*The Annals of Statistics*
1997, Vol. 25, No. 1, 38–76

# WAVELET THRESHOLDING IN ANISOTROPIC FUNCTION CLASSES AND APPLICATION TO ADAPTIVE ESTIMATION OF EVOLUTIONARY SPECTRA

BY MICHAEL H. NEUMANN AND RAINER VON SACHS

*Weierstrass Institute for Applied Analysis and Stochastics and University of Kaiserslautern*

We derive minimax rates for estimation in anisotropic smoothness classes. These rates are attained by a coordinatewise thresholded wavelet estimator based on a tensor product basis with separate scale parameter for every dimension. It is shown that this basis is superior to its one-scale multiresolution analog, if different degrees of smoothness in different directions are present.

As an important application we introduce a new adaptive wavelet estimator of the time-dependent spectrum of a locally stationary time series. Using this model which was recently developed by Dahlhaus, we show that the resulting estimator attains nearly the rate, which is optimal in Gaussian white noise, simultaneously over a wide range of smoothness classes. Moreover, by our new approach we overcome the difficulty of how to choose the right amount of smoothing, that is, how to adapt to the appropriate resolution, for reconstructing the local structure of the evolutionary spectrum in the time–frequency plane.

**1. Introduction.** There is a wide range of fields in which an observed time series shows a nonstationary behavior (by transients, amplitude or frequency modulation, quasi-oscillating behavior, etc.). These can be found, for example, in many physical phenomena (occurring in geophysics, in transmission problems like radio propagation or in speech and sound analysis), and from economical data analysis, also. A recent approach for modeling certain kinds of these nonstationarities is by the introduction of the class of locally stationary processes [Dahlhaus (1997)], which both controls the departure from stationary and gives a frame for asymptotic theory. As in the Cramér representation for stationary processes, the spectrum, which now becomes time dependent, controls the evolution of the variance–covariance distribution of the process over frequency and over time.

In the present paper we develop nonlinear wavelet estimators for this kind of time-varying spectral density: with this we address the problem of finding the right amount of smoothing of an estimator which should adaptively reconstruct the underlying structure of the spectrum in the time–frequency

Received February 1995; revised April 1996.

*AMS* 1991 *subject classifications*. Primary 62G07, 62M15; secondary 62E20, 62M10.

*Key words and phrases.* Anisotropic smoothness classes, adaptive estimation, optimal rate of convergence, wavelet thresholding, tensor product basis, time–frequency plane, locally stationary time series, evolutionary spectrum.

plane. Motivated by this problem, we study first a question of more general importance. Inference about the spectrum of a nonstationary time series is a two-dimensional estimation problem with two particular directions, time and frequency, on the plane. If, in this situation and, more generally for any multidimensional curve estimation problem, the underlying curve shows different degrees of smoothness in the different directions, then the construction of the estimator should properly take this into account. Hence, to establish a benchmark for our estimator, we derive first minimax rates for estimation in anisotropic smoothness classes. Because this question is of general interest, we do not assume any specific observation model, but we investigate this problem in Gaussian white noise. For simplicity, we consider the two-dimensional case and restrict ourselves to anisotropic Sobolev classes. Generalizations can be thought of for higher dimensions and other smoothness classes like Hölder and Besov, also. However, the technically challenging extension to anisotropic Besov classes goes beyond the scope of this paper, and should possibly be studied elsewhere.

We show that appropriately tuned wavelet estimators are able to attain the optimal rate of convergence in these classes. These estimators use coordinatewise nonlinear thresholding of empirical wavelet coefficients. The rate for the risk can be easily found by analyzing a certain complexity functional, which describes the amount of data compression of a basis in a given smoothness class. We show that we obtain a suitable higher-dimensional basis by taking respective tensor products of the one-dimensional wavelet basis. In contrast, the frequently used higher-dimensional multiresolution basis does not optimally compress the signal in anisotropic smoothness classes. This implies that any coordinatewise thresholded estimator based on such a basis is not able to attain the optimal rate of convergence.

The second part of this paper is devoted to the particular problem of spectral estimation. Throughout the paper we adopt the model of locally stationary time series developed in Dahlhaus (1997). To allow least restrictive assumptions on the smoothness of the spectrum and in order to embed it into the considered Sobolev class, we further relax the assumptions of Dahlhaus (1996a) to give a definition of the evolutionary spectrum as a function in the $L_2$-space over the time–frequency plane. Again, our main goal is to define an estimator that adapts to different degrees of smoothness in time and frequency direction, respectively. In contrast to Dahlhaus (1997) and von Sachs and Schneider (1996), who used a local periodogram on segments of length $N = N(T)$ (with $N \to \infty$ and $N/T \to 0$ as $T \to \infty$), here we define a periodogram-like pointwise statistic which can be considered as an empirical version of the local time-dependent spectrum. By this approach we avoid a kind of presmoothing in time direction and get rid of the additional smoothing parameter $N$, for which a theoretical approach to its optimal choice is still lacking. This overcomes the shortcoming of fixing with $N$ a lower bound for the ratio of the resolution in time and in frequency direction. Instead, to decide which degree of smoothing is appropriate, we project this time–frequency statistic on a suitable wavelet basis and use thresholding of the resulting coefficients. In view of the results in Section 2, in this construction, we use a tensor product basis. The appropriate tuning of the thresholds requires knowledge about the distribution of the empirical coefficients. Using cumulant techniques we prove asymptotic normality in terms of probabilities of large deviations. This implies the asymptotic risk equivalence of monotonic estimators to the case of normally distributed empirical coefficients and suggests the use of thresholding techniques prescribed by existing theory under Gaussian noise.

Finally, to obtain a fully defined threshold rule, it is natural to use some initial estimator of the standard deviation of the empirical coefficients. We show that rather weak assumptions on an initial estimator of the spectral density guarantee near-optimality of the final estimator.

The paper is organized as follows. In Section 2 we derive minimax rates in anisotropic Sobolev classes and examine the two mentioned different kinds of multidimensional wavelet bases w.r.t. their appropriateness in such function spaces. In Section 3, after introducing the model of local stationarity and an $L_2$-generalization of the definition of the evolutionary spectrum, we develop our new estimator and state theorems on rates for its risk. The proofs are contained in Section 4.

**2. Optimal estimation in anisotropic smoothness classes.** Before we develop a definite estimation method for the spectral density in the next section, we first consider a question of more general importance: we search for a basis that is appropriate for multidimensional estimation problems in situations, where we have possibly different degrees of smoothness in different directions. To do this, we consider balls in anisotropic Sobolev spaces and derive minimax rates in a Gaussian white-noise model. For simplicity, we only consider the two-dimensional case and restrict ourselves to anisotropic Sobolev spaces, although it is obvious that analogous results can be obtained in higher dimensions and for other function classes, such as, for example, anisotropic Hölder spaces. A meaningful generalization to anisotropic Besov classes seems to be possible, also, though some more careful lengthy considerations have to be taken into account. We feel that the gain in insight is too small compared to the amount of technical work, and we do not want to hide our main principal result by doing so. We show that thresholded wavelet estimators based on a tensor product wavelet basis in $L_2([0,1] \times [0,1])$ attain the optimal rate of convergence, whereas the one-scale multiresolution basis, which is often used in image analysis problems, does not share this property.

2.1. *Anisotropic Sobolev classes and multidimensional wavelet bases*. Following Nikol'skii (1975), an anisotropic Sobolev space $W_{p_1, p_2}^{m_1, m_2}$ is defined as

$$W_{p_1, p_2}^{m_1, m_2} = \left\{ f \,\middle|\, \sum_{i=1}^{2} \left( \|f\|_{p_i} + \left\| \frac{\partial^{m_i}}{\partial x_i^{m_i}} f \right\|_{p_i} \right) < \infty \right\}.$$

In the following we assume that our object of interest $f$ lies in the set

$$\mathscr{F}^{m_1, m_2}_{p_1, p_2} = \mathscr{F}(m_1, m_2, p_1, p_2, C) = \left\{ f \,\middle|\, \sum_{i=1}^{2} \left( \|f\|_{p_i} + \left\| \frac{\partial^{m_i}}{\partial x_i^{m_i}} f \right\|_{p_i} \right) \le C \right\}. \tag{2.1}$$

Whereas here this positive constant $C$ is, of course, fixed for the smoothness class under consideration, in what follows throughout the paper, $C$ is used to denote a generic positive constant, not necessarily the same in different contexts. Moreover, in the sequel, we restrict our considerations to $m_i \ge 1$, $p_i \ge 1$ and $m_i > 1/p_i$, which, in particular, implies continuity of $f$.

Assume we have an orthonormal basis of compactly supported wavelets of $L_2[0,1]$, where the functions $\phi$ and $\psi$ satisfy, for $m \ge \max\{m_1, m_2\}$, the following assumption.

ASSUMPTION 1. (i) $\phi$ and $\psi$ are in $C^m$;
(ii) $\int \phi(t)\, dt = 1$;
(iii) $\int \psi(t) t^k \, dt = 0$ for $0 \le k \le m-1$.

Such bases are given by Meyer (1991) and Cohen, Daubechies and Vial (1993).

Let $V_j$ be the subspace of $L_2[0,1]$, which is generated by $\{\phi_{jk}\}_k$. It is known that

$$L_2([0,1] \times [0,1]) = \overline{\bigcup_{j=l}^{\infty^i} V_j \otimes V_j},$$

which shows the possibility of building a basis of $L_2([0,1] \times [0,1])$ from tensor products of functions from a one-dimensional basis $\{\phi_{lk}\}_k \cup \{\psi_{jk}\}_{j \ge l;\, k}$. Let $W_j = \operatorname{span}\{\psi_{jk}\}_k$. We can write $V_{j^*}^{(2)} = V_{j^*} \otimes V_{j^*}$ as

$$\begin{aligned} V_{j^*}^{(2)} &= (V_l \oplus W_l \oplus \cdots \oplus W_{j^*-1}) \otimes (V_l \oplus W_l \oplus \cdots \oplus W_{j^*-1}) \\ &= V_l \otimes V_l \oplus \left( \bigoplus_{j=l}^{j^*-1} (W_j \otimes V_l) \right) \oplus \left( \bigoplus_{j=l}^{j^*-1} (V_l \otimes W_j) \right) \\ &\quad \oplus \left( \bigoplus_{j_1, j_2 = l}^{j^*-1} (W_{j_1} \otimes W_{j_2}) \right) \end{aligned} \tag{2.2}$$

as well as in the form

$$V_{j^*}^{(2)} = V_l \otimes V_l \oplus \bigoplus_{j=l}^{j^*-1} \left[ (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j) \right]. \tag{2.3}$$

According to (2.2), we obtain a basis $\mathscr{B}$ of $L_2([0,1] \times [0,1])$ as

$$\begin{aligned} \mathscr{B} = &\left\{\phi_{lk_1}(x_1)\phi_{lk_2}(x_2)\right\}_{k_1,k_2} \cup \left( \bigcup_{j_1 \geq l} \left\{\psi_{j_1k_1}(x_1)\phi_{lk_2}(x_2)\right\}_{k_1,k_2}\right) \\ &\cup \left( \bigcup_{j_2 \geq l} \left\{\phi_{lk_1}(x_1)\psi_{j_2k_2}(x_2)\right\}_{k_1,k_2}\right) \\ &\cup \left( \bigcup_{j_1, j_2 \geq l} \left\{\psi_{j_1k_1}(x_1)\psi_{j_2k_2}(x_2)\right\}_{k_1,k_2}\right). \end{aligned} \tag{2.4}$$

We like to compare that with another construction, which corresponds to decomposition (2.3) and which is not only widely used in image analysis: because it actually provides a two-dimensional multiresolution analysis, it is also of importance in operator theory, where it is called "nonstandard decomposition" [Beylkin, Coifman and Rokhlin (1993)]. It is given by

$$\begin{aligned} \overline{\mathscr{B}} = &\left\{\phi_{lk_1}(x_1)\phi_{lk_2}(x_2)\right\}_{k_1,k_2} \\ &\cup \bigcup_{j \geq l} \left\{\phi_{jk_1}(x_1)\psi_{jk_2}(x_2), \psi_{jk_1}(x_1)\phi_{jk_2}(x_2), \psi_{jk_1}(x_1)\psi_{jk_2}(x_2)\right\}_{k_1,k_2}. \end{aligned} \tag{2.5}$$

Note that for both $\mathscr{B}$ and $\overline{\mathscr{B}}$ we can also use different one-dimensional bases to build a two-dimensional basis, which is done in Section 3 in view of the special problem considered there.

It appears that, because of its more appealing structure, basis $\overline{\mathscr{B}}$ is more often used for two-dimensional estimation problems; see, for example, Delyon and Juditsky (1996), Tribouley (1995) and von Sachs and Schneider (1996). Its use seems to be appropriate in most frequently considered smoothness classes, such as, for example, isotropic Sobolev or Besov classes. However, in certain practical problems, for the curve we are interested in we could expect different smoothness properties in different directions. We will show that under such anisotropic smoothness priors basis $\overline{\mathscr{B}}$ is no longer appropriate.

For the sake of notational convenience, we slightly abuse the notation and define $\psi_{l-1,k} := \phi_{lk}$. Further, by $\mu_I$ we denote the basis functions in $\mathscr{B}$ using the multiindex $I = (j_1, j_2, k_1, k_2)$. Let $\Theta = \{(\theta_I) | \sum_I \theta_I \mu_I \in \mathscr{F}^{m_1, m_2}_{p_1, p_2}\}$. By Parseval's equality we see that the $L_2$-loss $\|\sum \hat{\theta}_I \mu_I - f\|^2$ of any estimator $\hat{f} = \sum \hat{\theta}_I \mu_I$ in the function space is equal to the $l_2$-loss $\sum_I (\hat{\theta}_I - \theta_I)^2$ in the sequence space, where

$$\theta_I = \iint \mu_I(x_1, x_2) f(x_1, x_2)\, dx_1\, dx_2$$

are the wavelet coefficients of $f$.

The reader who is primarily interested in the estimation of time-varying spectra, rather than in details concerning minimax rates of convergence in $\mathscr{F}^{m_1, m_2}_{p_1, p_2}$, can directly proceed to Section 3 on an important application of using the tensor product basis $\mathscr{B}$.

