k}, (j,k) \in \mathscr{J}_T\}$ is the set of empirical coefficients to be thresholded. Although this rule is slightly suboptimal w.r.t. the rate of convergence for the L_2-risk of the corresponding estimator, it has a number of interesting properties; see Donoho and Johnstone (1994).

Finally, one composes an estimator of $g(u) = \sum_{k\in I_l^0} \alpha_{l,k}\phi_{l,k}(u) + \sum_{j\geqslant l}\sum_{k\in I_j} \beta_{j,k} \psi_{j,k}(u)$ from the nonlinearly modified empirical coefficients as

$$\hat{g}(u) = \sum_{k\in I_l^0} \tilde{\alpha}_{l,k}\phi_{l,k}(u) + \sum_{(j,k)\in\mathscr{J}_T} \delta^{(\cdot)}(\tilde{\beta}_{j,k}, \lambda_T)\psi_{j,k}(u). \tag{4.6}$$

Besides some other properties that partly compensate for the slight suboptimality of wavelet estimators based on logarithmic thresholds as in (4.5), one major advantage of this method is its universality. The above scheme is neither restricted to specific models nor to specific smoothness classes. Actually, only some appropriate type of asymptotic normality for the empirical coefficients is necessary for the successful application of this method. That means in particular, that we neither need any specific structural assumptions on the data generating process nor on the joint distribution of the empirical wavelet coefficients.

Now, it is straightforward to transfer this nonlinear thresholding scheme to our particular context of estimating the parameter functions $\theta_0^{(i)}$.

Let $\mathscr{J}_T = \{(j,k) \mid k \in I_j; j \geqslant l,\ 2^j \leqslant T^{1-\gamma}\}$, for some $0 < \gamma \leqslant 1/3$, be the set of coefficients to be thresholded. Because of the exponentially decaying tails of the distribution of $\sqrt{T}(\tilde{\beta}_{j,k;i} - \beta_{j,k;i})$ stated in Theorem 3.1, it will turn out to be sufficient for purposes of "denoising" to apply the thresholds

$$\lambda_T = KT^{-1/2}\sqrt{\log T}, \tag{4.7}$$

where K is some sufficiently large constant. According to the above discussion, we define

$$\hat{\beta}_{j,k;i} = \delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T). \tag{4.8}$$

This leads to the estimator

$$\hat{\theta}^{(i)}(u) = \sum_{k\in I_l^0} \tilde{\alpha}_{l,k;i}\phi_{l,k}(u) + \sum_{(j,k)\in\mathscr{J}_T} \hat{\beta}_{j,k;i}\psi_{j,k}(u). \tag{4.9}$$

Now, we can state a theorem that characterizes the convergence properties of $\hat{\theta}^{(i)}$ in $\mathscr{F}_i$.

Theorem 4.1. *Suppose that* (A1)–(A5) *are fulfilled and that* $\theta_0^{(i)} \in \mathscr{F}_i$ *holds for all* $i = 1, \ldots, p$. *Then*

$$\mathbb{E}\|\hat{\theta}^{(i)} - \theta_0^{(i)}\|^2 = \mathrm{O}((\log T/T)^{2m_i/(2m_i+1)} + 2^{-2j^* s}).$$

According to (A4)(iii), this theorem has an immediate implication for the risk in estimating the best projection f_{θ_0}, measured in the Kullback–Leibler distance.

Corollary 4.1. *Suppose that* (A1)–(A5) *are fulfilled and that* $\theta_0^{(i)} \in \mathscr{F}_i$ *holds for all* $i = 1, \ldots, p$. *Then*

$$\mathbb{E}\int_0^1 \int_{-\pi}^{\pi} \left[\left\{ \log f_{\hat{\theta}(u)}(\lambda) + \frac{f(u,\lambda)}{f_{\hat{\theta}(u)}(\lambda)} \right\} - \left\{ \log f_{\theta_0(u)}(\lambda) + \frac{f(u,\lambda)}{f_{\theta_0(u)}(\lambda)} \right\} \right] \mathrm{d}\lambda\, \mathrm{d}u$$
$$= \mathrm{O}((\log T/T)^{2m/(2m+1)}),$$

where $m = \min\{m_1, \ldots, m_p\}$.

Theorem 4.1 and Corollary 4.1 basically say that the proposed estimators converge to the corresponding target quantities θ_0 and f_{θ_0} with a rate that matches the optimal rate in smoothness classes $\mathscr{F}_i$ in cases where such a minimax rate is known (regression, density estimation, etc.). This does not require that the fitted model is adequate. We decided to formulate these results under mixing conditions rather than under a more restrictive condition such as (2.7) on the cumulant sums. We did not include simulations. For the special case of fitting an AR(p)-process with time varying coefficients, some promising simulation results on a wavelet estimator are contained in Dahlhaus et al. (1999).

Appendix A Asymptotic normality of quadratic forms

As often in spectral analysis, theoretical results are based in large parts on an analysis of certain quadratic forms. In order to preserve a clear structure for the rest of the paper, we collect some technical results on quadratic forms in the subsection. We think that these results might also be of independent interest, and therefore we conclude this subsection with simple examples that classify Toeplitz matrices with respect to the asymptotic distributions of corresponding quadratic forms.

The derivation of the asymptotic normality is essentially based on upper estimates of the cumulants. Such estimates, which slightly generalize a result by Rudzkis (1978), are derived in Neumann (1996) under the assumptions $\mathbb{E}X_t = 0$ and

$$\sup_{t_1} \left\{ \sum_{t_2,\ldots,t_k=1}^{T} |\mathrm{cum}(X_{t_1}, \ldots, X_{t_k})| \right\} \leqslant C^k (k!)^{1+\gamma} \quad \text{for all } k = 2, 3, \ldots \tag{A.1}$$

and appropriate $C < \infty$, $\gamma \geqslant 0$. Even if (A.1) can be shown to hold under appropriate mixing conditions for some textbook distributions (see Neumann, 1996, Remark 3.1), it is somehow annoying to have such a *quantitative* restriction to hold simultaneously for all $k \geqslant 2$. In particular, for (A.1) to hold we have to assume that there exists constants $C < \infty$, $\gamma < \infty$ such that $E|X_t|^k \leqslant C^k (k!)^{\gamma}$ is satisfied for all k. Instead of (A.1), here we impose condition (A2) on the process $\{X_t\}$. Now, we have a *qualitative* restriction on the moments of the X_t's, that is instead of the explicit bounds for them we assume only their finiteness. The bridge to a cumulant estimate like (A.1), which is needed to apply Lemma 1 of Rudzkis et al. (1978) for proving the asymptotic normality of the quadratic form, is obtained via a simple truncation argument, that is the X_t will be replaced by certain truncated and recentered random variables $\tilde{X}_t$. Under (A2), we obtain a domain of attraction to the normal law of the form $[-\sqrt{C\log(T)}, \sqrt{C\log(T)}]$, which corresponds to "remaining tail probabilities" of order $T^{-C/2}$. In contrast, we obtained in Neumann (1996) a domain of attraction of the form $[-T^{\rho}, T^{\rho}]$ for some $\rho > 0$, which led to exponentially decaying tail probabilities.

We remind the reader that we use the convention that $\lambda < \infty$ denotes an arbitrarily large and $\delta > 0$ an arbitrarily small constant.