2.2. *Minimax rates of convergence in anisotropic Sobolev classes.* Since the problem investigated in this section seems to be of general interest in many statistical estimation problems, we do not want to specify any specific observation model. Instead, we assume that function-valued observations $Y(x_1, x_2)$ from the Gaussian white-noise model

$$Y(x_1, x_2) = \int_0^{x_2} \int_0^{x_1} f(z_1, z_2)\, dz_1\, dz_2 + \varepsilon W(x_1, x_2) \tag{2.6}$$

are available. Here $W$ is a Brownian sheet [cf., e.g., Walsh (1986)] and $\varepsilon > 0$ is the noise level.

REMARK 2.1. In the one-dimensional case it is well known that the difficulty in estimating $f$ in Gaussian white noise

$$Y(t) = \int_0^t f(s)\, ds + \varepsilon W_t, \tag{2.7}$$

where $W_t$ is a standard Wiener process, is closely related to the difficulty in estimating $f$ in non-Gaussian or non-i.i.d. situations, which is actually the interesting problem. Recently, this connection between nonparametric regression and model (2.7) has been established in a decision-theoretic manner by Brown and Low (1992). The equivalence between density estimation and some slightly modified version of (2.7) was shown by Nussbaum (1994).

For wavelet estimators this close connection often materializes also at the practical level. So it was shown in Neumann and Spokoiny (1995) for non-Gaussian regression and in Neumann (1994) for spectral density estimation that the empirical coefficients coming from these models are asymptotically normally distributed in a sufficiently strong sense. Then, for certain nonlinear wavelet estimators, it was possible to derive the risk equivalence between model (2.7) and the abovementioned models. We think that the two-dimensional continuous Gaussian model (2.6) will be again an appropriate counterpart for many practically relevant estimation problems.

We obtain empirical coefficients from the observation model (2.6) as

$$\tilde{\theta}_I = \iint \mu_I(x_1, x_2)\, dY(x_1, x_2) = \theta_I + \varepsilon \xi_I, \tag{2.8}$$

where $\xi_I \sim N(0, 1)$ are i.i.d.

First, we derive a lower bound for the minimax risk in model (2.6) under the assumption that $f \in \mathscr{F}_{p_1, p_2}^{m_1, m_2}$. Since we are only interested in the optimal *rate*, we can use a simple approach developed in Bretagnolle and Huber (1979). First, we establish the following lemma, which provides a lower bound for the complexity of the set $\Theta$.

LEMMA 2.1. *Suppose that Assumption* 1 *holds. The set* $\Theta$ *contains a hypercube of sidelength* $2\varepsilon$:

$$\Theta_\varepsilon = \{(\theta_I) | \theta_I \in [-\varepsilon, \varepsilon] \text{ for } I \in \mathscr{I}_\varepsilon \text{ and } \theta_I = 0 \text{ for } I \notin \mathscr{I}_\varepsilon\},$$

*with*

$$\dim(\Theta_\varepsilon) = \#\mathcal{I}_\varepsilon \asymp \varepsilon^{-2(m_1+m_2)/(2m_1m_2+m_1+m_2)}.$$

If we now take independent, uniformly distributed priors on $[-\varepsilon, \varepsilon]$ for $I \in \mathcal{I}_\varepsilon$, due to the independence of the $\tilde{\theta}_I$'s we obtain a Bayes risk of order $\varepsilon^{2-2(m_1+m_2)/(2m_1m_2+m_1+m_2)}$. This implies the following theorem.

THEOREM 2.1. *Denote by $\hat{f}$ any estimator of a member $f \in \mathcal{F}_{p_1,p_2}^{m_1,m_2}$. Then*

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_{p_1,p_2}^{m_1,m_2}} \left\{\mathbb{E}\|\hat{f} - f\|^2\right\} \geq C\varepsilon^{2\vartheta(m_1,m_2)},$$

*where*

$$\vartheta(m_1, m_2) = \frac{2m_1m_2}{2m_1m_2 + m_1 + m_2}.$$

Note that this rate can be written in the form $\varepsilon^{4\overline{m}/(2\overline{m}+2)}$ for $1/\overline{m} = \frac{1}{2}(1/m_1 + 1/m_2)$, which is the well-known optimal rate of convergence for two-dimensional isotropic smoothness classes of degree $\overline{m}$. Hence, this value of $\overline{m}$ can be interpreted as an appropriate notion of an "average smoothness" of our anisotropic class $\mathcal{F}_{p_1,p_2}^{m_1,m_2}$.

To show that this rate is actually attainable, we consider a certain complexity functional $\tilde{\Omega}_\varepsilon$ to be defined further below, which is similar to the modulus of continuity

$$\Omega_\varepsilon\left(\mathcal{B}, \mathcal{F}_{p_1,p_2}^{m_1,m_2}\right) = \sup_{f \in \mathcal{F}_{p_1,p_2}^{m_1,m_2}} \left\{\sum_I \min\left\{\varepsilon^2, \theta_I^2\right\}\right\} \tag{2.9}$$

considered in Donoho and Johnstone (1994a). There it was shown that $\Omega_\varepsilon$ gives an almost complete information about uniform rates for diagonal estimators in model (2.6).

Two commonly used rules to treat the coefficients are:

1. hard thresholding

$$\delta^{(h)}\left(\tilde{\theta}_I, \lambda\right) = \tilde{\theta}_I I\left(|\tilde{\theta}_I| \geq \lambda\right);$$

2. soft thresholding

$$\delta^{(s)}\left(\tilde{\theta}_I, \lambda\right) = \left(|\tilde{\theta}_I| - \lambda\right)_+ \operatorname{sgn}\left(\tilde{\theta}_I\right).$$

In the following, $\delta^{(\cdot)}$ is used to (somewhat sloppily) denote either $\delta^{(h)}$ or $\delta^{(s)}$.

Following the developments in Donoho and Johnstone (1994a), we can derive an estimator that attains the rate prescribed by the modulus of continuity $\Omega_\varepsilon(\mathcal{B}, \mathcal{F}_{p_1,p_2}^{m_1,m_2})$ up to a factor of $\log(1/\varepsilon)$. To prove that the rate $\varepsilon^{2\vartheta(m_1,m_2)}$ is exactly attainable, we have to modify $\Omega_\varepsilon$ slightly. First, by Lemma 1 of Donoho and Johnstone (1994a), we can prove that the relation

$$\mathbb{E}\left(\delta^{(\cdot)}\left(\tilde{\theta}_I, \lambda\right) - \theta_I\right)^2 \leq C\left(\varepsilon^2\left(\frac{\lambda}{\varepsilon} + 1\right)\varphi\left(\frac{\lambda}{\varepsilon}\right) + \min\left\{\lambda^2, \theta_I^2\right\}\right) \tag{2.10}$$

holds uniformly in $\lambda \geq 0$ and $\theta_I \in \mathbb{R}$, where $\varphi$ denotes the standard normal density. This motivates us to define the complexity functional

$$\tilde{\Omega}_{\varepsilon}\left(\mathscr{B}, \mathscr{F}_{p_1, p_2}^{m_1, m_2}\right) = \inf_{(\lambda_I)} \sup_{f \in \mathscr{F}_{p_1, p_2}^{m_1, m_2}} \left\{ \sum_I \left( \varepsilon^2 \left( \frac{\lambda_I}{\varepsilon} + 1 \right) \varphi\left( \frac{\lambda_I}{\varepsilon} \right) + \min\{\lambda_I^2, \theta_I^2\} \right) \right\}. \tag{2.11}$$

The essential reason why the modulus of continuity $\Omega_{\varepsilon}$ does not immediately provide an attainable rate for estimators is that it does not take the possible sparsity of the signal into account. In cases where we have a too large number of potentially important coefficients, we lose an additional log-term as we do not know which are the really important ones. To get rid of the high amount of variability caused by this large set of empirical coefficients, one chooses the thresholds $\lambda_I$ by a logarithmic factor larger than the noise level. Accordingly, the functional $\tilde{\Omega}_{\varepsilon}$ penalizes such cases of extreme sparsity by the additional terms $(\lambda_I/\varepsilon + 1)\varphi(\lambda_I/\varepsilon)$, which arise from upper estimates of tail probabilities of Gaussian random variables.

The next lemma shows a particular choice of the vector $(\lambda_I)$, which provides the rate $\varepsilon^{2\vartheta(m_1, m_2)}$ for the right-hand side of (2.11).

LEMMA 2.2. *Suppose that Assumption* 1 *holds. Let* $\lambda_{I,\varepsilon}$ *be such that*

$$\lambda_{I,\varepsilon} = \begin{cases} 0, & \text{if } j_1 \leq j_1^* \text{ and } j_2 \leq j_2^*, \\ \varepsilon K_{m_1, m_2} \sqrt{\max\{(j_1 - j_1^*)/m_2, (j_2 - j_2^*)/m_1\}}, & \text{otherwise}, \end{cases}$$

*where*

$$2^{j_1^*} = \varepsilon^{-2/(2m_1 + 1 + m_1/m_2)}, \qquad 2^{j_2^*} = \varepsilon^{-2/(2m_2 + 1 + m_2/m_1)}$$

*and* $K_{m_1, m_2} > \sqrt{2(m_1 + m_2)\log(2)}$ *is fixed. Then*

$$\sup_{f \in \mathscr{F}_{p_1, p_2}^{m_1, m_2}} \left\{ \sum_I \left( \varepsilon^2 \left( \frac{\lambda_{I,\varepsilon}}{\varepsilon} + 1 \right) \varphi\left( \frac{\lambda_{I,\varepsilon}}{\varepsilon} \right) + \min\{\lambda_{I,\varepsilon}^2, \theta_I^2\} \right) \right\} = O(\varepsilon^{2\vartheta(m_1, m_2)}).$$

Let the $\lambda_{I,\varepsilon}$'s be chosen as in Lemma 2.2 and let

$$\hat{f}_{\varepsilon} = \sum \delta^{(\cdot)}\left(\tilde{\theta}_I, \lambda_{I,\varepsilon}\right) \mu_I. \tag{2.12}$$

Using Lemma 2.2 in conjunction with (2.10), we can immediately derive the following theorem, which, together with Theorem 2.1, tells us that $\hat{f}_{\varepsilon}$ is minimax in the class $\mathscr{F}_{p_1, p_2}^{m_1, m_2}$.

THEOREM 2.2. *If Assumption* 1 *is satisfied*, *then*

$$\sup_{f \in \mathscr{F}_{p_1, p_2}^{m_1, m_2}} \left\{ \mathbb{E}\|\hat{f}_{\varepsilon} - f\|^2 \right\} = O(\varepsilon^{2\vartheta(m_1, m_2)}).$$

Although this theorem provides an interesting theoretical result, it turns out to be of limited practical use. The proposed estimator $\hat{f}_{\varepsilon}$ requires an appropriate tuning of the thresholds $\lambda_{I,\varepsilon}$, which strongly depend on the unknown $m_1$ and $m_2$. Even if it would be possible to adapt these parameters in our idealized Gaussian white-noise model, it is often not obvious how to transfer such a procedure to other noise structures (i.e., with dependencies, non-Gaussianity) which occur in practically relevant estimation problems. One could try to find specific procedures for each particular case; however, it seems to be difficult to find a universal recipe.

An alternative approach that is much less dependent on prior knowledge of $m_1$ and $m_2$ is proposed in a series of papers by Donoho and Johnstone, also contained in Donoho, Johnstone, Kerkyacharian and Picard (1995). First, we analyze the analog of the tail-$n$-widths [see Donoho, Johnstone, Kerkyacharian and Picard (1995)] in our two-dimensional function classes.

LEMMA 2.3. *Suppose that Assumption* 1 *holds. Let* $\tilde{V}_J = \bigoplus_{j_1+j_2=J}(V_{j_1} \otimes V_{j_2})$. *Then*

$$\sup_{f \in \mathscr{F}^{m_1, m_2}_{p_1, p_2}} \left\{ \|f - \mathrm{Proj}_{\tilde{V}_J} f\|^2 \right\} = O\left(2^{-J\gamma(m_1, m_2, p_1, p_2)}\right),$$

*where*

$$\gamma(m_1, m_2, p_1, p_2) = \frac{2m_1m_2 + m_1 + m_2 - 2m_1/\tilde{p}_2 - 2m_2/\tilde{p}_1}{m_1 + m_2},$$

$$\tilde{p}_i = \min\{p_i, 2\}.$$

If we now choose $J_\varepsilon$ sufficiently large, we are able to obtain

$$\sum_{I:\, j_1+j_2 > J_\varepsilon} \theta_I^2 = O\left(\varepsilon^{2\vartheta(m_1, m_2)}\right), \tag{2.13}$$

that is, the truncation of the wavelet series does not affect the desired rate of the estimator. Define $\mathscr{K}_\varepsilon = \{I = (j_1, j_2, k_1, k_2) | j_1 + j_2 \le J_\varepsilon\}$. We consider the estimator

$$\hat{\hat{f}}_\varepsilon = \sum_{I \in \mathscr{K}_\varepsilon} \delta^{(\cdot)}\left(\tilde{\theta}_I, \lambda_\varepsilon\right)\mu_I, \tag{2.14}$$

where

$$\lambda_\varepsilon = \varepsilon\sqrt{2\log(\#\mathscr{K}_\varepsilon)}\,.$$

Using Lemmas 2.2 and 2.3 and (2.10), we obtain the following theorem.

THEOREM 2.3. *Suppose that Assumption* 1 *holds and* $2^{J_\varepsilon} = O(\varepsilon^{-\eta})$ *for any* $\eta < \infty$. *Then*

$$\sup_{f \in \mathscr{F}^{m_1, m_2}_{p_1, p_2}} \left\{ \mathbb{E}\|\hat{\hat{f}}_\varepsilon - f\|^2 \right\} = O\left(\left(\varepsilon^2 \log(1/\varepsilon)\right)^{\vartheta(m_1, m_2)}\right)$$

*over all* $(m_1, m_2, p_1, p_2)$ *satisfying* $2^{-J_\varepsilon\gamma(m_1, m_2, p_1, p_2)} \le \varepsilon^{2\vartheta(m_1, m_2)}$.

Hence, the estimator $\hat{\hat{f}}_\varepsilon$ is minimax up to a factor of $\log(1/\varepsilon)$ over a wide range of function classes.

In the rest of this section we will briefly examine the basis $\overline{\mathscr{B}}$ w.r.t. its capability of data compression in anisotropic Sobolev spaces. The following lemma states a result on the decay of the modulus of continuity $\Omega_\varepsilon$ for this basis.

LEMMA 2.4. *It holds that*

$$\Omega_\varepsilon\Big(\overline{\mathscr{B}}, \mathscr{F}_{p_1, p_2}^{m_1, m_2}\Big) \asymp \varepsilon^{2\overline{\vartheta}(m_1, m_2)},$$

*where*

$$\overline{\vartheta}(m_1, m_2) = \min\left\{\frac{m_1}{m_1+1}, \frac{m_2}{m_2+1}\right\}.$$

It can be easily shown that $\overline{\vartheta}(m_1, m_2) = \vartheta(m_1, m_2)$ if $m_1 = m_2$ and $\overline{\vartheta}(m_1, m_2) < \vartheta(m_1, m_2)$ if $m_1 \neq m_2$. The rate $\overline{\vartheta}(m_1, m_2)$ is the usual one for a two-dimensional estimation problem in isotropic smoothness classes with degree of smoothness $m_1 \wedge m_2$. To give a simple explanation as to why the tensor product basis $\mathscr{B}$ gives a better data compression than the one-scale multiresolution basis $\overline{\mathscr{B}}$, consider the extreme case of a function $f(x_1, x_2) = f(x_1)$, which is constant in the $x_2$-direction. Assume we wish to get a certain accuracy of approximation by certain subsets of $\mathscr{B}$ and $\overline{\mathscr{B}}$, respectively. In $\mathscr{B}$, we achieve this by the set of basis functions

$$\bigcup_{l-1 \le j_1 \le J} \big\{\psi_{j_1 k_1}(x_1)\,\psi_{l-1, k_2}(x_2)\big\}_{k_1, k_2},$$

where $J$ depends on the desired degree of approximation. This set has a cardinality of $O(2^J)$. In contrast, we obtain the same accuracy of approximation in $\overline{\mathscr{B}}$ by the set of functions

$$\big\{\phi_{lk_1}(x_1)\,\phi_{lk_2}(x_2)\big\}_{k_1, k_2} \cup \bigcup_{l \le j \le J} \big\{\psi_{jk_1}(x_1)\,\phi_{jk_2}(x_2)\big\}_{k_1, k_2},$$

which has a cardinality of order $O(2^{2J})$. In other words, basis $\overline{\mathscr{B}}$ is not really able to adapt coordinatewise to the right degree of resolution.

We have already seen that basis $\mathscr{B}$ provides an optimal data compression in the sense that $\tilde{\Omega}_\varepsilon(\mathscr{B}, \mathscr{F}_{p_1, p_2}^{m_1, m_2})$ decays at the same rate as the minimax risk in $\mathscr{F}_{p_1, p_2}^{m_1, m_2}$. To make a comparison between the two bases in statistical terms, we restrict our consideration to thresholded diagonal estimators in both cases. Let $\overline{\mathscr{B}} = \{\overline{\mu}_I\}$ and let $\overline{\theta}_I$ and $\tilde{\overline{\theta}}_I$ denote the corresponding true and empirical coefficients, respectively. By simple calculations we can show that

$$\inf_\lambda \left\{\mathbb{E}\Big(\delta^{(\cdot)}\Big(\tilde{\overline{\theta}}_I, \lambda\Big) - \overline{\theta}_I\Big)^2\right\} \ge C \min\big\{\varepsilon^2, \overline{\theta}_I^2\big\}. \tag{2.15}$$

Hence, we will get a lower bound for the risk of thresholded diagonal estimators simply by observing the rate of decay of $\Omega_\varepsilon$. The following theorem is an immediate consequence of Lemma 2.4 and (2.15).

THEOREM 2.4. *Suppose that Assumption* 1 *holds. Then*

$$\sup_{f \in \mathscr{F}^{m_1, m_2}_{p_1, p_2}} \inf_{(\lambda_I)} \left\{ \mathbb{E} \left\| \sum \delta^{(\cdot)}\left(\tilde{\bar{\theta}}_I, \lambda_I\right) \bar{\mu}_I - f \right\|^2 \right\} \geq C \varepsilon^{2\bar{\vartheta}(m_1, m_2)}.$$

Hence, we get that diagonal estimators based on basis $\overline{\mathscr{B}}$ are never better than those based on $\mathscr{B}$, and they are worse if $m_1 \neq m_2$. At this point we want to remark that there exists an attempt to construct higher-dimensional multiresolution bases for anisotropic smoothness classes. Berkolajko and Novikov (1992) obtained such a basis by properly connecting levels $j_1$ and $j_2$ in dependence on the relation between $m_1$ and $m_2$. However, as this approach depends strongly on the latter relation, it does not provide a *universal* basis which is optimal for a greater range of smoothness classes. The adaptive choice of an appropriate basis, which, in principle, seems to be possible in view of results by Donoho and Johnstone (1994b), would call for another step in the estimation process. Another approach was proposed by Donoho (1995), who uses CART methodology to choose the best *adapted* basis functions among all anisotropic Haar bases. The method based on the tensor product basis described here is certainly somewhat easier to apply in that we have a one-step procedure which is good enough for our purpose of obtaining (nearly) minimax results.

**3. Adaptive estimation of evolutionary spectra.** In this section we apply the two-dimensional smoothing method developed in the previous section to the particular problem of estimating the evolutionary spectrum. First, we introduce the model of local stationarity and provide an $L_2$-generalization of the definition of the evolutionary spectrum. Then we define a new version of a localized periodogram and propose the general estimation scheme. The fine tuning of the estimator, that is, the specification of the parameters involved in its definition, is given after a deeper study of the stochastic behavior of the empirical wavelet coefficients.

3.1. *The model of local stationarity.* To address the problem of adaptively estimating the time-dependent spectrum of a nonstationary time series, we start by citing the definition of a locally stationary process, as given in Dahlhaus (1993). Note that this generalizes the Cramér representation of a stationary stochastic process [see, e.g., Priestley (1981)].

DEFINITION 3.1. A sequence of stochastic processes $\{X_{t,T}\}_{t=1,\dots,T}$ is called *locally stationary* with transfer function $A^o$ and trend $\mu$ if there exists a representation

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} A^o_{t,T}(\omega) \exp(i\omega t)\, d\xi(\omega), \tag{3.1}$$

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, that is, $\operatorname{cov}(d\xi(\omega), d\xi(\omega')) = \delta(\omega - \omega')\,d\omega$, $\operatorname{cum}\{d\xi(\omega_1), \dots, d\xi(\omega_k)\} = \eta(\sum_{j=1}^{k}\omega_j)h_k(\omega_1, \dots, \omega_{k-1})\,d\omega_1 \dots d\omega_k$, where $\operatorname{cum}\{\dots\}$ denotes the cumulant of order $k$, $|h_k(\omega_1, \dots, \omega_{k-1})| \le \text{const}_k$ 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 $2\pi$ extension of the Dirac delta function;

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

$$\sup_{t,\,\omega}|A^o_{t,T}(\omega) - A(t/T, \omega)| \le KT^{-1}. \tag{3.2}$$

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

In this model the smoothness of $A$ in $u$ restricts the departure from stationarity and ensures the locally stationary behavior of the process (as will be discussed in the remarks below). The exact smoothness assumptions on $A(u, \omega)$ will be given in Assumptions 2 and 3 below. This model also allows us to define a *unique* underlying time-varying spectrum of $\{X_{t,T}\}$, as follows.

Consider first, for $u \in (0, 1)$,

$$f_T(u, \omega) = \frac{1}{2\pi}\sum_{s=-\infty}^{\infty}\operatorname{cov}\{X_{[uT-s/2],T}; X_{[uT+s/2],T}\}\exp(-i\omega s), \tag{3.3}$$

where the $X_{t,T}$'s are given by (3.1), with $A^o_{t,T}(\omega) = A(0, \omega)$ for $t < 1$ and $A^o_{t,T}(\omega) = A(1, \omega)$ for $t > T$.

This quantity, for fixed $T$, is similar to the so-called *Wigner–Ville spectrum* [see, e.g., Martin and Flandrin (1985)].

Then, by the following (3.4) and (3.5), $f_T(u, \omega)$ will be related to the smooth amplitude function $A(u, \omega)$, which defines the "evolutionary spectrum."

DEFINITION 3.2. As *evolutionary spectrum* of $\{X_{t,T}\}$ given in (3.1) we define, for $u \in (0, 1)$,

$$f(u, \omega) := |A(u, \omega)|^2. \tag{3.4}$$

This $f(u, \omega)$ is, in general, in some mean-square sense as shown below, the limit of $f_T(u, \omega)$ as $T \to \infty$.

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

$$\int_{-\pi}^{\pi}|f_T(u, \omega) - f(u, \omega)|^2\,d\omega = o(1). \tag{3.5}$$

Moreover, by Theorem 2.2 in a previous version of Dahlhaus (1993), if $A(u, \omega)$ is differentiable in $u$ and $\omega$ with uniformly bounded derivatives, then (3.5) holds pointwise in $u$ a.e. in $\omega$.

However, in order to be able to treat less regular evolutionary spectra, basically those defined by the smoothness assumptions (Assumptions 2 and 3) on $A(u, \psi)$ below, we will show in Theorem 3.1 that (3.5) continues to hold in some weaker $L_2(du, d\omega)$-sense on $[0, 1] \times [-\pi, \pi]$.

Before that, however, we give some comments on this class of locally stationary processes. For more details we refer to the discussion in the cited papers of Dahlhaus, in von Sachs and Schneider (1996) and Neumann and von Sachs (1995).

REMARK 3.1. The main purpose of this approach of rescaling in time is the modeling of nonstationary time series in a way which allows asymptotic inference on its second-order structure from a single realization. To achieve this, one has to bound the complexity of the spectrum as the object which defines the underlying model: without any rescaling (or related further assumptions), it is not possible to make statistical inference on a spectrum $f(t, \omega)$ which depends on a growing number of time points $t = 1, \dots, T$, even with $T$ tending to $\infty$. Rescaling, however, parallels nonparametric regression with an asymptotically denser and denser design: assuming smoothness, as $T \to \infty$ the complexity of the spectrum remains bounded, whereas the amount of statistical information increases by a growing number of observations.

Moreover, Definition 3.2, under the smoothness assumptions on $A(u, \omega)$, turns out to be unique [cf. the uniqueness of the Wigner–Ville spectrum, pointed out in Martin and Flandrin (1985), Section 2, B.7]. This is an inherent advantage of the locally stationary approach when trying to define what is meant by "*the* spectrum of a nonstationary process $X_1, \dots, X_T$ at a fixed time point $t_0$."

REMARK 3.2. A simple example for a locally stationary process is $X_{t,T} = \mu(t/T) + \sigma(t/T)Y_t$, with a stationary process $Y_t$ and smooth $\mu(u)$ and $\sigma(u)$ [see, also for more examples, Dahlhaus (1997)]. More generally, this class includes ARMA processes with time-varying coefficients, and, of course, if $A$ and $\mu$ do not depend on $t$ and $T$, ordinary stationary processes. More specifically, the representation (3.1) is based on a sequence of functions $A^o_{t,T}(\omega)$ instead of the smooth function $A(t/T, \omega)$ itself. For some simple examples, like time-dependent moving average processes, $A^o_{t,T}(\omega) = A(t/T, \omega)$. However, in general, the time-varying second-order structure of the process is only assumed to be *coupled* to some "asymptotically" smooth behavior. In particular, this is necessary to include the class of autoregressive processes with time-varying coefficients, as shown in Dahlhaus (1996a), Theorem 2.3.

In our work, for reasons of notational convenience, we do not want to adopt the more general definition based on the sequence $A^o_{t,T}(\omega)$, but rather restrict ourselves to the smooth function $A(t/T, \omega)$ directly. However, in view of the fast rate of approximation in (3.2), all results will continue to hold for the original class as in Definition 3.1.

Note that, as in Dahlhaus (1997) and von Sachs and Schneider (1996), for simplicity we assume that $\mu(u) = 0$; that is, we do not treat the problem of estimating the mean of the time series. In comparison to Dahlhaus (1996a, 1997), here, our smoothness assumptions on $A(u, \omega)$ are relaxed: basically, we like to impose minimal smoothness as being of bounded variation on $U \times \Pi := [0,1] \times [-\pi, \pi]$ (which is made precise in Assumption 2). For technical reasons, in order to facilitate the proofs, we impose an additional smoothness condition on the decay of the Fourier coefficients of $A(u, \omega)$ as a function of $\omega$, which implies continuity of $A$ in $\omega$.

DEFINITION 3.3 (Total variation on $U \times \Pi := [0,1] \times [-\pi, \pi]$).

$$TV_{U\times\Pi}(f) := \sup \sum_i \sum_j |f(u_i, \omega_j) - f(u_i, \omega_{j-1}) - f(u_{i-1}, \omega_j) + f(u_{i-1}, \omega_{j-1})|,$$

where the supremum is to be taken over all partitions of $U \times \Pi$.

Now we impose the following assumptions.

ASSUMPTION 2. (a) $A(u, \omega)$ has bounded total variation on $U \times \Pi$, that is, $TV_{U\times\Pi}(A) < \infty$;
(b) $\sup_u TV_{[-\pi,\pi]}(A(u, \cdot)) < \infty$ and $\sup_\omega TV_{[0,1]}(A(\cdot, \omega)) < \infty$;
(c) $\sup_{u,\omega}|A(u, \omega)| < \infty$;
(d) $\inf_{u,\omega}|A(u, \omega)| \geq \kappa$ for some $\kappa > 0$.

ASSUMPTION 3. Let $\hat{A}(u, s) := 1/(2\pi)\int A(u, \omega)\exp(i\omega s)\, d\omega$, $s \in \mathbb{Z}$, $u \in [0,1]$. Then $\sup_u \sum_s |\hat{A}(u, s)| < \infty$.

Note that these are somewhat minimal conditions, part of which might be fulfilled simply by restricting $A$ to be a member of the specific smoothness class under consideration (anisotropic Sobolev, Hölder, etc.). In our case of Sobolev restrictions Assumptions 2(b) and (c) and 3 are implications of the considered Sobolev smoothness. Now this minimal smoothness of $A$ is sufficient to ensure the locally stationary behavior of the process with a spectrum which is uniquely defined in the $L_2(du, d\omega)$-sense on $U \times \Pi$. It is precisely this general definition which is needed to embed our object of interest into the smoothness class of functions for our kind of estimation problem, as considered in Section 2. The following theorem shows that (3.5) continues to hold in this mean-square sense.