Lemma A.1. *Suppose that* $\mathbb{E}X_t = 0$ *and that* (A2) *is fulfilled. Further, let* $\delta, \lambda > 0$ *be arbitrary. Then there exists random variables* $\tilde{X}_t$ *with* $\mathbb{E}\tilde{X}_t = 0$,

$$P(X_t \neq \tilde{X}_t) = O(T^{-\lambda}),$$
$$\mathbb{E}\tilde{X}_{t_1}\cdots\tilde{X}_{t_k} = \mathbb{E}X_{t_1}\cdots X_{t_k} + O(T^{-\lambda})$$

and

$$\sup_{t_1}\left\{\sum_{t_2,\ldots,t_k=1}^{T} |\mathrm{cum}(\tilde{X}_{t_1},\ldots,\tilde{X}_{t_k})|\right\} \leqslant C^k \quad \textit{for all } k = 2,3,\ldots .$$

Moreover, there exists a unique constant C such that

$$\sup_{t_1}\left\{\sum_{t_2,\ldots,t_k=1}^{T} |\mathrm{cum}(\tilde{X}_{t_1},\ldots,\tilde{X}_{t_k})|\right\} \leqslant C^k T^{\delta k}(k!)^3 \quad \textit{for all } k = 2,3,\ldots .$$

Proof. (i) *Construction of* $\{\tilde{X}_t\}$: We define appropriate $\tilde{X}_t$'s by truncation at $T^{\delta}/2$ and recentering of these random variables.

Let

$$X_t^* = \begin{cases} X_t & \text{if } |X_t| \leqslant T^{\delta}/2, \\ 0 & \text{otherwise.} \end{cases}$$

From Markov's inequality we have that

$$P(X_t \neq X_t^*) = P(|X_t| > T^{\delta}/2) = O(T^{-\lambda}\mathbb{E}|X_t|^{\lambda/\delta}) = O(T^{-\lambda})$$

holds for arbitrary $\lambda < \infty$. Furthermore, we have

$$|\mathbb{E}X_t^*| = |\mathbb{E}X_t I(|X_t| > T^{\delta}/2)|$$
$$\leqslant \sqrt{\mathbb{E}X_t^2}\sqrt{P(|X_t| > T^{\delta}/2)} = O(T^{-\lambda}).$$

We define, for some set Ω_t with $P(\Omega_t) = 2\mathbb{E}X_t^*/T^{\delta}$,

$$\tilde{X}_t = \begin{cases} X_t^* - T^{\delta}/2 & \text{if } X_t^* \leqslant \Omega_t, \\ X_t^* & \text{otherwise.} \end{cases}$$

Then $\mathbb{E}\tilde{X}_t = 0$, $|\tilde{X}_t| \leqslant T^{\delta}$ a.s., and

$$P(X_t \neq \tilde{X}_t) = O(T^{-\lambda}). \tag{A.2}$$

Since $\tilde{X}_t$ is a function of X_t, the mixing property (A2)(ii) remains true for the sequence $\{\tilde{X}_t\}$ as well.

Further, we have that

$$\begin{aligned}
&\mathbb{E}\tilde{X}_{t_1}\cdots\tilde{X}_{t_k} - \mathbb{E}X_{t_1}\cdots X_{t_k} \\
&\quad = \sum_{i=1}^{k} \mathbb{E}X_{t_1}\cdots X_{t_{i-1}}(\tilde{X}_{t_i} - X_{t_i})\tilde{X}_{t_{i+1}}\cdots\tilde{X}_{t_k} \\
&\quad \leqslant \sum_{i=1}^{k} \sqrt{P(\tilde{X}_{t_i} \neq X_{t_i})}\sqrt{\mathbb{E}X_{t_1}^2\cdots X_{t_{i-1}}^2(\tilde{X}_{t_i} - X_{t_i})^2\tilde{X}_{t_{i+1}}^2\cdots\tilde{X}_{t_k}^2} \\
&\quad = \mathrm{O}(T^{-\lambda}).
\end{aligned}$$

(ii) *Estimation of the cumulant sums*: Using (1)(a) of Theorem 3 in Statulevicius and Jakimavicius (1988) we have, for $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_k$, that

$$\begin{aligned}
|\mathrm{cum}(\tilde{X}_{t_1},\ldots,\tilde{X}_{t_k})| &\leqslant \min_i \left\{(k-1)!2^k \sup_t\{\|\tilde{X}_t\|_\infty^k\}\alpha(|t_i - t_{i+1}|)\right\} \\
&\leqslant (k-1)!2^k T^{\delta k} \prod_{i=1}^{k-1} \alpha^{1/(k-1)}(|t_i - t_{i+1}|).
\end{aligned}$$

Now, we obtain

$$\begin{aligned}
&\sum_{t_2,\ldots,t_k} |\mathrm{cum}(\tilde{X}_{t_1},\ldots,\tilde{X}_{t_k})| \\
&\quad = \sum_{s_2,\ldots,s_k} |\mathrm{cum}(\tilde{X}_{t_1},\tilde{X}_{t_1+s_2},\ldots,\tilde{X}_{t_1+s_k})| \\
&\quad \leqslant k!(k-1)!2^k T^{\delta k}\left[\sum_{s_2}\alpha^{1/(k-1)}(|s_2|)\right]\cdot\ldots\cdot\left[\sum_{s_k}\alpha^{1/(k-1)}(|s_k|)\right] \\
&\quad \leqslant C^k(k!)^2 T^{\delta k}(k-1)^{k-1} \\
&\quad \leqslant C^k T^{\delta k}(k!)^3.
\end{aligned}$$

The last inequality follows from the relation $k^k = k(1+1/(k-1))^{k-1}(k-1)^{k-1} \leqslant ke(k-1)^{k-1} \leqslant \cdots \leqslant k!e^k$. $\square$

In the following, we consider the stochastic behavior of quadratic forms $\underline{X}'M\underline{X}$, where $\underline{X} = (X_1,\ldots,X_T)'$ and M is a symmetric matrix. According to Lemma A.1, we can replace this quadratic form by $\underline{\tilde{X}}'M\underline{\tilde{X}}$, where $\underline{\tilde{X}} = (\tilde{X}_1,\ldots,\tilde{X}_T)'$, which will be considered first.

Lemma A.2. *Suppose that* $\mathbb{E}X_t = 0$ *and* (A2) *are fulfilled. Further, let* M *be a symmetric matrix. Then there exist random variables* $\tilde{X}_t$ *with*

$$P(X_t \neq \tilde{X}_t) = \mathrm{O}(T^{-\lambda})$$

and

$$|\mathrm{cum}_k(\underline{\tilde{X}}'M\underline{\tilde{X}})| \leqslant (k-1)!2^{k-1}[\mathrm{tr}([M\,\mathrm{Cov}(\underline{\tilde{X}})]^2)]^{k/2} + R_k,$$

where

$$R_k \leqslant C^k T^{\delta k}((2k)!)^3 \max_{s,t}\{|M_{s,t}|\}\tilde{M}\|M\|_\infty^{k-2},$$

$$\tilde{M} = \sum_s \max_t\{|M_{s,t}|\}, \quad \|M\|_\infty = \max_s \left\{\sum_t |M_{s,t}|\right\}.$$

Proof. Follows from the previous lemma and Lemma 3.1 in Neumann (1996). □

Lemma A.3. *Suppose that* $\mathbb{E}X_t = 0$ *and* (A2) *are fulfilled. Further, let* M *be a symmetric matrix. Then*

$$\sigma_T^2 = \operatorname{var}(\underline{X}'M\underline{X}) = 2\operatorname{tr}(M\operatorname{Cov}(\underline{X})M\operatorname{Cov}(\underline{X})) + \sum_{t_1,t_2,t_3,t_4} M_{t_1,t_2}M_{t_3,t_4}\operatorname{cum}(X_{t_1},X_{t_2},X_{t_3},X_{t_4}),$$

where

$$\operatorname{cum}(X_{t_1},X_{t_2},X_{t_3},X_{t_4}) = \mathbb{E}X_{t_1}X_{t_2}X_{t_3}X_{t_4} - \mathbb{E}X_{t_1}X_{t_2}\mathbb{E}X_{t_3}X_{t_4} - \mathbb{E}X_{t_1}X_{t_3}\mathbb{E}X_{t_2}X_{t_4} - \mathbb{E}X_{t_1}X_{t_4}\mathbb{E}X_{t_2}X_{t_3}$$

is the fourth-order cumulant.