THEOREM 3.1. *Under Assumptions* 2 *and* 3,

$$\lim_{T\to\infty} \int_0^1 \int_{-\pi}^{\pi} |f_T(u, \omega) - f(u, \omega)|^2 \, d\omega \, du = 0.$$

For the particular context of our work, we now restrict our consideration to the anisotropic Sobolev class as introduced in Section 2; that is, we assume that $f$ is a member of this class by assuming that $A(u, \omega)$ is:

$$A \in \mathscr{F}_{p_1, p_2}^{m_1, m_2}(C) \quad \text{with } m_i \geq 1,\ p_i \geq 1 \text{ and } m_i > 1/p_i. \tag{3.6}$$

We note that with (3.6) $f$ is in any $L_p(U \times \Pi)$-space (due to the continuity in each argument), that is, in particular, in $L_2$.

3.2. *The new adaptive estimator*. Now we turn to the problem of estimating the evolutionary spectrum $f$. The first step in our inference about $f$ is to transfer the information $\{X_{1,T}, \dots, X_{T,T}\}$ given in the time domain to the time–frequency domain. One possibility, as chosen by Dahlhaus (1997) and also in von Sachs and Schneider (1996), is to consider a localized periodogram, localized by introducing segments of length $N = N(T)$, where $N \to \infty$ and $N/T \to 0$ as $T \to \infty$. One problem with this approach is that the segment length $N$ is an additional parameter, whose optimal choice depends on the relation between the smoothness in time and frequency direction. Here we intend to develop a fully adaptive approach: by wavelet thresholding the procedure should be able to automatically adapt to the right degree of resolution in both time and frequency direction. Note that these are, of course, not independent, but stand in a reciprocal relationship due to the uncertainty principle: the more accurately we try to estimate $f(u, \omega)$ in the time direction, the less accurately can we estimate it in the frequency direction and vice versa; cf. Priestley (1981), page 835.

To this end, by a straightforward analogy to (3.3), we introduce a periodogram-like statistic $I_{t,T}$, $1 \leq t \leq T$, which is different from the localized periodogram of von Sachs and Schneider (1996) and which we like to call a "pre-periodogram:"

$$I_{t,T}(\omega) = \frac{1}{2\pi} \sum_{s:\, 1 \leq [t-s/2],\, [t+s/2] \leq T} X_{[t-s/2],T} X_{[t+s/2],T} \exp(-i\omega s). \tag{3.7}$$

Note that $I_{t,T}$ can be considered as a preliminary "estimate" which is even more fluctuating than the classical periodogram is. However, its expected value (asymptotically) equals the evolutionary spectrum $f(t/T, \omega)$, and for each fixed $T$ its expectation coincides with a Wigner–Ville spectrum.

In von Sachs and Schneider (1996) part of the localization was delivered by summation over certain time points in segments of chosen length $N$ before the actual local smoothing was performed by wavelet thresholding. Thus, inherently a lower bound was fixed for the resolution in time which obviously had consequences also for the performance in the frequency direction: the larger $N$ the worse is the time resolution, but the better can low-frequency components be detected and vice versa. Here, in our new approach, we avoid a twofold smoothing: projection of these pre-periodograms $I_{t,T}$ on an appropriate wavelet basis will do the whole task of adaptive local smoothing!

To give the link to the previous section on anisotropic smoothness classes, with this particular task, we are confronted with a two-dimensional estimation problem, where the axes have a special meaning, time and frequency, respectively. It seems reasonable to design the estimation method in such a way that it takes different degrees of smoothness in these two directions into account.

As we have seen in the preceding section, we obtain an appropriate wavelet basis according to the definition of the basis $\mathscr{B} = \{\mu_I(u, \omega)\}$. We get such a basis as a tensor product of two bases, where in the time direction we choose a wavelet basis on the interval $\{\phi_{lk}\}_k \cup \{\psi_{jk}\}_{j \geq l, k}$ [e.g., boundary-corrected Meyer wavelets; see Meyer (1991) or those of Cohen, Daubechies and Vial (1993)]. In the frequency direction a periodic basis $\{\tilde{\phi}_{lk}\}_k \cup \{\tilde{\psi}_{jk}\}_{j \geq l, k}$ is used [as proposed in Daubechies (1992), Chapter 9.3]. As an example, we like to mention the orthogonal periodized Battle–Lemarié spline wavelets [as in von Sachs and Schneider (1996)], though these have "numerical compact support," only, but our proofs will only slightly change with these. For notational convenience we write again $\psi_{l-1,k}$ and $\tilde{\psi}_{l-1,k}$ for $\phi_{lk}$ and $\tilde{\phi}_{lk}$, respectively. Whenever it is not misleading, we use the multiindex $I = (j_1, j_2, k_1, k_2)$. In order to derive our asymptotic results, we have to impose the following conditions on the chosen wavelet bases.

Assumption 4. (a) $\phi(u)$, $\psi(u)$, $\tilde{\phi}(\omega)$ and $\tilde{\psi}(\omega)$ are real functions of bounded total variation on $[0, 1]$ and $[-\pi, \pi]$, respectively.

(b) Further, $\sum_s |\hat{\tilde{\phi}}(s)| < \infty$ and $\sum_s |\hat{\tilde{\psi}}(s)| < \infty$.

In addition to the "true" wavelet coefficients $\theta_I$ of $f(u, \omega)$:

$$\begin{aligned} \theta_I &= \int_{U \times \Pi} f(u, \omega) \mu_I(u, \omega)\, du\, d\omega \\ &= \int_{U \times \Pi} f(u, \omega) \psi_{j_1 k_1}(u) \tilde{\psi}_{j_2 k_2}(\omega)\, du\, d\omega, \end{aligned} \tag{3.8}$$

we define empirical wavelet coefficients as follows:

$$\tilde{\theta}_I = \sum_{t=1}^{T} \int_{(t-1)/T}^{t/T} \psi_{j_1 k_1}(u)\, du \int_{-\pi}^{\pi} \tilde{\psi}_{j_2 k_2}(\omega) I_{t,T}(\omega)\, d\omega. \tag{3.9}$$

In the special case of a stationary time series, the advantage of the tensor product basis over the multiresolution basis becomes apparent. Then all coefficients $\theta_I$ with $j_1 \neq l-1$ are equal to 0, whereas $\theta_{(l-1, j_2, k_1, k_2)} \asymp 2^{-l/2} \theta_{j_2 k_2}$, where $\theta_{j_2 k_2} = \int f(\omega) \tilde{\psi}_{j_2 k_2}(\omega)\, d\omega$ are the wavelet coefficients of the (one-dimensional) spectral density $f(\omega) = f(u, \omega)$. In view of the results from Section 2, it is obvious that in estimating $f(u, \omega)$, which is constant in $u$, we can obtain the same rate as in Neumann (1994) in the stationary case.

Let $\mathscr{I}_T$ be a set of indices specified in the next section in (3.11) and

$$\mathscr{I}_T^0 = \{I \in \mathscr{I}_T | (j_1, j_2) \neq (l-1, l-1)\}.$$

We consider the estimator

$$\hat{f}(u, \omega) = \sum_{I:(j_1, j_2)=(l-1,l-1)} \tilde{\theta}_I \mu_I(u, \omega) + \sum_{I \in \mathscr{I}_T^0} \delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,T}\big)\mu_I(u, \omega), \tag{3.10}$$

where the thresholds $\lambda_{I,T}$ are also specified below. As usually done, we do not shrink the coefficients from the coarsest level $(j_1, j_2) = (l-1, l-1)$. This seems to be reasonable in view of Assumption 2(d), which implies that the spectrum is bounded away from 0.

3.3. *Asymptotic properties of empirical wavelet coefficients*. In the following we intend to derive asymptotic normality of the empirical coefficients by the method of cumulants. It turns out that a simple central limit theorem would not be sufficient for proving risk equivalence between our thresholded wavelet estimator and the case of Gaussian noise. In view of quite a large number of coefficients which cannot be a priori neglected in cases of "inhomogeneous smoothness," we have to choose the threshold somewhat higher than the noise level, that is, of larger order than the standard deviation of the empirical coefficients. Accordingly, we need some formulation of asymptotic normality, which puts special emphasis on moderate and large deviations.

In contrast to a central limit theorem, where it would be sufficient to show that $\mathrm{cum}_n(\tilde{\theta}_I/\sqrt{\mathrm{var}(\tilde{\theta}_I)}) = o(1)$ holds for each particular $n \geq 3$, here we need a stronger estimate for the higher-order cumulants. For the reader's convenience we quote a lemma from Neumann (1994), which holds under the following assumption and provides appropriate estimates for general quadratic forms.

ASSUMPTION 5.

$$\sup_{1 \le t_1 \le T} \left\{ \sum_{t_2, \dots, t_k = 1}^{T} \left|\mathrm{cum}\big(X_{t_1,T}, \dots, X_{t_k,T}\big)\right| \right\} \le C^k (k!)^{1+\gamma}$$

for all $k = 2, 3, \dots,$ where $\gamma \geq 0$.

LEMMA 3.1. *Let*

$$\eta_T = \mathbf{X}' M \mathbf{X},$$

*where*

$$\mathbf{X} = (X_{1,T}, \dots, X_{T,T})', \qquad M = ((M_{st}))_{s,t=1,\dots,T}, \qquad M_{st} = M_{ts}.$$

*Further, let*

$$\xi_T = \mathbf{Y}' M \mathbf{Y},$$

*where* $\mathbf{Y} = (Y_1, \dots, Y_T)' \sim N(0, \mathrm{Cov}(\mathbf{X}))$. *Then, under Assumption* 5,

$$\mathrm{cum}_n(\eta_T) = \mathrm{cum}_n(\xi_T) + R_n$$

*holds for* $n \geq 2$, *where*:

(i) $|\mathrm{cum}_n(\xi_T)| \leq \mathrm{var}(\xi_T) 2^{n-2}(n-1)! [\lambda_{\max}(M)\lambda_{\max}(\mathrm{Cov}(\mathbf{X}))]^{n-2}$,

(ii) $R_n \leq 2^{n-2} C^{2n}((2n)!)^{1+\gamma} \max_{s,t}\{|M_{st}|\} \tilde{M} \|M\|_\infty^{n-2}$,

$$\tilde{M} = \sum_s \max_t \{|M_{st}|\}, \qquad \|M\|_\infty = \max_s \left\{ \sum_t |M_{st}| \right\}.$$

In the following we are able to show asymptotic normality for all coefficients $\tilde{\theta}_I$ with $2^{j_1+j_2} = o(T)$ and $j_2^{-1} = o(1)$.

Fix some $\delta > 0$. We define

$$\mathscr{I}_T = \{ I | 2^{j_1+j_2} \leq T^{1-\delta} \}. \tag{3.11}$$

Making use of Lemma 3.1, we obtain the following result for the empirical coefficients.

LEMMA 3.2. *Suppose that Assumptions* 1–5 *hold. Then*:

(i) $\mathbb{E}\tilde{\theta}_I = \theta_I + o(T^{-1/2})$,

(ii) $$\mathrm{var}(\tilde{\theta}_I) = 2\pi T^{-1} \int_{U\times\Pi} \{f(u,\omega)\psi_{j_1k_1}(u)\}^2 \, du \, \tilde{\psi}_{j_2k_2}(\omega) \times [\tilde{\psi}_{j_2k_2}(\omega) + \tilde{\psi}_{j_2k_2}(-\omega)] \, d\omega + o(T^{-1}) + O(2^{-j_2}T^{-1}),$$

(iii) $\left|\mathrm{cum}_n(\tilde{\theta}_I)\right| \leq (n!)^{2+2\gamma} C^n T^{-1} (T^{-1} 2^{(j_1+j_2)/2} \log(T))^{n-2}$

*for* $n \geq 3$ *and appropriate* $C > 0$ *uniformly in* $I \in \mathscr{I}_T$.

Note that the last term in the remainder of $\mathrm{var}(\tilde{\theta}_I)$ includes the term composed of fourth-order cumulants of the time series data, and, as usual in nonparametric spectrum estimation, it is asymptotically negligible if the spectral window shrinks asymptotically, that is, if $j_2 \to \infty$.

By (ii) and (iii) of this lemma we can show that the relation

$$\left|\mathrm{cum}_n\left((\tilde{\theta}_I - \theta_I)/\sqrt{\mathrm{var}(\tilde{\theta}_I)}\right)\right| \leq (n!)^{2+2\gamma}(CT)^{-\mu(n-2)}$$

holds for an appropriate choice of $C$ and $\mu > 0$, where the constant $C$ is uniform in $n \geq 3$, $I \in \mathscr{I}_T \cap \{I | 2^{j_2} \geq T^\rho\}$ for any $\rho > 0$.

Using Lemma 1 in Rudzkis, Saulis and Statulevicius (1978), we now obtain the desired version of asymptotic normality. Let $\sigma_I^2$ denote the variance of $\tilde{\theta}_I$.

PROPOSITION 3.1. *Suppose that Assumptions* 1–5 *hold. Let* $\Delta_T = C\sqrt{\log T}$ *for any fixed* $C < \infty$. *Then both*

$$P\Big(\big(\tilde{\theta}_I - \theta_I\big)/\sigma_I \geq x\Big) = (1 - \Phi(x))(1 + o(1))$$

*and*

$$P\Big(-\big(\tilde{\theta}_I - \theta_I\big)/\sigma_I \geq x\Big) = (1 - \Phi(x))(1 + o(1))$$

*hold uniformly in* $-\infty \leq x \leq \Delta_T$ *and* $I \in \mathscr{I}_T \cap \{I | 2^{j_2} \geq T^{\rho}\}$ *for* $\rho > 0$ *arbitrarily small, where* $\Phi(x) = \int_{-\infty}^{x} \varphi(t)\,dt$ *denotes the standard normal cumulative distribution function.*

In order to establish the equivalence to the case of Gaussian noise, we consider the following approximating model for our empirical coefficients:

$$\tilde{\xi}_I = \theta_I + \sigma_I \varepsilon_I, \qquad I \in \mathscr{I}_T,$$

where $\varepsilon_I \sim N(0,1)$.

Essentially by integration by parts, due to Proposition 3.1 we obtain the following assertion.