Moreover, we assume that

$$\|M\|_\infty = O(T^{-1/2-\delta})$$

and

$$\max_{s,t}\{|M_{st}|\}\tilde{M}\|M\|_\infty = O(T^{-3/2-\delta}):$$

(i) *If* $\sigma_T \geqslant C_0 T^{-1/2}$ *for some* $C_0 > 0$, *then*

$$P\left(\pm\frac{\underline{X}'M\underline{X} - \mathbb{E}\underline{X}'M\underline{X}}{\sigma_T} \geqslant x\right) = (1-\Phi(x))(1+o(1)) + O(T^{-\lambda})$$

holds uniformly in $x \in \mathbb{R}$.

(ii) *If* $\sigma_T = O(T^{-1/2})$, *then, for* $\bar{\sigma}_T = \max\{\sigma_T, C_0/\sqrt{T}\}$ *and arbitrary* $C_0 > 0$,

$$P\left(\pm\frac{\underline{X}'M\underline{X} - \mathbb{E}\underline{X}'M\underline{X}}{\bar{\sigma}_T} \geqslant x\right) \leqslant 2(1-\Phi(x))(1+o(1)) + O(T^{-\lambda})$$

holds uniformly in $x \in \mathbb{R}$.

Proof. According to Lemma A.1, we have with a probability exceeding $1 - O(T^{-\lambda})$ that

$$\underline{X}'M\underline{X} - \mathbb{E}\underline{X}'M\underline{X} = \underline{\tilde{X}}'M\underline{\tilde{X}} - \mathbb{E}\underline{\tilde{X}}'M\underline{\tilde{X}} + \tilde{O}(T^{-\lambda-1/2}), \tag{A.3}$$

which allows to consider the $\tilde{X}_t$'s instead of the X_t's. By Lemma A.2 we have, for $k \geqslant 2$, that

$$|\operatorname{cum}_k(\underline{\tilde{X}}'M\underline{\tilde{X}})| \leqslant k!C^k[\operatorname{tr}(M^2)]^{k/2} + (k!)^6 C^k \max_{s,t}\{|M_{s,t}|\}\tilde{M}(T^\delta\|M\|_\infty)^{k-2}, \tag{A.4}$$

which implies the first assertion by Lemma 1 of Rudzkis et al. (1978).

The second assertion is based on the fact that the quantity $\tilde{\underline{X}}'M\tilde{\underline{X}} + \xi_T$, where $\xi_T \sim \mathrm{N}(0, \bar{\sigma}_T^2 - \sigma_T^2)$ is independent of $\tilde{\underline{X}}'M\tilde{\underline{X}}$, has variance $\bar{\sigma}_T^2$ and the same cumulants of order $k = 3, 4, \ldots$ as $\tilde{\underline{X}}'M\tilde{\underline{X}}$. Hence, (ii) follows from (i) and

$$P\left(\pm\frac{\underline{X}'M\underline{X} - \mathbb{E}\underline{X}'M\underline{X}}{\bar{\sigma}_T} \geqslant x\right) \leqslant 2P\left(\pm\frac{\underline{X}'M\underline{X} + \xi_T - \mathbb{E}\underline{X}'M\underline{X}}{\bar{\sigma}_T} \geqslant x\right). \qquad \square$$

Remark 2. It can be seen from the proof of Lemma A.1 that a certain finite number L of moments in (A1)(ii) would be enough to guarantee a domain of attraction $[-\sqrt{C_L \log(T)}, \sqrt{C_L \log(T)}]$, where C_L has to be chosen in accordance with L. This is in analogy to the situation for sums of independent random variables, where also a certain finite number of moments is enough to guarantee asymptotic normality on a domain of attraction $[-\sqrt{C_L \log(T)}, \sqrt{C_L \log(T)}]$; cf. Amosova (1972).

Remark 3. As a simple example, we briefly consider quadratic forms with Toeplitz matrices M, that is $M_{s,t} = M_{|s-t|}$. We consider the simple case of $M_{s,t} \in \{0, 1\}$ for all s, t.

Case 1: If $M_{s,t} = I(s=t)$, then we have asymptotic normality, where $\lim_{T\to\infty}\{T^{-1}\sigma_T^2\}$ usually depends on the fourth-order cumulants as well.

Case 2: If $M_{s,t} = I(|s-t| \leqslant \Delta_T)$, where $\Delta_T \asymp T^\eta$ for some $0 < \eta < 1$, then we have asymptotic normality where $\lim_{T\to\infty}\{T^{-1}\sigma_T^2\}$ does not depend on the fourth-order cumulants.

Case 3: If $M_{s,t} = 1\ \forall s, t$, then $T^{-1}\underline{X}'M\underline{X}$ is asymptotically χ_1^2 distributed.

These three cases have their approximate counterparts in spectral density estimation for stationary processes. Case 1 corresponds to the case of a smoothed periodogram with fixed window, whereas case 2 corresponds to the case with a converging window. Finally, case 3 corresponds roughly to a single value of the periodogram.

The following lemma gives an upper estimate in terms of $\tilde{\mathrm{O}}$ for linear forms.

Lemma A.4. *Suppose that* $\mathbb{E}X_t = 0$ *and* (A1) *are fulfilled.*
Then

$$\sum_{t=1}^{T} w_t X_t = \tilde{\mathrm{O}}(\|w\|_{l_2}\{\sqrt{\log T} + [\|w\|_{l_1}\|w\|_{l_\infty}^2/\|w\|_{l_2}^3]T^\delta\}).$$

Proof. From

$$\mathrm{cum}_k\left(\sum w_t X_t\right) = \sum_{1\leqslant t_1,\ldots,t_k\leqslant T} \mathrm{cum}(w_{t_1}X_{t_1}, \ldots, w_{t_k}X_{t_k}),$$

we obtain that

$$\left|\mathrm{cum}_k\left(\sum w_k X_t\right)\right| \leqslant \sup_t\{|w_t|^{k-1}\} \sup_{t_1}\left\{\sum_{t_2,\ldots,t_k} |\mathrm{cum}(X_{t_1}, \ldots, X_{t_k})|\right\} \sum_t |w_t|$$
$$\leqslant C^k k! \|w\|_1 \|w\|_\infty^{k-1} T^{\delta k}. \tag{A.5}$$

On the other hand, we have that

$$\mathrm{cum}_2\left(\sum w_t X_t\right) = \mathrm{var}\left(\sum w_t X_t\right) = w' \,\mathrm{Cov}(\underline{X}) w = \mathrm{O}(\|w\|_2^2). \tag{A.6}$$

Because of $\|w\|_2^2 \leqslant \|w\|_1 \|w\|_\infty$ we get by (A.5) and (A.6) that

$$\left|\mathrm{cum}_k\left(\sum w_t X_t / \|w\|_2\right)\right| \leqslant \left(\frac{k!}{2}\right) C^{k-1} \left(\frac{\|w\|_1 \|w\|_\infty^2 T^\delta}{\|w\|_2^3}\right)^{k-2}$$

holds for $k=2,3,\ldots$, which yields the assertion by Lemma 2.1 in Bentkus and Rudzkis (1980). □

Appendix B. Proofs of the assertions

Proof of Lemma 2.1. (i) and (ii): Let $Y_{t,T} = X_{t,T} - \mu(t/T)$. We have

$$\begin{aligned} F(J_T^\mu) &= \sum_k \int_0^1 q(u) Y_{[uT+0.5-k/2],T} Y_{[uT+0.5+k/2],T} \left[\frac{1}{2\pi}\int p(u,\lambda)\exp(-\mathrm{i}k\lambda)\,\mathrm{d}\lambda\right] \mathrm{d}u \\ &= \sum_{s,t} M_{s,t} Y_{s,T} Y_{t,T}, \end{aligned}$$

where

$$M_{s,t} = \frac{1}{2\pi}\int_{(t+s-1)/(2T)}^{(t+s+1)/(2T)} q(u) \int_{-\pi}^{\pi} p(u,\lambda)\exp(-\mathrm{i}(t-s)\lambda)\,\mathrm{d}\lambda\,\mathrm{d}u$$

and $p(u,\lambda)$ is supposed to be 0 for $u \notin [0,1]$. For the matrix $M = ((M_{s,t}))_{s,t=1,\ldots,T}$, we get the relations

$$\max_{s,t}\{|M_{s,t}|\} = \mathrm{O}(T^{-1}\|q\|_\infty),$$

$$\tilde{M} \leqslant \sum_{s,t} |M_{s,t}| = \mathrm{O}(\|q\|_1)$$

and

$$\|M\|_\infty = \mathrm{O}(T^{-1}\|q\|_\infty).$$

Hence, we have $\|M\|_\infty = \mathrm{O}(T^{-1/2-\delta})$ and $\max_{s,t}\{|M_{s,t}|\}\tilde{M}\|M\|_\infty = \mathrm{O}(T^{-3/2-\delta})$, which yields (i) and (ii) by Lemma A.3.

(iii) Using

$$\mathrm{cov}(X_{[t+0.5-s/2],T}; X_{[t+0.5+s/2],T}) = \int_{-\pi}^{\pi} A^0_{[t+0.5-s/2],T}(\mu)\overline{A^0_{[t+0.5+s/2],T}(\mu)}\exp(\mathrm{i}\mu s)\,\mathrm{d}\mu$$

and

$$f(u,\lambda) = \frac{1}{2\pi}\sum_{s=-\infty}^{\infty}\int_{-\pi}^{\pi} f(u,\mu)\exp(\mathrm{i}\mu s)\,\mathrm{d}\mu\exp(-\mathrm{i}\lambda s),$$

we obtain

$$\begin{aligned}
&\mathbb{E}F(J_T^{\mu}) - F(f) \\
&= \frac{1}{2\pi}\int_0^1 q(u)\int_{-\pi}^{\pi} \sum_{s:1\leqslant[uT+0.5-s/2],[uT+0.5+s/2]\leqslant T} \int_{-\pi}^{\pi} p(u,\lambda) \\
&\quad \times \left[A^0_{[uT+0.5-s/2],T}(\mu)\overline{A^0_{[uT+0.5+s/2],T}(\mu)} - f(u,\mu)\right] \\
&\quad \times \exp(\mathrm{i}(\mu-\lambda)s)\,\mathrm{d}\mu\,\mathrm{d}\lambda\,\mathrm{d}u - \frac{1}{2\pi}\int_0^1 q(u)\int_{-\pi}^{\pi} p(u,\lambda) \\
&\quad \times \sum_{s:[uT+0.5-|s|/2]<1 \text{ or } [uT+0.5+|s|/2]>T} \int_{-\pi}^{\pi} f(u,\mu)\exp(\mathrm{i}(\mu-\lambda)s)\,\mathrm{d}\mu\,\mathrm{d}\lambda\,\mathrm{d}u \\
&= R_1 + R_2.
\end{aligned}$$

Now we have, by (A1), that

$$\begin{aligned}
R_1 &= \mathrm{O}\Bigg(\|q\|_\infty \sum_s \sup_u \left\{ \left| \int p(u,\lambda)\exp(-\mathrm{i}\lambda s)\,\mathrm{d}\lambda \right| \right\} \\
&\quad \times \sup_\mu \left\{ \int_0^1 |A^0_{[uT+0.5-s/2],T}(\mu)\overline{A^0_{[uT+0.5+s/2],T}(\mu)} - f(u,\mu)|\,\mathrm{d}u \right\} \Bigg) \\
&= \mathrm{O}(\|q\|_\infty T^{-1})
\end{aligned}$$

and

$$\begin{aligned}
R_2 &= \mathrm{O}\Bigg(\|q\|_\infty \sum_s \sup_u \left\{ \left| \int_{-\pi}^{\pi} p(u,\lambda)\exp(-\mathrm{i}\lambda s)\,\mathrm{d}\lambda \right| \left| \int_{-\pi}^{\pi} f(u,\mu)\exp(\mathrm{i}\mu s)\,\mathrm{d}\mu \right| \right\} \\
&\quad \times \int_0^1 I([uT+0.5-|s|/2]<1 \text{ or } [uT+0.5+|s|/2]>T)\,\mathrm{d}u \Bigg) \\
&= \mathrm{O}(\|q\|_\infty T^{-1}).
\end{aligned}$$

(iv) We define the vectors $\mu = (\mu(1/T),\ldots,\mu(T/T))'$, $Y = (Y_{1,T},\ldots,Y_{T,T})'$, and the smoothing matrix W such that $WY = (\tilde{\mu}(1/T),\ldots,\tilde{\mu}(T/T))'$. Now we split up

$$\begin{aligned}
F(J_T^{\tilde{\mu}}) &= (Y + (\mu - W\mu) - WY)'M(Y + (\mu - W\mu) - WY) \\
&= F(J_T^{\mu}) - 2Y'MWY + Y'W'MWY \\
&\quad + 2(\mu - W\mu)'MY - 2(\mu - W\mu)'MWY + (\mu - W\mu)'M(\mu - W\mu) \\
&= F(J_T^{\mu}) + T_1 + \cdots + T_5. \qquad \text{(B.1)}
\end{aligned}$$

We get easily that

$$\begin{aligned}
\mathbb{E}T_1 &= -2\sum_{s,t} M_{s,t}\left[\sum_u w_u(t/T)\mathbb{E}Y_{s,T}Y_{u,T}\right] \\
&= \mathrm{O}\left((Tb)^{-1}\sum_{s,t}|M_{s,t}| \right) = \mathrm{O}((Tb)^{-1}\|q\|_1). \qquad \text{(B.2)}
\end{aligned}$$