PROPOSITION 3.2. *Suppose that Assumptions* 1–5 *hold. Then, for arbitrary nonrandom thresholds* $\lambda_{I,T} = O(T^{-1/2}\sqrt{\log(T)}\,)$,

$$\sum_{I \in \mathscr{I}_T} \mathbb{E}\Big(\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,T}\big) - \theta_I\Big)^2 = (1 + o(1)) \sum_{I \in \mathscr{I}_T} \mathbb{E}\Big(\delta^{(\cdot)}\big(\tilde{\xi}_I, \lambda_{I,T}\big) - \theta_I\Big)^2 + O\big(T^{-\vartheta(m_1, m_2)}\big).$$

This asymptotic risk equivalence enables us to derive the following theorem. Recall that $\vartheta(m_1, m_2)$ was defined in Theorem 2.1 and $\gamma(m_1, m_2, p_1, p_2)$ in Lemma 2.3, respectively.

THEOREM 3.2. *Suppose that Assumptions* 1–5 *hold and* $(1 - \delta)\gamma(m_1, m_2, p_1, p_2) \geq \vartheta(m_1, m_2)$ *with* $\delta$ *as in* (3.11). *Further, assume that, for some* $\gamma_T \to 1$ *and appropriate* $C$,

$$\gamma_T \sigma_I \sqrt{2\log\big(\#\mathscr{I}_T^0\big)} \leq \lambda_{I,T} \leq C T^{-1/2} \sqrt{\log(T)}$$

*holds for* $I \in \mathscr{I}_T^*$, $\mathscr{I}_T^* \subseteq \mathscr{I}_T^0$, *where* $\#(\mathscr{I}_T^0 \setminus \mathscr{I}_T^*) = O(T^{1-\vartheta(m_1, m_2)})$. *Then*

$$\mathbb{E}\|\hat{f} - f\|^2_{L_2(U \times \Pi)} = O\Big(\big(\log(T)/T\big)^{\vartheta(m_1, m_2)}\Big).$$

Note that this rate for the estimator $\hat{f}$ corresponds to the rate derived in Theorem 2.3 if we set $\varepsilon^2 = T^{-1}$.

There are many possibilities for $m_1$, $m_2$, $p_1$ and $p_2$ to fulfill $\gamma(m_1, m_2, p_1, p_2) > \vartheta(m_1, m_2)$, for example, if $m_i \geq 2/p_i$. Then we can find some sufficiently small $\delta > 0$, such that the assumption of Theorem 3.2 is satisfied. Hence, our estimator is simultaneously nearly optimal over a wide range of smoothness classes.

Although Theorem 3.2 is of certain theoretical interest, it is not very helpful for practical purposes, because the definition of the estimator $\hat{f}$ depends on the unknown quantities $\sigma_I$. It is a natural ideal to use some initial estimates of them to construct a fully adaptive procedure. Let $\hat{\sigma}_I$ be any estimate of $\sigma_I$. With the thresholds $\hat{\lambda}_I = \hat{\sigma}_I \sqrt{2\log(\#\mathcal{I}_T^0)}$, we define the following estimator, which is a fully adaptive version of (3.10):

$$\hat{\hat{f}}(u,\omega) = \sum_{I:\,(j_1,j_2)=(l-1,l-1)} \tilde{\theta}_I \mu_I(u,\omega) + \sum_{I\in\mathcal{I}_T^0} \delta^{(\cdot)}\big(\tilde{\theta}_I, \hat{\lambda}_I\big)\mu_I(u,\omega). \tag{3.12}$$

The next proposition provides a sufficient condition, slightly stronger than $\sqrt{T}$-consistency, for the estimates $\hat{\sigma}_I$, which ensures the desired rate of convergence for $\hat{\hat{f}}$.

Proposition 3.3. *Suppose that Assumptions* 1–5 *hold. Let* $(1-\delta)\gamma(m_1, m_2, p_1, p_2) \ge \vartheta(m_1, m_2)$. *Assume that, for any sequence* $\{c_T\}$ *tending to* 0 *as* $T \to \infty$, *the relation*

$$P\big(|\hat{\sigma}_I - \sigma_I| > c_T T^{-1/2}\big) = O(T^{-4}) \tag{3.13}$$

*holds in a uniform manner in* $I \in \mathcal{I}_T^*$, *where* $\#(\mathcal{I}_T^0 \setminus \mathcal{I}_T^*) = O(T^{1-\vartheta(m_1,m_2)})$. *Then*

$$\mathbb{E}\|\hat{\hat{f}} - f\|^2_{L_2(U\times\Pi)} = O\Big(\big(\log(T)/T\big)^{\vartheta(m_1,m_2)}\Big).$$

This proposition is an immediate consequence of Theorem 3.3 formulated below. In practice, one could, for example, plug some bivariate kernel estimator $\tilde{f}$ of $f$ based on $I_{t,T}$ into the asymptotic formula for the variance of the empirical coefficients, which is given in Lemma 3.2(ii). It turns out that (3.13) will be satisfied under weak assumptions on the time series and the estimator $\tilde{f}$, as indicated by related calculations in Neumann (1994), Section 6, for the case of stationary time series.

Some preliminary numerical experiments seem to indicate the stability of our approach in practice. Some further work necessary to explore a fully automatic parameter choice is to be reported elsewhere.

For the sake of greater generality, we formulate the following theorem, which also allows for thresholding schemes being different to that described above. Compared to Proposition 3.3, the conditions on the thresholds are more general, and can be viewed as some kind of necessary conditions which still ensure the desired rate of convergence.

Theorem 3.3. *Suppose that Assumptions* 1–5 *hold. Let* $(1-\delta)\gamma(m_1, m_2, p_1, p_2) \ge \vartheta(m_1, m_2)$. *Assume that the thresholds in* (3.12) *satisfy, for any* $\gamma_T \to 1$ *and appropriate* $C < \infty$,

$$\sum_{I\in\mathcal{I}_T^0} \mathbb{E}\big(\tilde{\theta}_I^2 + 1\big) I\Big(\hat{\lambda}_I \notin \Big[\gamma_T \sigma_I \sqrt{2\log(\#\mathcal{I}_T^0)}, C T^{-1/2}\sqrt{\log(T)}\Big]\Big) = O(T^{-\vartheta(m_1,m_2)}). \tag{3.14}$$

*Then*

$$\mathbb{E}\|\hat{\hat{f}} - f\|^2_{L_2(U\times\Pi)} = O\big((\log(T)/T)^{\vartheta(m_1, m_2)}\big).$$

REMARK 3.3. (i) We remark that if the assumptions of Theorem 3.2, Proposition 3.3 and Theorem 3.3 are to hold uniformly, then all assertions will hold uniformly in the class $\mathscr{F}^{m_1, m_2}_{p_1, p_2}$.

(ii) An obvious alternative to our nonlinear wavelet estimator would be a bivariate kernel estimator based on the pre-periodogram $I_{t,T}(\omega)$. For "homogeneous smoothness classes," that is, $p_i \ge 2$, we can expect the same rates of convergence as for our wavelet estimator. This has been derived, for example, by Dahlhaus (1996b) for the situation of a segmented periodogram estimator. The motivation for choosing the slightly more involved wavelet approach was twofold: first, we are able to treat less homogeneous smoothness classes appropriately, which mimic, for example, situations with a sudden change in the covariance structure of a nonstationary time series. Second, our wavelet threshold approach provides immediately an automatic choice of the right degree of smoothing in both directions, and spares us the technically more challenging task of choosing two bandwidths of a kernel estimator in a data-dependent way.

## 4. Proofs.

PROOF OF LEMMA 2.1. Let $j_1^*$ and $j_2^*$ be chosen such that

$$2^{j_1^*} \le C_0 \varepsilon^{-2m_2/(2m_1m_2+m_1+m_2)} < 2^{j_1^*+1},$$

$$2^{j_2^*} \le C_0 \varepsilon^{-2m_1/(2m_1m_2+m_1+m_2)} < 2^{j_2^*+1}$$

hold for some $C_0$ chosen at the end of this proof. Define

$$\mathscr{I}_\varepsilon = \{I = (j_1, j_2, k_1, k_2) | (j_1, j_2) = (j_1^*, j_2^*)\}.$$

It is obvious that $\mathscr{I}_\varepsilon$ satisfies

$$\#\mathscr{I}_\varepsilon \asymp 2^{j_1^*+j_2^*} \asymp \varepsilon^{-2(m_1+m_2)/(2m_1m_2+m_1+m_2)}.$$

It remains to show that, for an appropriate choice of $C_0$, the relation $\Theta_\varepsilon \subseteq \Theta$ holds. Let $f = \sum \theta_I \mu_I$ be arbitrary with $(\theta_I) \in \Theta_\varepsilon$. Then we obtain

$$(4.1) \qquad \|f\|_{p_i} \le \|f\|_\infty \le C\varepsilon 2^{(j_1^*+j_2^*)/2} \le CC_0 \varepsilon^{2m_1m_2/(2m_1m_2+m_1+m_2)}$$

and

$$(4.2) \qquad \left\| \frac{\partial^{m_i}}{\partial x_i^{m_i}} f \right\|_{p_i} \le \left\| \frac{\partial^{m_i}}{\partial x_i^{m_i}} f \right\|_\infty \le C\varepsilon 2^{j_i^* m_i} 2^{(j_1^*+j_2^*)/2} \le CC_0.$$

For $C_0$ small enough we obtain $f \in \mathscr{F}^{m_1, m_2}_{p_1, p_2}$, which implies $\Theta_\varepsilon \subseteq \Theta$. □

PROOF OF LEMMA 2.2. Let $\mathscr{I}_\varepsilon$ be chosen as in the proof of Lemma 2.1 and let

$$\theta_I^* = \begin{cases} \varepsilon, & \text{if } I \in \mathscr{I}_\varepsilon, \\ 0, & \text{otherwise.} \end{cases}$$

We have seen in the proof of Lemma 2.1 that $(\theta_I^*) \in \Theta$ holds, which implies

$$\Omega_\varepsilon\left(\mathscr{B}, \mathscr{F}_{p_1, p_2}^{m_1, m_2}\right) \geq \varepsilon^2 \# \mathscr{I}_\varepsilon \geq C \varepsilon^{2\vartheta(m_1, m_2)}.$$

Since $\Omega_\varepsilon(\mathscr{B}, \mathscr{F}_{p_1, p_2}^{m_1, m_2}) \leq C \tilde{\Omega}_\varepsilon(\mathscr{B}, \mathscr{F}_{p_1, p_2}^{m_1, m_2})$ for some appropriate constant $C$, we have a lower bound for $\tilde{\Omega}_\varepsilon(\mathscr{B}, \mathscr{F}_{p_1, p_2}^{m_1, m_2})$.

Now let $f \in \mathscr{F}_{p_1, p_2}^{m_1, m_2}$ be arbitrary. Let $j_1 \geq l$ and $x_{(j_1, k_1)} \in \operatorname{supp}(\psi_{j_1 k_1})$. Then, by Taylor's formula, which holds for $L_p$-Sobolev functions due to integration by parts,

$$\begin{aligned}
&\int \psi_{j_1 k_1}(x_1) f(x_1, x_2)\, dx_1 \\
&\quad = \int \psi_{j_1 k_1}(x_1) \left[ \int_{x_{(j_1, k_1)}}^{x_1} \frac{(x_1 - z)^{m_1 - 1}}{(m_1 - 1)!} \frac{\partial^{m_1}}{\partial x_1^{m_1}} f(z, x_2)\, dz \right] dx_1 \\
&\quad = O(2^{-j_1(m_1 - 1)}) \int |\psi_{j_1 k_1}(x_1)|\, dx_1 \int_{\operatorname{supp}(\psi_{j_1 k_1})} \left| \frac{\partial^{m_1}}{\partial x_1^{m_1}} f(z, x_2) \right| dz \\
&\quad = O(2^{-j_1(m_1 - 1/2)}) \int_{\operatorname{supp}(\psi_{j_1 k_1})} \left| \frac{\partial^{m_1}}{\partial x_1^{m_1}} f(z, x_2) \right| dz,
\end{aligned}$$

which implies, since every basis function $\mu_I$ overlaps only with a finite number of basis functions from the same scale $(j_1, j_2)$, that

$$\begin{aligned}
\sum_{k_1, k_2} \left| \theta_{(j_1 j_2 k_1 k_2)} \right|^p &= \sum_{k_1, k_2} \left| \int \psi_{j_2 k_2}(x_2) \int \psi_{j_1 k_1}(x_1) f(x_1, x_2)\, dx_1\, dx_2 \right|^p \\
&\leq C \sum_{k_1, k_2} 2^{-j_1(m_1 - 1/2)p} 2^{j_2 p/2} \\
&\qquad \times \left[ \iint_{\operatorname{supp}(\mu_{(j_1 j_2 k_1 k_2)})} \left| \frac{\partial^{m_1}}{\partial x_1^{m_1}} f(x_1, x_2) \right| dx_1\, dx_2 \right]^p \\
&\leq C \sum_{k_1, k_2} 2^{-j_1 m_1 p} 2^{(j_1 + j_2)p/2} \\
&\qquad \times \iint_{\operatorname{supp}(\mu_{(j_1 j_2 k_1 k_2)})} \left| \frac{\partial^{m_1}}{\partial x_1^{m_1}} f(x_1, x_2) \right|^p dx_1\, dx_2 \\
&\qquad \times \left( \operatorname{mes}\left( \operatorname{supp}(\mu_{(j_1 j_2 k_1 k_2)}) \right) \right)^{p-1} \\
&\leq C 2^{-j_1 m_1 p} 2^{(j_1 + j_2)(1 - p/2)} \left\| \frac{\partial^{m_1}}{\partial x_1^{m_1}} f \right\|_{p_1}^p \\
&= O(2^{-j_1 m_1 p} 2^{(j_1 + j_2)(1 - p/2)})
\end{aligned} \tag{4.3}$$

for all $p \le p_1$. By analogous calculations we can show that

$$\sum_{k_1,k_2} |\theta_{(j_1 j_2 k_1 k_2)}|^p = O(2^{-j_2 m_2 p} 2^{(j_1+j_2)(1-p/2)}) \tag{4.4}$$

holds for $j_2 \ge l$, $p \le p_2$.