To find an upper estimate for $T_1 - \mathbb{E}T_1$, we write $T_1 = \sum_k T_{1k}$, where $T_{1k} = -2Y'M^{(k)}WY$ and

$$M_{s,t}^{(k)} = I(|t-s| = k)M_{s,t}.$$

Since Lemma A.3 requires a symmetric matrix, we simply write $T_{1k} = -2Y'N^{(k)}Y$, where $N^{(k)} = (M^{(k)}W + (M^{(k)}W)')/2$. Now, we get

$$\begin{aligned}
&(M^{(k)}W)_{s,t}\\
&\quad = M_{s,s+k}W_{s+k,t} + M_{s,s-k}W_{s-k,t}\\
&\quad = \begin{cases} O((Tb)^{-1}(|M_{s,s+k}| + |M_{s,s-k}|)) & \text{if } |s+k-t| \leqslant CTb \text{ or } |s-k-t| \leqslant CTb,\\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

which implies that

$$\begin{aligned}
\operatorname{tr}((N^{(k)})^2) &\leqslant \sum_{s,t} |(M^{(k)}W)_{s,t}|^2\\
&= O\left(\sum_s (|M_{s,s+k}| + |M_{s,s-k}|)^2 (Tb)^{-1}\right) = O(c_k^2 \|q\|_2^2 T^{-1}(Tb)^{-1}),
\end{aligned}$$

where $c_k = \sup_u\{|\int p(u,\lambda)\exp(-\mathrm{i}k\lambda)\,\mathrm{d}\lambda|\}$. Further, we have

$$\max_{s,t}\{|N_{s,t}^{(k)}|\} = O(c_k T^{-1}\|q\|_\infty (Tb)^{-1}),$$

$$\widetilde{N^{(k)}} = O(c_k T^{-1}\|q\|_\infty)$$

and

$$\|N^{(k)}\|_\infty = O(c_k T^{-1}\|q\|_\infty).$$

By Lemma A.3 we get the estimate

$$T_{1k} - \mathbb{E}T_{1k} = \tilde{O}(c_k[\|q\|_2 T^{-1/2} + \|q\|_\infty T^{-1}](Tb)^{-1/2}\sqrt{\log T} + c_k T^{-1}\|q\|_\infty T^\delta).$$

Because of $\sum c_k = O(1)$ we get

$$T_1 - \mathbb{E}T_1 = \tilde{O}(\|q\|_2 T^{-1/2}(Tb)^{-1/2}\sqrt{\log T} + \|q\|_\infty T^{\delta-1}). \tag{B.3}$$

T_2 can be estimated analogously.

To get an estimate for T_3, observe that

$$\begin{aligned}
\|M(\mu - W\mu)\|_2 &\leqslant \sum_k \|M^{(k)}(\mu - W\mu)\|_2\\
&= \sum_k O(c_k\|q\|_2 T^{-1/2}b^r) = O(\|q\|_2 T^{-1/2}b^r),
\end{aligned}$$

$$\begin{aligned}
\|M(\mu - W\mu)\|_1 &\leqslant \sum_k \|M^{(k)}(\mu - W\mu)\|_1\\
&= \sum_k O(c_k\|q\|_1 b^r) = O(\|q\|_1 b^r)
\end{aligned}$$

and

$$\|M(\mu - W\mu)\|_\infty \leqslant \sum_k \|M^{(k)}(\mu - W\mu)\|_\infty$$
$$= \sum_k \mathrm{O}(c_k\|q\|_\infty T^{-1}b^r) = \mathrm{O}(\|q\|_\infty T^{-1}b^r).$$

By Lemma A.4 we obtain that

$$T_3 = \tilde{\mathrm{O}}(\|q\|_2 T^{-1/2}b^r[\sqrt{\log T} + T^{\delta-1/2}\|q\|_1\|q\|_\infty^2/\|q\|_2^3]). \tag{B.4}$$

The term T_4 can be estimated analogously.

Since $(\mu - W\mu)_t = \mathrm{O}(b^r)$, we get

$$T_5 = \mathrm{O}\left(b^{2r}\sum_{s,t}|M_{s,t}|\right) = \mathrm{O}(\|q\|_1 b^{2r}). \tag{B.5}$$

Collecting the upper estimates from (B.2)–(B.5), we get the assertion. □

Proof of Lemma 3.1. Let $\alpha_k = ((\alpha_{j^*,k-N}^{[-N]})', \ldots, (\alpha_{j^*,k-N}^{[-N]})')'$. A two-fold application of (A4)(iii) provides that

$$\sum_{k=N+1}^{2^{j^*}} \|\alpha_{j^*,k,\inf}^{[0]} - \alpha_{j^*,k}\|^2 + \sum_{k=1}^{N} \|\alpha_{j^*,N+1-k;\inf}^{[-k]} - \alpha_{j^*,N+1-k}^{[-k]}\|^2$$
$$+\|\alpha_{j^*,2^{j^*}+k;\inf}^{[k]} - \alpha_{j^*,2^{j^*}+k}^{[k]}\|^2$$
$$\leqslant \sum_k \|\alpha_{k;\inf} - \alpha_k\|^2$$
$$\leqslant \sum_k \int_{\Delta_k} \left\| \sum_{l=-N}^{N} \alpha_{j^*,k+l;\inf}^{[l]}\phi_{j^*,k+l}^{[l]}(u) - \theta_0(u)\right\|^2 \mathrm{d}u$$
$$\asymp \sum_k \int_{\Delta_k}\int_{-\pi}^{\pi}\left[\left\{\log f_{\alpha_{k;\inf}}(u,\lambda) - \frac{f(u,\lambda)}{f_{\alpha_{k;\inf}}(u,\lambda)}\right\}\right.$$
$$\left. - \left\{\log f_{\theta_0(u)}(\lambda) - \frac{f(u,\lambda)}{f_{\theta_0(u)}(\lambda)}\right\}\right] \mathrm{d}\lambda\,\mathrm{d}u$$
$$\leqslant \sum_k \int_{\Delta_k}\int_{-\pi}^{\pi}\left[\left\{\log f_{\alpha_k}(u,\lambda) - \frac{f(u,\lambda)}{f_{\alpha_k}(u,\lambda)}\right\}\right.$$
$$\left. - \left\{\log f_{\theta_0(u)}(\lambda) - \frac{f(u,\lambda)}{f_{\theta_0(u)}(\lambda)}\right\}\right] \mathrm{d}\lambda\,\mathrm{d}u$$
$$\asymp \sum_k \int_{\Delta_k} \left\| \sum_{l=-N}^{N} \alpha_{j^*,k+l}^{[l]}\phi_{j^*,k+l}^{[l]}(u) - \theta_0(u)\right\|^2 \mathrm{d}u.$$

According to Theorem 2 in Donoho and Johnstone (1998) we obtain that

$$\sum_k \int_{\Delta_k} \left\| \sum_l \alpha^{[l]}_{j^*,k+l} \phi^{[l]}_{j^*,k+l}(u) - \theta_0(u) \right\|^2 \, \mathrm{d}u = \mathrm{O}(2^{-2j^* s}),$$

which proves the assertion. □

Proof of Theorem 3.1. (i) Since $\theta_0^{(i)}$ is uniformly continuous on [0,1] we obtain that

$$\max_k \left\{ \sup_{u \in \Delta_k} \{\theta_0^{(i)}(u)\} - \inf_{u \in \Delta_k} \{\theta_0^{(i)}(u)\} \right\} = \mathrm{o}(1).$$

Moreover, since one can choose $(\alpha^{[-N]}_{j^*,k-N}, \ldots, \alpha^{[N]}_{j^*,k+N})'$ in such a way that $\sum_{l=-N}^{N} \alpha^{[l]}_{j^*,k+l;i} \phi^{[l]}_{j^*,k+l}(u) = 1$ for all $u = \Delta_k$, we obtain by (A4)(ii) that

$$\begin{aligned} &D^{(k)}(\alpha_{k;\inf}) - \frac{1}{4\pi} \int_{\Delta_k} \int_{-\pi}^{\pi} \left\{ \log f_{\theta_0(u)}(\lambda) - \frac{f(u,\lambda)}{f_{\theta_0(u)}(\lambda)} \right\} \mathrm{d}\lambda \, \mathrm{d}u \\ &\quad \asymp \int_{\Delta_k} \left\| \sum_{l=-N}^{N} (\alpha^{[l]}_{j^*,k+l;\inf})_i \phi^{[l]}_{j^*,k+l}(u) - \theta_0^{(i)}(u) \right\|^2 \\ &\quad = \mathrm{o}(2^{-j^*}). \end{aligned}$$

On the other hand, if α lies on the boundary of the corresponding set of admissible values, then we get by (A4)(ii)

$$D^{(k)}(\alpha) - \frac{1}{4\pi} \int_{\Delta_k} \int_{-\pi}^{\pi} \left\{ \log f_{\theta_0(u)}(\lambda) - \frac{f(u,\lambda)}{f_{\theta_0(u)}(\lambda)} \right\} \mathrm{d}\lambda \, \mathrm{d}u \geqslant C 2^{-j^*},$$

for some constant $C > 0$. Hence, for T sufficiently large, all $\alpha_{k;\inf}$ are interior points of the corresponding sets of admissible values.