Let $j_1^*$ and $j_2^*$ be such that $2^{j_1^*} = \varepsilon^{-2/(2m_1+1+m_1/m_2)}$ and $2^{j_2^*} = \varepsilon^{-2/(2m_2+1+m_2/m_1)}$. We decompose the set $\mathscr{J} = \{(j_1, j_2) | j_1 \ge l,\ j_2 \ge l\}$ into the following three sets:

$$\mathscr{J}_1 = \{(j_1, j_2) \in \mathscr{J} | j_1 \le j_1^* \text{ and } j_2 \le j_2^*\},$$
$$\mathscr{J}_2 = \{(j_1, j_2) \in \mathscr{J} | j_1 m_1 \le j_2 m_2 \text{ and } j_2 > j_2^*\},$$
$$\mathscr{J}_3 = \{(j_1, j_2) \in \mathscr{J} | j_1 m_1 > j_2 m_2 \text{ and } j_1 > j_1^*\}.$$

Then

$$\begin{aligned}\sum_{(j_1,j_2)\in\mathscr{J}_1} \sum_{k_1,k_2} \varepsilon^2 \left(\frac{\lambda_{I,\varepsilon}}{\varepsilon} + 1\right) \varphi\left(\frac{\lambda_{I,\varepsilon}}{\varepsilon}\right) + \min\{\lambda_{I,\varepsilon}^2, \theta_I^2\} &= O(\varepsilon^2 2^{j_1^*+j_2^*}) \\ &= O(\varepsilon^{2\vartheta(m_1,m_2)}).\end{aligned} \tag{4.5}$$

Further

$$\begin{aligned}&\sum_{(j_1,j_2)\in\mathscr{J}_2} \sum_{k_1,k_2} \varepsilon^2 \left(\frac{\lambda_{I,\varepsilon}}{\varepsilon} + 1\right) \varphi\left(\frac{\lambda_{I,\varepsilon}}{\varepsilon}\right) \\ &\quad = \sum_{j_2 > j_2^*} \sum_{j_1 : j_1 m_1 \le j_2 m_2} O\left(2^{j_1+j_2} \varepsilon^2 \sqrt{j_2 - j_2^*} \exp\left\{-\frac{K_{m_1,m_2}^2 (j_2 - j_2^*)}{2m_1}\right\}\right) \\ &\quad = \varepsilon^2 2^{j_2^*(m_1+m_2)/m_1} \sum_{j_2 > j_2^*} O\Big(\exp\Big\{\big(\log(2)(m_1+m_2) - K_{m_1,m_2}^2/2\big) \\ &\qquad\qquad \times (j_2 - j_2^*)/m_1\Big\} \sqrt{j_2 - j_2^*}\Big) \\ &\quad = O(\varepsilon^2 2^{j_2^*(m_1+m_2)/m_1}) = O(\varepsilon^{2\vartheta(m_1,m_2)}).\end{aligned} \tag{4.6}$$

Here the next-to-the-last equality follows due to the convergence of the geometric series. Let $(j_1, j_2) \in \mathscr{J}_2$ be fixed. We choose $p = 1$ if $p_2 = 1$ or $1 < p < 2$, $p \le p_2$ if $p_2 > 1$. By (4.4) we obtain

$$\#\{(k_1,k_2) \| \theta_{(j_1 j_2 k_1 k_2)}| > \lambda_{I,\varepsilon}\} = O(\lambda_{I,\varepsilon}^{-p} 2^{-j_2 m_2 p} 2^{(j_1+j_2)(1-p/2)}),$$

which implies that

$$\begin{aligned}&\sum_{k_1,k_2} \min\{\lambda_{I,\varepsilon}^2, \theta_I^2\} \\ &\quad = \lambda_{I,\varepsilon}^2 \#\{(k_1,k_2) \| \theta_{(j_1 j_2 k_1 k_2)}| > \lambda_{I,\varepsilon}\} + \sum_{(k_1,k_2):\, |\theta_{(j_1 j_2 k_1 k_2)}| \le \lambda_{I,\varepsilon}} \theta_I^p \lambda_{I,\varepsilon}^{2-p} \\ &\quad = O(\lambda_{I,\varepsilon}^{2-p} 2^{-j_2 m_2 p} 2^{(j_1+j_2)(1-p/2)}) \\ &\quad = O\big(\varepsilon^{2-p} (j_2 - j_2^*)^{1-p/2} 2^{-j_2 m_2 p} 2^{(j_1+j_2)(1-p/2)}\big).\end{aligned}$$

By $m_2 > 1/p$ we obtain that $[m_1 m_2 + (m_1 + m_2)(p/2 - 1)] > 0$, which yields

$$
\begin{aligned}
&\sum_{(j_1, j_2) \in \mathscr{J}_2} \sum_{k_1, k_2} \min\{\lambda_{I,\varepsilon}^2, \theta_I^2\} \\
&\quad = \varepsilon^{2-p} 2^{-j_2^* m_2 p} \\
&\qquad \times \sum_{j_2 > j_2^*} \sum_{j_1 :\, j_1 m_1 \le j_2 m_2} O\Big( (j_2 - j_2^*)^{1-p/2} 2^{(j_2^* - j_2) m_2 p} 2^{(j_1 + j_2)(1 - p/2)} \Big) \\
&\quad = \varepsilon^{2-p} 2^{-j_2^* [m_2 - ((m_1 + m_2)/2m_1)] p} \\
&\qquad \times \sum_{j_2 > j_2^*} O\Big( (j_2 - j_2^*)^{1-p/2} 2^{(j_2^* - j_2)[m_1 m_2 + (m_1 + m_2)(p/2 - 1)]/m_1} \Big) \\
&\quad = O(\varepsilon^{2\vartheta(m_1, m_2)}).
\end{aligned} \tag{4.7}
$$

The sum over $\mathscr{J}_3$ can be treated analogously to (4.6) and (4.7), which completes the proof. □

PROOF OF LEMMA 2.3. It is easy to see that

$$
\big\| f - \mathrm{Proj}_{\tilde{V}_J} f \big\|^2 = \sum_{j_1 + j_2 > J - 2} \sum_{k_1, k_2} \theta_I^2 \le \sum_{(j_1, j_2) \in \mathscr{J}_4} \sum_{k_1, k_2} \theta_I^2 + \sum_{(j_1, j_2) \in \mathscr{J}_5} \sum_{k_1, k_2} \theta_I^2,
$$

where

$$
\begin{aligned}
\mathscr{J}_4 &= \{(j_1, j_2) | L_1 j_1 \ge L_2 j_2 \text{ and } j_1 > (J - 2) L_2/(L_1 + L_2)\}, \\
\mathscr{J}_5 &= \{(j_1, j_2) | L_1 j_1 < L_2 j_2 \text{ and } j_2 > (J - 2) L_1/(L_1 + L_2)\},
\end{aligned}
$$

with

$$
L_1 = m_1 - 1/\tilde{p}_1 + 1/\tilde{p}_2, \qquad L_2 = m_2 - 1/\tilde{p}_2 + 1/\tilde{p}_1.
$$

For the sake of a clear presentation, we introduce the following notation:

$$
\theta_{(\psi, j_1, k_1), (\phi, j_2, k_2)} = \iint \psi_{j_1 k_1}(x_1) \phi_{j_2 k_2}(x_2) f(x_1, x_2)\, dx_1\, dx_2. \tag{4.8}
$$

Now we get by Parseval's equality, Jensen's inequality and (4.3) that

$$
\begin{aligned}
\sum_{j_2 :\, j_2 \le j_1 L_1/L_2} \sum_{k_1, k_2} \theta_I^2 &= \big\| \mathrm{Proj}_{(W_{j_1} \otimes V_{[j_1 L_1/L_2 + 1]})} f \big\|^2 \\
&= \sum_{k_1, k_2} \theta^2_{(\psi, j_1, k_1), (\phi, [j_1 L_2/L_2 + 1], k_2)} \\
&\le \Bigg( \sum_{k_1, k_2} \big| \theta_{(\psi, j_1, k_1), (\phi, [j_1 L_1/L_2 + 1], k_2)} \big|^{\tilde{p}_1} \Bigg)^{2/\tilde{p}_1} \\
&= O\big( 2^{-2 j_1 m_1} 2^{(j_1 + j_1 L_1/L_2)(2/\tilde{p}_1 - 1)} \big) \\
&= O\big( 2^{-j_1 [2 m_1 m_2 + m_1 + m_2 - 2 m_1/\tilde{p}_2 - 2 m_2/\tilde{p}_1]/L_2} \big).
\end{aligned}
$$

Since $[2m_1m_2 + m_1 + m_2 - 2m_1/\tilde{p}_2 - 2m_2/\tilde{p}_1] > 0$, we have

$$\sum_{(j_1, j_2)\in\mathscr{J}_4} \sum_{k_1, k_2} \theta_I^2 = O\big(2^{-J[2m_1m_2+m_1+m_2-2m_1/\tilde{p}_2-2m_2/\tilde{p}_1]/(L_1+L_2)}\big)$$
$$= O(2^{-J\gamma(m_1, m_2, p_1, p_2)}).$$

The sum over $\mathscr{J}_5$ can be treated analogously, which proves the assertion. □

PROOF OF THEOREM 2.3. Using (2.10) and Lemmas 2.2 and 2.3, we obtain, with $\delta^2 = \varepsilon^2 \log(1/\varepsilon)$, that

$$\begin{aligned}\mathbb{E}\|\hat{\hat{f}} - f\|^2 &= \sum_{I\in\mathscr{K}_\varepsilon} \mathbb{E}\Big(\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_\varepsilon\big) - \theta_I\Big)^2 + \sum_{I\notin\mathscr{K}_\varepsilon} \theta_I^2 \\ &= O\bigg(\#\mathscr{K}_\varepsilon \varepsilon^2\Big(\sqrt{2\log(\#\mathscr{K}_\varepsilon)} + 1\Big)\varphi\Big(\sqrt{2\log(\#\mathscr{K}_\varepsilon)}\Big) + \sum_{I\in\mathscr{K}_\varepsilon} \min\{\delta^2, \theta_I^2\}\bigg) \\ &\quad + O(2^{-J_\varepsilon\gamma(m_1,m_2, p_1, p_2)}) \\ &= O\Big(\Big(\varepsilon^2\sqrt{2\log(\#\mathscr{K}_\varepsilon)} + \Omega_\delta\big(\mathscr{B}, \mathscr{F}_{p_1, p_2}^{m_1, m_2}\big)\Big) + \varepsilon^{2\vartheta(m_1, m_2)}\Big) \\ &= O\Big(\big(\varepsilon^2\log(1/\varepsilon)\big)^{\vartheta(m_1, m_2)}\Big). \qquad \square\end{aligned}$$

PROOF OF LEMMA 2.4. Without loss of generality, let $m_1 \le m_2$. Let $j^*$ be such that $2^{j^*} \le C_0\varepsilon^{-1/(m_1+1)} < 2^{j^*+1}$. To get a lower bound for $\Omega_\varepsilon(\overline{\mathscr{B}}, \mathscr{F}_{p_1, p_2}^{m_1, m_2})$, consider the function

$$f_{\varepsilon,1}(x_1, x_2) = \varepsilon\sum_{k_1} 2^{j^*/2}\psi_{j^*, k_1}(x_1).$$

We have

$$\|f_{\varepsilon,1}\|_{p_i} \le \|f_{\varepsilon,1}\|_\infty \le C\varepsilon 2^{j^*} \le CC_0\varepsilon^{m_1/(m_1+1)},$$
$$\left\|\frac{\partial^{m_1}}{\partial x_1^{m_1}} f_{\varepsilon,1}\right\|_{p_i} \le \left\|\frac{\partial^{m_1}}{\partial x_1^{m_1}} f_{\varepsilon,1}\right\|_\infty \le C\varepsilon 2^{j^*(m_1+1)} \le CC_0$$

and

$$\frac{\partial^{m_2}}{\partial x_2^{m_2}} f_{\varepsilon,1} \equiv 0,$$

which implies that $f_{\varepsilon,1} \in \mathscr{F}_{p_1, p_2}^{m_1, m_2}$ for an appropriate choice of $C_0$. With the exception of a negligible number of boundary wavelets, we have, using (4.8),

$$\theta_{(\psi, j^*, k_1), (\phi, j^*, k_2)} = C\varepsilon,$$

which implies that

$$\sum_{k_1, k_2} \min\big\{\varepsilon^2, \theta^2_{(\psi, j^*, k_1), (\phi, j^*, k_2)}\big\} \ge C\varepsilon^{2\overline{\vartheta}(m_1, m_2)}. \tag{4.9}$$

Now let $f \in \mathscr{F}_{p_1, p_2}^{m_1, m_2}$ be arbitrary. Then we have

$$\sum_{l-1 \le j \le j^*} \sum_{k_1, k_2} \min\left\{\varepsilon^2, \theta^2_{(\psi, j^*, k_1), (\phi, j^*, k_2)}\right\} = O(\varepsilon^2 2^{2j^*}) = O(\varepsilon^{2\bar{\vartheta}(m_1, m_2)}). \tag{4.10}$$

By (4.3) we get

$$\#\left\{(k_1, k_2) \| \theta_{(\psi, j, k_1), (\phi, j, k_2)}| > \varepsilon\right\} = O\left(\varepsilon^{-\tilde{p}_1} 2^{-j[(m_1+1)\tilde{p}_1 - 2]}\right),$$

which, by $(m_1 + 1)\tilde{p}_1 > 2$, implies that

$$\begin{aligned} &\sum_{j > j^*} \sum_{k_1, k_2} \min\left\{\varepsilon^2, \theta^2_{(\psi, j, k_1), (\phi, j, k_2)}\right\} \\ &\quad = \sum_{j > j^*} \varepsilon^2 \#\left\{(k_1, k_2) \| \theta_{(\psi, j, k_1), (\phi, j, k_2)}| > \varepsilon\right\} \\ &\qquad + \sum_{j > j^*} \sum_{(k_1, k_2) : |\theta_{(\psi, j, k_1), (\phi, j, k_2)}| \le \varepsilon} \theta^2_{(\psi, j, k_1), (\phi, j, k_2)} \\ &\quad = \sum_{j > j^*} O\left(\varepsilon^{2 - \tilde{p}_1} 2^{-j[(m_1+1)\tilde{p}_1 - 2]}\right) \\ &\quad = O\left(\varepsilon^{2 - \tilde{p}_1} 2^{-j^*[(m_1+1)\tilde{p}_1 - 2]}\right) = O(\varepsilon^{2\bar{\vartheta}(m_1, m_2)}). \end{aligned} \tag{4.11}$$

The terms corresponding to the basis functions $\phi_{jk_1}(x_1)\psi_{jk_2}(x_2)$ as well as to $\psi_{jk_1}(x_1)\psi_{jk_2}(x_2)$ can be treated analogously. □

PROOF OF THEOREM 3.1. Neglecting the factor $1/(2\pi)$, we show that

$$\begin{aligned} R_T &:= \int_0^1 \int_{-\pi}^{\pi} \left\{ \sum_{s=-\infty}^{\infty} c_T(u, s) \exp(-i\omega s) - f(u, \omega) \right\}^2 du\, d\omega \\ &\to 0 \quad \text{as } T \to \infty, \end{aligned}$$

where

$$\begin{aligned} c_T(u, s) &:= \operatorname{cov}\{X_{[uT - s/2], T}; X_{[uT + s/2], T}\} \\ &= \int_{-\pi}^{\pi} A([uT - s/2]/T, \lambda) \overline{A([uT + s/2]/T, \lambda)} \exp(i\lambda s)\, d\lambda. \end{aligned}$$

Using the relation $\sum_s \exp(i(\lambda - \omega)s) = \delta(\lambda - \omega)$, we obtain

$$\begin{aligned} R_T = \int_0^1 \int_{-\pi}^{\pi} \Bigg\{ \sum_{s=-\infty}^{\infty} \big[ & A([uT - s/2]/T, \lambda) \\ & \times \overline{A([uT + s/2]/T, \lambda)} - f(u, \lambda) \big] \exp(i(\lambda - \omega)s)\, d\lambda \Bigg\}^2 du\, d\omega. \end{aligned}$$

Proceeding quite similarly as in the proof of Lemma 3.2(i) (on the rate of the bias), we have to estimate two terms of similar form. Hence, we only treat the first one which is

$$\hat{\Delta}_s(u,s) = \int \Delta_s(u,\lambda)\exp(i\lambda s)\,d\lambda,$$

with

$$\Delta_s(u,\lambda) := \{A([uT - s/2]/T,\lambda) - A(u,\lambda)\}\overline{A(u,\lambda)},$$

$$\iint \Big|\sum_s \hat{\Delta}_s(u,s)\exp(-i\omega s)\Big|^2\,d\omega\,du$$

$$= \int du \sum_s \sum_v \hat{\Delta}_s(u,s)\overline{\hat{\Delta}_v(u,v)} \int \exp(-i\omega(s-v))\,d\omega$$

$$= \sum_s \int_0^1 du\,|\hat{\Delta}_s(u,s)|^2 = R_T^{(1)} + R_T^{(2)},$$

with

$$R_T^{(1)} = \sum_{|s|\le 2T} \sum_{n=1}^{[s_T^{-1}]+1} \int_{(n-1)s_T}^{(ns_T)\wedge 1} du\,|\hat{\Delta}_s(u,s)|^2$$

and

$$R_T^{(2)} = \sum_{|s|>2T} \int_0^1 du\,|\hat{\Delta}_s(u,s)|^2,$$

where $s_T := |s|/(2T)$, $0 \le |s| \le 2T$.

Similarly to the proof of Lemma 3.2(i), we can show that

$$|s| \sup_{u\in[(n-1)s_T,\,ns_T]} |\hat{\Delta}_s(u,s)|$$

$$\le C_1\Big[\sup_{u,\lambda}|\overline{A(u,\lambda)}| + \sup_u TV_{[-\pi,\pi]}(\overline{A(u,\cdot)})\Big]\sup_\lambda TV_{I_n(s_T)}(A(\cdot,\lambda))$$

$$+ C_2 \sup_{u,\lambda}\{|\overline{A(u,\lambda)}|\} TV_{[-\pi,\pi]}\{TV_{I_n(s_T)}(A(\cdot,\cdot))\},$$

where $C_1$ and $C_2$ denote some positive constants and where, on the right-hand side, the $\sup_u$ is taken over $u \in I_n(s_T) := [(n-1)s_T - 1/T, (n+1)s_T]$ and the $\sup_\lambda$ over $\lambda \in [-\pi,\pi]$.

Note that

$$\sum_{n=1}^{[s_T^{-1}]+1} \sup_{u\in I_n(s_T)} |\hat{\Delta}_s(u,s)| = O(|s|^{-1}),$$

due to Assumption 2(a)–(c) [as $\sum_n TV_{I_n(s_T)}(A(\cdot,\lambda)) \le TV_{[0,1]}(A(\cdot,\lambda))$]. Hence,

$$|R_T^{(1)}| \le \sum_{|s|\le 2T} |s_T| \sum_{n=1}^{[s_T^{-1}]+1} \sup_{u\in I_n(s_T)} |\hat{\Delta}_s(u,s)|^2$$
$$\le (2T)^{-1} \sum_{|s|\le 2T} |s| \left\{ \sum_{n=1}^{[s_T^{-1}]+1} \sup_{u\in I_n(s_T)} |\hat{\Delta}_s(u,s)| \right\}^2 = O\big(T^{-1}\log(T)\big).$$

Further,

$$|R_T^{(2)}| = O\bigg(\sum_{|s|>2T} s^{-2}\bigg) = O(T^{-1})$$

as, by Definition (3.2) for $|s| > 2T$, $\Delta_s(u,\lambda) = \{A(0,\lambda)\overline{A(1,\lambda)} - |A(u,\lambda)|^2\}$ independent of $s$. Hence, $\sup_{0\le u\le 1}|\hat{\Delta}_s(u,s)| = O(|s|^{-1})$. □

PROOF OF LEMMA 3.2. (i) We show that

$$R_T := |\mathbb{E}\tilde{\theta}_I - \theta_I| = O\big(2^{(j_1+j_2)/2}T^{-1}\log(T)\big).$$

By (3.1) and with $A_t(\lambda) := A(t/T,\lambda)$, neglecting the factor $1/(2\pi)$,

$$\mathbb{E}I_{t,T}(\omega) = \sum_{|s/2|\le \min\{t-1,T-t\}} \operatorname{cov}\big(X_{[t-s/2],T}; X_{[t+s/2],T}\big)\exp(-i\omega s)$$
$$= \sum_{|s/2|\le \min\{t-1,T-t\}} \int_{-\pi}^{\pi} A_{[t-s/2]}(\lambda)\overline{A_{[t+s/2]}(\lambda)}\exp(i(\lambda-\omega)s)\,d\lambda.$$

Let $t_T := \min\{t-1, T-t\}$. According to the decomposition

$$A_{[t-s/2]}(\lambda)\overline{A_{[t+s/2]}(\lambda)} - A_t(\lambda)\overline{A_t(\lambda)}$$
$$= \big[A_{[t-s/2]}(\lambda) - A_t(\lambda)\big]\overline{A_t(\lambda)} + A_{[t-s/2]}(\lambda)\big[\overline{A_{[t+s/2]}(\lambda)} - \overline{A_t(\lambda)}\big],$$

we have $R_T = R_T^{(1)} + R_T^{(2)}$ with

$$R_T^{(1)} = \sum_t \int_{(t-1)/T}^{t/T} du\, \psi_{j_1k_1}(u) \sum_{s=-\infty}^{\infty} \int_{-\pi}^{\pi} d\omega\, \tilde{\psi}_{j_2k_2}(\omega)$$
$$\times \int_{-\pi}^{\pi} \big\{|A_t(\lambda)|^2 - f(u,\lambda)\big\}\exp(i(\lambda-\omega)s)\,d\lambda$$
$$- \sum_t \int_{(t-1)/T}^{t/T} du\, \psi_{j_1k_1}(u) \sum_{|s/2|>t_T} \int d\omega\, \tilde{\psi}_{j_2k_2}(\omega)$$
$$\times \exp(-i\omega s)\int |A_t(\lambda)|^2 \exp(i\lambda s)\,d\lambda$$
$$= R_T^{(1,1)} + R_T^{(1,2)}$$

and

$$R_T^{(2)} = \sum_t \int_{(t-1)/T}^{t/T} du\, \psi_{j_1 k_1}(u) \sum_{|s/2| \le t_T} \int d\omega\, \tilde{\psi}_{j_2 k_2}(\omega) \exp(-i\omega s)$$
$$\times \int \overline{A_t(\lambda)} \big[ A_{[t-s/2]}(\lambda) - A_t(\lambda) \big] \exp(i\lambda s)\, d\lambda,$$

where, with $R_T^{(2)}$, we only treat the first part of two similar differences, w.l.o.g.
Now, by Assumptions 2(b), 3 and 4(b),

$$\begin{aligned} |R_T^{(1,1)}| &\le \int d\omega\, |\tilde{\psi}_{j_2 k_2}(\omega)| \sum_t \int_{(t-1)/T}^{t/T} du\, |\psi_{j_1 k_1}(u)| TV_{[(t-1)/T,\, t/T]}(f(\cdot,\omega)) \\ &\le 2^{-j_2/2} 2^{j_1/2} T^{-1} \sup_\omega TV_{[0,1]}(f(\cdot,\omega)) \\ &= O(2^{-j_2/2} 2^{j_1/2} T^{-1}) \end{aligned}$$

and

$$\begin{aligned} |R_T^{(1,2)}| &\le \sup_u \{ |\psi_{j_1 k_1}(u)| \} T^{-1} \sum_t \sum_{|s/2| > t_T} |\widehat{\tilde{\psi}_{j_2 k_2}}(s)| \sup_t |\hat{f}(t/T, s)| \\ &= O(2^{j_1/2} T^{-1}) \sum_t \sum_{|s/2| > t_T} O(2^{j_2/2} s^{-2}) \\ &= O\big(2^{(j_1+j_2)/2} T^{-1} \log(T)\big). \end{aligned}$$

Further, with $s$ being even, w.l.o.g.,

$$R_T^{(2)} = -\sum_s \widehat{\tilde{\psi}_{j_2 k_2}}(s) \sum_t \int_{(t-1)/T}^{t/T} du\, \psi_{j_1 k_1}(u)$$
$$\times \int \overline{A_t(\lambda)} \sum_{n=0}^{s/2-1} \{ A_{t-n}(\lambda) - A_{t-n-1}(\lambda) \} \exp(i\lambda s)\, d\lambda,$$

such that, by Assumption 2(a)–(c), for some positive constant $C$,

$$\begin{aligned} |R_T^{(2)}| &\le C \sum_s |\widehat{\tilde{\psi}_{j_2 k_2}}(s)| \sup_u |\psi_{j_1 k_1}(u)| T^{-1} |s|/2 |s|^{-1} \\ &\quad \times \Big[ \sup_{u,\lambda} \{ |A(u,\lambda)| \} \sup_\lambda TV_{[0,1]}(A(\cdot,\lambda)) \\ &\quad\quad + \sup_{u,\lambda} \{ |\overline{A(u,\lambda)}| \} TV_{U \times \Pi}(A) \\ &\quad\quad + \sup_u TV_{[-\pi,\pi]}(A(u,\cdot)) \sup_\lambda TV_{[0,1]}(A(\cdot,\lambda)) \Big] \\ &= O(2^{j_1/2} 2^{j_2/2} T^{-1}). \end{aligned}$$

The proof of the last estimate (for $R_T^{(2)}$) is delivered by some lengthy, but straightforward algebra using elementary generalizations of total variation estimates and partial summation. Roughly speaking, we proceed as follows: the integral w.r.t. $\lambda$ delivers $s/2$ terms which are all of order $O(s^{-1})$, as for each of the differences labeled by $n$ we use estimates like [cf. Edwards (1979), page 34f.]

$$\int \Delta_t(\lambda)\exp(i\lambda s)\,d\lambda \sim \sum_k \Delta_t(\lambda_k)\{g_s(\lambda_k) - g_s(\lambda_{k-1})\}$$
$$\sim -\sum_k \{\Delta_t(\lambda_{k+1}) - \Delta_t(\lambda_k)\} g_s(\lambda_k),$$

with $\Delta_t(\lambda) := \overline{A_t(\lambda)}(A_t(\lambda) - A_{t-1}(\lambda))$, $g_s(\lambda) := \exp(i\lambda s)/(is)$ and with a sufficiently fine partition $(\lambda_k)_k$ of $[-\pi, \pi]$. Note that $g_s(\lambda) = O(s^{-1})$.

The sum over $t$ can be bounded from above by the bounded total variation of $\Delta_t(\lambda)$ as a function of $u$. Putting both (simultaneously) together, in order to strictly bound all occurring terms, we need Assumption 2(a)–(c), as $\Delta_t(\lambda)$ is a product of the two functions of time and frequency.

(ii) To apply cumulant techniques, we write $\tilde{\theta}_I$ as a quadratic form with a symmetric matrix $N_I$:

$$\tilde{\theta}_I = \mathbf{X}' N_I \mathbf{X},$$

where $N_I = (M_I + \overline{M}_I)/2$ and, with $w_{j_1 k_1}(t/T) := T\int_{(t-1)/T}^{t/T} \psi_{j_1 k_1}(u)\,du$ and $\tilde{w}_{j_2 k_2}(s) := \widehat{\tilde{\psi}_{j_2 k_2}}(s) = (2\pi)^{-1}\int_{-\pi}^{\pi} \tilde{\psi}_{j_2 k_2}(\omega)\exp(i\omega s)\,d\omega$,

$$(M_I)_{tv} = \begin{cases} T^{-1} w_{j_1 k_1}\left(\dfrac{t+v}{2T}\right)\tilde{w}_{j_2 k_2}(t-v), & \text{if } t+v \text{ even},\\[2ex] T^{-1} w_{j_1 k_1}\left(\dfrac{t+v+1}{2T}\right)\tilde{w}_{j_2 k_2}(t-v), & \text{if } t+v \text{ odd}. \end{cases}$$

In the following, for reasons of notational convenience, we use $w_{j_1 k_1}((s+t)/2T)$ to denote $w_{j_1 k_1}([(s+t+1)/2]/T)$. Note that, by the approximations used in the course of the proof, this does not lead to any problems.