(ii) According to (A4)(iii), we have

$$D^{(k)}(\tilde{\alpha}_k) - D^{(k)}(\alpha_{k;\inf}) \geqslant C \|\tilde{\alpha}_k - \alpha_{k;\inf}\|^2,$$

which yields, in conjunction with $D_T^{(k)}(\alpha_{k;\inf}) - D_T^{(k)}(\tilde{\alpha}_k) \geqslant 0$, that

$$[D_T^{(k)}(\alpha_{k;\inf}) - D^{(k)}(\alpha_{k;\inf})] - [D_T^{(k)}(\tilde{\alpha}_k) - D_T^{(k)}(\tilde{\alpha}_k)] \geqslant C \|\tilde{\alpha}_k - \alpha_{k;\inf}\|^2. \tag{B.6}$$

By (A4)(i) it is easy to see that

$$\sum_s \sup_{u \in \Delta_k} \left\{ \left| \int_{-\pi}^{\pi} \left[\frac{1}{f_{\alpha_{k;\inf}}(u,\lambda)} - \frac{1}{f_{\tilde{\alpha}_k}(u,\lambda)} \right] \exp(-\mathrm{i}s\lambda) \, \mathrm{d}\lambda \right| \right\} \leqslant C 2^{j^*/2} \|\alpha_{k;\inf} - \tilde{\alpha}_k\|.$$

Hence, we obtain, by applying Lemma A.3 and (ii) and (iii) of Lemma 2.1 on a sufficiently fine grid, that

$$\begin{aligned}
&[D_T^{(k)}(\alpha_{k;\inf}) - D^{(k)}(\alpha_{k;\inf})] - [D_T^{(k)}(\alpha_k) - D_T^{(k)}(\alpha_k)] \\
&\quad = \frac{1}{4\pi} \int_{\Delta_k} \int_{-\pi}^{\pi} \left[\frac{1}{f_{\alpha_{k;\inf}}(u,\lambda)} - \frac{1}{f_{\alpha_k}(u,\lambda)} \right] [J_T^{\tilde{\mu}}(u,\lambda) - f(u,\lambda)] \, \mathrm{d}\lambda \, \mathrm{d}u \\
&\quad \leqslant C_\lambda T^{-1/2} \sqrt{\log(T)} \|\alpha_k - \alpha_{k;\inf}\| \qquad \text{(B.7)}
\end{aligned}$$

is simultaneously satisfied for all α_k in a compact set with a probability of $1 - \mathrm{O}(T^{-\lambda})$. (B.6) and (B.7) imply

$$\|\tilde{\alpha}_k - \alpha_{k;\inf}\| = \tilde{\mathrm{O}}(T^{-1/2}\sqrt{\log(T)}). \qquad \text{(B.8)}$$

Therefore, we have

$$\sup_{u \in \Delta_k} \left\{ \left\| \sum_{l=-N}^{N} \alpha_{j^*,k+l;\inf}^{(l)} \phi_{j^*,k+l}^{(l)}(u) - \theta_0(u) \right\| \right\} = \tilde{\mathrm{O}}(2^{j^*/2} T^{-1/2} \sqrt{\log T}), \qquad \text{(B.9)}$$

that means, with a probability exceeding $1 - \mathrm{O}(T^{-\lambda})$, the $\tilde{\alpha}_k$ are also interior points of the corresponding sets of admissible values.

(iii) Assume in the following that both $\alpha_{k;\inf}$ and $\tilde{\alpha}_k$ are interior points. Then $\nabla D_T^{(k)}(\tilde{\alpha}_k) = \nabla D^{(k)}(\alpha_{k;\inf}) = 0$, which implies that

$$\begin{aligned}
0 &= \nabla D_T^{(k)}(\tilde{\alpha}_k) - \nabla D_T^{(k)}(\alpha_{k;\inf}) + \nabla D_T^{(k)}(\alpha_{k;\inf}) - \nabla D^{(k)}(\alpha_{k;\inf}) \\
&= \nabla^2 D^{(k)}(\alpha_{k;\inf})(\tilde{\alpha}_k - \alpha_{k;\inf}) + \nabla D_T^{(k)}(\alpha_{k;\inf}) - \nabla D^{(k)}(\alpha_{k;\inf}) + R_k, \qquad \text{(B.10)}
\end{aligned}$$

where

$$R_k = [\nabla D_T^{(k)}(\tilde{\alpha}_k) - \nabla D_T^{(k)}(\alpha_{k;\inf})] - \nabla^2 D^{(k)}(\alpha_{k;\inf})(\tilde{\alpha}_k - \alpha_{k;\inf}). \qquad \text{(B.11)}$$

In other words, we have

$$\tilde{\alpha}_k = \alpha_{k;\inf} - (\nabla^2 D^{(k)}(\alpha_{k;\inf}))^{-1} [\nabla D_T^{(k)}(\alpha_{k;\inf}) - \nabla D^{(k)}(\alpha_{k;\inf}) + R_k]. \qquad \text{(B.12)}$$

It is clear from the mean value theorem that

$$\|R_k\| = \mathrm{O}\left(\sup_{\alpha: \|\alpha - \alpha_{k;\inf}\| \leqslant \|\tilde{\alpha}_k - \alpha_{k;\inf}\|} \max_{i_1,i_2,l_1,l_2} \left\{ \left| \frac{\partial^2}{\partial \alpha_{l_1}^{(i_1)} \partial \alpha_{l_2}^{(i_2)}} [D_T^{(k)}(\alpha) - D^{(k)}(\alpha_{k;\inf})] \right| \right\} \|\tilde{\alpha}_k - \alpha_{k;\inf}\| \right). \qquad \text{(B.13)}$$

We have that, for $\|\alpha - \alpha_{k;\inf}\| \leqslant \|\tilde{\alpha}_k - \alpha_{k;\inf}\|$,

$$\begin{aligned}
&\frac{\partial^2}{\partial\alpha_{l_1}^{(i_1)}\partial\alpha_{l_2}^{(i_2)}} D_T^{(k)}(\alpha) \\
&\quad = \frac{1}{4\pi}\int_{\Delta_k} \phi_{j^*,k+l_1}^{[l_1]}(u)\phi_{j^*,k+l_2}^{[l_2]}(u) \\
&\qquad \times \int_{-\pi}^{\pi}\left\{ \frac{f_\alpha^{(i_1,i_2)}(u,\lambda)}{f_\alpha(u,\lambda)} - \frac{f_\alpha^{(i_1)}(u,\lambda) f_\alpha^{(i_2)}(u,\lambda)}{f_\alpha^2(u,\lambda)} - \frac{J_T^{\tilde{\mu}}(u,\lambda) f_\alpha^{(i_1,i_2)}(u,\lambda)}{f_\alpha^2(u,\lambda)} \right. \\
&\qquad \left. + 2\,\frac{J_T^{\tilde{\mu}}(u,\lambda) f_\alpha^{(i_1)}(u,\lambda) f_\alpha^{(i_2)}(u,\lambda)}{f_\alpha^3(u,\lambda)} \right\} \mathrm{d}\lambda\,\mathrm{d}u \\
&\quad = \frac{\partial^2}{\partial\alpha_{l_1}^{(i_1)}\partial\alpha_{l_2}^{(i_2)}} D^{(k)}(\alpha) \\
&\qquad + \frac{1}{4\pi}\int_{\Delta_k} \phi_{j^*,k+l_1}^{[l_1]}(u)\phi_{j^*,k+l_2}^{[l_2]}(u) \int_{-\pi}^{\pi}\left\{ \frac{2 f_\alpha^{(i_1)}(u,\lambda) f_\alpha^{(i_2)}(u,\lambda)}{f_\alpha^3(u,\lambda)} - \frac{f_\alpha^{(i_1,i_2)}(u,\lambda)}{f_\alpha^2(u,\lambda)} \right\} \\
&\qquad \times [J_T^{\tilde{\mu}}(u,\lambda) - f(u,\lambda)]\,\mathrm{d}\lambda\,\mathrm{d}u \\
&\quad = \frac{\partial^2}{\partial\alpha_{l_1}^{(i_1)}\partial\alpha_{l_2}^{(i_2)}} D^{(k)}(\alpha) + \tilde{O}(2^{j^*/2} T^{-1/2}\sqrt{\log T} + 2^{j^*} T^{-1}) \\
&\quad = \frac{\partial^2}{\partial\alpha_{l_1}^{(i_1)}\partial\alpha_{l_2}^{(i_2)}} D^{(k)}(\alpha_{k;\inf}) + \tilde{O}(2^{j^*/2} T^{-1/2}\sqrt{\log T}). \qquad \text{(B.14)}
\end{aligned}$$

This implies that

$$\|R_k\| = \tilde{O}(2^{j^*/2} T^{-1}\log T). \qquad \text{(B.15)}$$

(iv) According to (3.13) we obtain

$$\tilde{\beta}_{j,k} - \beta_{j,k;\inf} = \Gamma'_{j,k}(\nabla^2 D(\underline{\alpha}_{\inf}))^{-1}[\nabla D_T(\underline{\alpha}_{\inf}) - \nabla D(\underline{\alpha}_{\inf})] + S_{j,k}, \qquad \text{(B.16)}$$

where $\nabla^2 D(\underline{\alpha}_{\inf}) = \mathrm{Diag}(\nabla^2 D^{(1)}(\alpha_{1;\inf}), \ldots, \nabla^2 D^{(2^{j^*}+N)}(\alpha_{2^{j^*}+N;\inf}))$, $\nabla D_T(\underline{\alpha}_{\inf}) = (\nabla D_T^{(1)}(\alpha_{1;\inf}), \ldots, \nabla D_T^{(2^{j^*}+N)}(\alpha_{2^{j^*}+N;\inf}))'$, $\nabla D(\underline{\alpha}_{\inf}) = (\nabla D^{(1)}(\alpha_{1;\inf}), \ldots, \nabla D^{(2^{j^*}+N)}(\alpha_{2^{j^*}+N;\inf}))'$, $\underline{\alpha}_{\inf} = (\alpha_{1;\inf}, \ldots, \alpha_{2^{j^*}+N;\inf})'$, and

$$S_{j,k} = \Gamma'_{j,k}(\nabla^2 D(\underline{\alpha}_{\inf}))^{-1}(R_1, \ldots, R_{2^{j^*}+N})'. \qquad \text{(B.17)}$$

We will show that $S_{j,k}$ is of negligible order for most of the k.

It is easy to see that

$$\|\Gamma_{j,k}\|_1 = O(2^{(j^*-j)/2}\|\Gamma_{j,k}\|_2) = O(2^{(j^*-j)/2}). \qquad \text{(B.18)}$$

Observe that the matrix

$$F(u) = ((\phi^{[l_1]}_{j^*,k+l_1}(u)\phi^{[l_2]}_{j^*,k+l_2}(u)))_{l_1,l_2=1,\dots,2N+1}$$
$$= \begin{pmatrix} \phi^{[-N]}_{j^*,k-N}(u) \\ \vdots \\ \phi^{[N]}_{j^*,k+N}(u) \end{pmatrix} (\phi^{[-N]}_{j^*,k-N}(u) \cdots \phi^{[N]}_{j^*,k+N}(u))$$

is positive semidefinite, which implies by (A4) (v) that

$$\nabla^2 D^{(k)}(\alpha) = \int_{\Delta_k} \left[F(u) \otimes \nabla^2 \int_{-\pi}^{\pi} \left\{ \log f_\alpha(u,\lambda) + \frac{f(u,\lambda)}{f_\alpha(u,\lambda)} \right\} \mathrm{d}\lambda \right] \mathrm{d}u$$
$$\geqslant \int_{\Delta_k} F(u) \otimes M \, \mathrm{d}u = I_{2N+1} \otimes M. \tag{B.19}$$

Since $\nabla^2 D(\underline{\alpha}_{\inf})$ is a block diagonal matrix, we have

$$\|(\nabla^2 D(\underline{\alpha}_{\inf}))^{-1}\|_\infty \leqslant \max_k \{\|(\nabla^2 D^{(k)}(\alpha_{k;\inf}))^{-1}\|_\infty\}$$
$$= \mathrm{O}\left(\max_k \{1/\lambda_{\min}(\nabla^2 D^{(k)}(\alpha_{k;\inf}))\} \right) = \mathrm{O}(1). \tag{B.20}$$

This yields, in conjunction with (B.15), that

$$S_{j,k} \leqslant \|\Gamma_{j,k}\|_1 \|(\nabla^2 D(\underline{\alpha}_{\inf}))^{-1}\|_\infty \|(R_1,\dots,R_{2^{j^*}+N})'\|_\infty$$
$$= \tilde{\mathrm{O}}(2^{j^*-j/2}T^{-1}\log T, T^{-\lambda}), \tag{B.21}$$

that is the remainder terms $S_{j,k}$ are negligible. Moreover, asymptotic subgaussianity of the leading term of (B.16), $\Gamma'_{j,k}(\nabla^2 D(\underline{\alpha}_{\inf}))^{-1}[\nabla D_T(\underline{\alpha}_{\inf}) - \nabla D(\underline{\alpha}_{\inf})]$, follows from Lemma 2.1. □

Proof of Theorem 4.1. First, we obtain by Parseval's identity that

$$\mathbb{E}\|\hat{\theta}^{(i)} - \theta^{(i)}\|^2 = \sum_{k \in I_l^0} E(\tilde{\alpha}_{l,k;i} - \alpha_{l,k;i})^2$$
$$+ \sum_{(j,k)\in \mathscr{J}_T} E(\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2$$
$$+ \sum_{(j,k)\notin \mathscr{J}_T} \beta^2_{j,k;i}$$
$$= T_1 + T_2 + T_3. \tag{B.22}$$

It is obvious that

$$T_1 = \mathrm{O}(T^{-1}). \tag{B.23}$$

Using Theorem 3.1 we may reduce the problem of estimating $E(\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2$ to the case of normally distributed empirical coefficients for which there are appropriate results available. Let

$$\bar{\beta}_{j,k;i} \sim \mathrm{N}(\beta_{j,k;i}, \bar{\sigma}_T^2). \tag{B.24}$$

Since $\delta^{(\cdot)}$ is monotonic in its first argument we have that

$$(\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > x$$

if and only if

$$\tilde{\beta}_{j,k;i} - \beta_{j,k;i} \geqslant / > f_{j,k;i}(x) \quad \text{or} \quad \tilde{\beta}_{j,k;i} - \beta_{j,k;i} \leqslant / < g_{j,k;i}(x)$$

for appropriate functions $f_{j,k;i}$ and $g_{j,k;i}$, where the equality sign has to be included if and only if $\delta^{(\cdot)}(\cdot, \lambda_T)$ is not left- or right-continuous, respectively, at the corresponding points. Hence, we obtain from Theorem 3.1 that

$$P((\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > x)$$
$$\leqslant 2P((\delta^{(\cdot)}(\bar{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > x)(1 + \mathrm{o}(1)) + \mathrm{O}(T^{-\lambda}). \tag{B.25}$$

Let

$$\rho_{j,k;i} = \sup\{x\colon P((\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > x) \geqslant T^{-\lambda}\}.$$

Then

$$P((\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > \rho_{j,k;i}) \leqslant T^{-\lambda},$$

which yields by the Cauchy–Schwarz inequality that

$$\int_{\rho_{j,k;i}}^{\infty} P((\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > x)\,\mathrm{d}x$$
$$\leqslant \mathbb{E}(\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 I((\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > \rho_{j,k;i})$$
$$\leqslant \sqrt{\mathbb{E}(\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^4}\sqrt{P((\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > \rho_{j,k;i})} = \mathrm{O}(T^{-\lambda}),$$

and, analogously,

$$\int_{\rho_{j,k;i}}^{\infty} P((\delta^{(\cdot)}(\bar{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > x)\,\mathrm{d}x = \mathrm{O}(T^{-\lambda}).$$

Therefore, we obtain, in conjunction with (B.25), that

$$\mathbb{E}(\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2$$
$$= \int_0^{\rho_{j,k;i}} P((\delta^{(\cdot)}(\tilde{\beta}_{j,k;i,} \lambda_T) - \beta_{j,k;i})^2 > x)\,\mathrm{d}x$$
$$+ \int_{\rho_{j,k;i}}^{\infty} P((\delta^{(\cdot)}(\tilde{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 > x)\,\mathrm{d}x$$
$$\leqslant 2(1 + \mathrm{o}(1))\mathbb{E}(\delta^{(\cdot)}(\bar{\beta}_{j,k;i}, \lambda_T) - \beta_{j,k;i})^2 + \mathrm{O}(T^{-\lambda}). \tag{B.26}$$

By Lemma 1 of Donoho and Johnstone (1994) we have that

$$\mathbb{E}(\delta^{(\cdot)}(\bar{\beta}_{j,k;i}, \lambda) - \beta_{j,k;i})^2 \leqslant C\left(\bar{\sigma}_T^2\left(\frac{\lambda}{\bar{\sigma}_T} + 1\right)\varphi\left(\frac{\lambda}{\bar{\sigma}_T}\right) + \min\{\lambda^2, \beta_{j,k;i}^2\}\right) \tag{B.27}$$

holds uniformly in $\lambda \geqslant 0$ and $\beta_{j,k;i} \in \mathbb{R}$, where φ denotes the standard normal density.

Therefore, we obtain analogously to the proof of Theorem 5.2 in Neumann (1996) that

$$T_2 = \mathrm{O}((\log T/T)^{2m/(2m+1)}) \tag{B.28}$$

and, by Theorem 2 of Donoho and Johnstone (1998),

$$T_3 = \mathrm{O}(2^{-2j^*\{m+1/2-1/(p\wedge 2)\}}) = \mathrm{O}(T^{-2m/(2m+1)}), \tag{B.29}$$

which completes the proof. □

References

Amosova, N.N., 1972. On limit theorems for probabilities of moderate deviations. Vestnik Leningrad. Univ. 13, 5–14 (in Russian).

Bentkus, R., Rudzkis, R., 1980. On exponential estimates of the distribution of random variables. Liet Mat. Rinkinys 20, 15–30 (in Russian).

Bergh, J., Löfström, J., 1976. Interpolation Spaces: An Introduction. Springer, Berlin.

Bloomfield, P., 1973. An exponential model for the spectrum of scalar time series. Biometrika 80, 817–822.

Cohen, A., Daubechies, I., Vial, P., 1993. Wavelets on the interval and fast wavelet transform. Appl. Comput. Harmonic Anal. 1, 54–81.

Dahlhaus, R., 1996a. On the Kullback-Leibler information divergence of locally stationary processes. Stochastic Process. Appl. 62, 139–168.

Dahlhaus, R., 1996b. Asymptotic Statistical Inference for Nonstationary Processes with Evolutionary Spectra. Lecture Notes in Statistics, vol. 115, Springer, New York, Athens Conference on Applied Probability and Time Series, Vol. II, pp. 145–159.

Dahlhaus, R., 1997. Fitting time series models to nonstationary processes. Ann. Statist. 25, 1–37.

Dahlhaus, R., 2000. A likelihood approximation for locally stationary processes. Ann. Statist., to appear.

Dahlhaus, R., Neumann, M.H., von Sachs, R., 1999. Nonlinear wavelet estimation of time-varying autoregressive processes. Bernoulli 5, 873–906.

Daubechies, I., 1988. Orthogonal bases of compactly supported wavelets. Comm. Pure Appl. Math. 41, 909–996.

Donoho, D.L., Johnstone, I.M., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, 425–455.

Donoho, D.L., Johnstone, I.M., 1998. Minimax estimation via wavelet shrinkage. Ann. Statist. 26, 879–921.

Gasser, T., Müller, H.G., 1979. Kernel estimation of regression functions. In Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics, vol. 757. Springer, Berlin, pp. 23–68.

Hall, P., McKay, I., Turlach, B.A., 1996. Performance of wavelet methods for functions with many discontinuities. Ann. Statist. 24, 2462–2476.

Hall, P., Patil, P., 1995. Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. Ann. Statist. 23, 905–928.

Martin, W., Flandrin, P., 1985. Wigner-Ville spectral analysis of nonstationary processes. IEEE Trans. Signal Process. 33, 1461–1470.

Meyer, Y., 1991. Ondelettes sur l'intervalle. Rev. Math. Ibero-Americana 7 (2), 115–133.

Neumann, M.H., 1996. Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. J. Time Ser. Anal. 17, 601–633.

Neumann, M.H., von Sachs, R., 1997. Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. Ann. Statist. 25, 38–76.

Rudzkis, R., 1978. Large deviations for estimates of spectrum of stationary series. Lithuanian Math. J. 18, 214–226.

Rudzkis, R., Saulis, L., Statulevicius, V., 1978. A general lemma on probabilities of large deviations. Lithuanian Math. J. 18, 226–238.

von Sachs, R., Schneider, K., 1996. Wavelet smoothing of evolutionary spectra by non-linear thresholding. Appl. Comput. Harmonic Anal. 3, 268–282.

Saulis, L., Statulevicius, V.A., 1991. Limit Theorems for Large Deviations. Kluwer, Dordrecht.

Statulevicius, V., Jakimavicius, D., 1988. Estimates of semi-invariants and centered moments of stochastic processes with mixing I. Lithuanian Math. J. 28, 67–80.

Triebel, H., 1990. Theory of Function Spaces II. Birkhäuser, Basel.

Zani, M., 1999. Large deviations for quadratic forms of Gaussian locally stationary processes. Manuscript.