Since $\tilde{\phi}$ and $\tilde{\psi}$ are of bounded variation, we get, by integration by parts,

$$\tilde{w}_{j_2 k_2}(t-v) = O\left(2^{-j_2/2} \wedge \left(2^{j_2/2}|t-v|^{-1}\right)\right),$$

which implies that

$$(N_I)_{tv} = \begin{cases} O\left(T^{-1} 2^{j_1/2}\left[2^{-j_2/2} \wedge \left(2^{j_2/2}|t-v|^{-1}\right)\right]\right), \\ \qquad\qquad \text{if } \dfrac{t+v}{2T} \in \left[2^{-j_1}(k_1 - C), 2^{-j_1}(k_1 + C)\right], \\ 0, \quad \text{otherwise}. \end{cases}$$

Hence, we obtain the estimates

$$\max_{t,v}\{|(N_I)_{tv}|\} = O(T^{-1} 2^{j_1/2} 2^{-j_2/2}),$$
$$\|N_I\| \le \|N_I\|_\infty = O\left(T^{-1} 2^{j_1/2} 2^{j_2/2} \log(T)\right)$$

and

$$\begin{aligned}
\tilde{N}_I &= \sum_s \max_t \{|(N_I)_{st}|\} \\
&= \sum_{s:\,|s/T-2^{-j_1}k_1|\le C2^{-j_1}} O(T^{-1}2^{j_1/2}2^{-j_2/2}) \\
&\quad + \sum_{s:\,|s/T-2^{-j_1}k_1|> C2^{-j_1}} O\Big(T^{-1}2^{j_1/2}2^{j_2/2} \max_{t:\,(N_I)_{st}\neq 0} \{|s-t|^{-1}\}\Big) \\
&= O(2^{-j_1/2}2^{-j_2/2}) + O\big(T^{-1}2^{j_1/2}2^{j_2/2}\log(T)\big) \\
&= O(2^{-j_1/2}2^{-j_2/2}).
\end{aligned}$$

Let $\mathbf{Y} \sim N(0, \mathrm{Cov}(\mathbf{X}))$. Since

$$\max_{t,v} \{|(N_I)_{tv}|\}\tilde{N}_I = O(T^{-1}2^{-j_2}),$$

by Lemma 3.1 we obtain that

$$\mathrm{var}\big(\tilde{\theta}_I\big) = \mathrm{var}(\mathbf{Y}'N_I\mathbf{Y}) + O(2^{-j_2}T^{-1}). \tag{4.12}$$

Now, with

$$\begin{aligned}
\mathrm{var}(\mathbf{Y}'N_I\mathbf{Y}) &= 2\,\mathrm{tr}(N_I\Sigma_T N_I\Sigma_T) \\
&= \tfrac{1}{2}\Big[\mathrm{tr}(M_I\Sigma_T M_I\Sigma_T) + \mathrm{tr}\big(\overline{M}_I\Sigma_T\overline{M}_I\Sigma_T\big) \\
&\qquad + 2\,\mathrm{tr}\big(\overline{M}_I\Sigma_T M_I\Sigma_T\big)\Big],
\end{aligned} \tag{4.13}$$

we have to show that

$$\begin{aligned}
&\mathrm{tr}\big(\overline{M}_I\Sigma_T M_I\Sigma_T\big) \\
&\quad = 2\pi T^{-1}\int_{U\times\Pi}\{f(u,\omega)\psi_{j_1k_1}(u)\}^2\,du\,\tilde{\psi}_{j_2k_2}(\omega)\tilde{\psi}_{j_2k_2}(\omega)\,d\omega + o(T^{-1})
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{tr}(M_I\Sigma_T M_I\Sigma_T) &= \mathrm{tr}\big(\overline{M}_I\Sigma_T\overline{M}_I\Sigma_T\big) \\
&= 2\pi T^{-1}\int_{U\times\Pi}\{f(u,\omega)\psi_{j_1k_1}(u)\}^2\,du\,\tilde{\psi}_{j_2k_2}(\omega)\tilde{\psi}_{j_2k_2}(-\omega)\,d\omega \\
&\quad + o(T^{-1}).
\end{aligned}$$

As this runs quite analogously for all terms under consideration, we treat the first one only:

$$
\begin{aligned}
&\operatorname{tr}\left(\overline{M}_I \Sigma_T M_I \Sigma_T\right) \\
&\quad = \sum_{s,v,w,t} \left(\overline{M}_I\right)_{vw} (\Sigma_T)_{ws} (M_I)_{st} (\Sigma_T)_{tv} \\
&\quad = T^{-2} \sum_{s=1}^{T} \sum_{v=1}^{T} \sum_{w=1}^{T} w_{j_1 k_1}\left(\frac{w+v}{2T}\right) \tilde{w}_{j_2 k_2}(w-v) c_T\left(\frac{s+w}{2T}, s-w\right) \\
&\qquad\qquad \times \sum_{t=1}^{T} w_{j_1 k_1}\left(\frac{s+t}{2T}\right) \tilde{w}_{j_2 k_2}(s-t) c_T\left(\frac{t+v}{2T}, t-v\right),
\end{aligned}
$$

where we use the convention given in the beginning of our proof which allows us to proceed regardless to the parity of the arguments of $w_{j_1 k_1}$, and where $\Sigma_T = \operatorname{Cov}(\mathbf{X}) = (c_T(\cdot, \cdot))$, with

$$
\begin{aligned}
c_T\left(\frac{t}{T}, n\right) &= \operatorname{cov}\left\{X_{[t-n/2],T}; X_{[t+n/2],T}\right\} \\
&= \int_{-\pi}^{\pi} A_{[t-n/2]}(\lambda) \overline{A_{[t+n/2]}(\lambda)} \exp(i\lambda n)\, d\lambda.
\end{aligned}
$$

Note that with this,

$$
c_T\left(\frac{t+v}{2T}, t-v\right) = \operatorname{cov}\{X_{t,T}; X_{v,T}\}.
$$

Further, let $\Sigma = (c(\cdot, \cdot))$ with

$$
c\left(\frac{t}{T}, n\right) := (2\pi)^{-1} \int_{-\pi}^{\pi} f\left(\frac{t}{T}, \lambda\right) \exp(i\lambda n)\, d\lambda.
$$

For smooth $A$,

$$
c_T\left(\frac{t}{T}, n\right) = c\left(\frac{t}{T}, n\right) + c'\left(\frac{\cdot}{T}, n\right) O\left(\frac{n}{T}\right)
$$

with both $\sup_{t/T} \sum_n |c(t/T, n)| < \infty$ and $\sup_{t/T} \sum_n |c'(t/T, n)| < \infty$.

If $A$ is not smooth, but fulfills Assumptions 2 and 3, then we proceed as in the proof of part (i) of this lemma, with the same quality of approximation (i.e., the same resulting rates).

In the following, for the sake of notational simplicity, we give the proof of (ii) only for the functions $A(u, \lambda)$ and $\psi_{j_1 k_1}(u)$, which are smooth in $u$.

To motivate the idea how to derive the leading term of the asymptotic variance, we briefly sketch the stationary situation [for details, cf. Gao (1993), page 19, but note the missing symmetrization of the Hermitian matrix $M$ in that reference, which leads to a slight mistake in the resulting asymptotic expression for the whole variance]:

$$
\begin{aligned}
T^{-2} &\sum_{s,t,w,v} \tilde{w}_{jk}(w-v)\tilde{w}_{jk}(s-t)c(s-w)c(t-v) \\
&= T^{-2} \sum_{l,m} \tilde{w}_{jk}(l)\tilde{w}_{jk}(m) \sum_{s} c(s)c(s+l-m) + o(T^{-1}) \\
&= 2\pi T^{-1} \int_{-\pi}^{\pi} f^2(\omega)\tilde{\psi}_{jk}^2(\omega)\, d\omega + o(T^{-1})
\end{aligned}
$$

[similarly to

$$
\sum_{s=-T}^{T} \widehat{\tilde{\psi}_{jk}}(s)c(s) - \int \tilde{\psi}_{jk}(\omega)f(\omega)\, d\omega = \sum_{|s|>T} \widehat{\tilde{\psi}_{jk}}(s)c(s) = O(T^{-1})\Bigg].
$$

To treat all of the occurring remainders, we use estimates like $\sum_n |\tilde{w}_{j_2k_2}(n)| = O(2^{j_2/2})$ [due to Assumption 4(b)] and $\sum_n |c(u,n)| < \infty$ uniformly in $u \in [0,1]$ (due to Assumption 3).

Note, in particular, that $\sum_n |n|\,|l_1(n)|\,|l_2(n)| < \infty$, where $l_i(n)$ is any of $\tilde{w}(n)$, $c(\cdot, n)$ or even of $\sum_w \tilde{w}(w-v)c(s-w) = l(s-v)$, say, with again $\sum_n |l(n)| < \infty$.

Our proof proceeds by three different approximations. The first is replacing $c_T(t/T, n)$ by $c(t/T, n)$ with an error of order $O(n/T)$ (see above). The second one is to replace $c((s+w)/2T, \cdot)$ by $c(s/2T, \cdot) + O(w/T)$ and $w_{j_1k_1}((w+v)/2T)$ by $w_{j_1k_1}(v/2T) + O(2^{j_1/2}w/T)$. With this,

$$
\begin{aligned}
&\operatorname{tr}\big(\overline{M}_I \Sigma M_I \Sigma\big) \\
&\quad = T^{-2} \sum_{s} \sum_{v} \sum_{w} \left[ w_{j_1k_1}\!\left(\frac{v}{2T}\right) + O\!\left(2^{j_1/2}\frac{w}{T}\right)\right] \\
&\qquad\qquad \times \tilde{w}_{j_2k_2}(w-v)\left[c\!\left(\frac{s}{2T}, s-w\right) + O\!\left(\frac{w}{T}\right)\right] \\
&\qquad \times \sum_{t} \left[ w_{j_1k_1}\!\left(\frac{s}{2T}\right) + O\!\left(2^{j_1/2}\frac{t}{T}\right)\right] \tilde{w}_{j_2k_2}(s-t)\left[c\!\left(\frac{v}{2T}, t-v\right) + O\!\left(\frac{t}{T}\right)\right].
\end{aligned}
$$

The leading term of $\operatorname{tr}(\overline{M}_I \Sigma M_I \Sigma)$ turns out to be

$$
\begin{aligned}
T^{-2} \sum_{s} \sum_{v} w_{j_1k_1}\!\left(\frac{v}{2T}\right) w_{j_1k_1}\!\left(\frac{s}{2T}\right) \iint d\lambda\, d\tilde{\lambda}\, f\!\left(\frac{v}{2T}, \lambda\right) f\!\left(\frac{s}{2T}, \tilde{\lambda}\right) \\
\times \tilde{\psi}_{j_2k_2}(\lambda)\tilde{\psi}_{j_2k_2}(\tilde{\lambda}) \exp\big(i(s-v)(\lambda-\tilde{\lambda})\big).
\end{aligned}
$$

The occurring remainders of both the first [i.e., replacing $c_T(\cdot,\cdot)$ by $c(\cdot,\cdot)$] and the second approximations are of the following kind (or even of higher order):

$$
\begin{aligned}
T^{-2} &\sum_{s,v,w,t} w_{j_1k_1}\!\left(\frac{v}{2T}\right) w_{j_1k_1}\!\left(\frac{s}{2T}\right) \tilde{w}_{j_2k_2}(w-v) \\
&\qquad \times \tilde{w}_{j_2k_2}(s-t)\, c\!\left(\frac{s}{2T}, s-w\right) c'\!\left(\frac{\cdot}{2T}, t-v\right) O\!\left(\frac{t}{T}\right).
\end{aligned}
$$

In each of these remainders use estimates like

$$\sum_s \sum_t \frac{|t|}{T} |\tilde{w}_{j_2 k_2}(s-t)| \left| c'\left( \frac{\cdot}{2T}, t-s \right) \right| = o(2^{j_2/2}),$$

and respectively,

$$T^{-2} \sum_s \sum_v \sum_t \left| w_{j_1 k_1}\left( \frac{v}{2T} \right) w_{j_1 k_1}\left( \frac{s}{2T} \right) l(s-v) \frac{|t|}{T} \tilde{w}_{j_1 k_1}(s-t) c'\left( \frac{\cdot}{2T}, t-s \right) \right|$$
$$= O(2^{(j_1+j_2)} T^{-2}) = o(T^{-1}).$$

Finally, the third approximation, which is

$$f\left( \frac{v}{2T}, \tilde{\lambda} \right) = f\left( \frac{s}{2T}, \tilde{\lambda} \right) + f'\left( \frac{\cdot}{T}, \tilde{\lambda} \right) O\left( \frac{s-v}{T} \right)$$

and

$$w_{j_1 k_1}\left( \frac{v}{2T} \right) = w_{j_1 k_1}\left( \frac{s}{2T} \right) + w'_{j_1 k_1}\left( \frac{\cdot}{T} \right) O\left( 2^{j/2} \frac{s-v}{T} \right),$$

delivers a leading term, with $n := s - v$,

$$T^{-2} \sum_s \sum_{|n| \le T} w^2_{j_1 k_1}\left( \frac{s}{2T} \right) \left[ \iint d\lambda \, d\tilde{\lambda} \, \tilde{\psi}_{j_2 k_2}(\lambda) \tilde{\psi}_{j_2 k_2}(\tilde{\lambda}) \right.$$
$$\left. \times f\left( \frac{s}{2T}, \lambda \right) f\left( \frac{s}{2T}, \tilde{\lambda} \right) \exp\left( in(\lambda - \tilde{\lambda}) \right) + R_T^{(3)}(n) \right],$$

with

$$\sum_n |R_T^{(3)}(n)| = \sum_n \frac{|n|}{T} |\hat{\mathscr{F}}(\cdot, n)|^2 = O(2^{j_2} T^{-1}),$$

where

$$\hat{\mathscr{F}}(\cdot, n) = \int \tilde{\psi}_{j_2 k_2}(\lambda) f(\cdot, \lambda) \exp(in\lambda) \, d\lambda$$

is again absolutely summable as a function on $n$, uniformly in its first argument, and with

$$T^{-2} \sum_s w^2_{j_1 k_1}\left( \frac{s}{2T} \right) = O(2^{j_1} T^{-1}).$$

We complete the proof by a technique similar to the proof of part (i), that is, replacing $\sum_{|n| \le T} \dots$ by $\sum_{n=-\infty}^{\infty} \dots$, noting that $\sum_{|n| \ge cT} |\hat{\mathscr{F}}(\cdot, n)|^2 = O(T^{-1})$.

Hence, we end up with the following overall leading term of $\operatorname{tr}(\overline{M}_I \Sigma M_I \Sigma)$:

$$2\pi T^{-2} \sum_s w^2_{j_1 k_1}\left( \frac{s}{2T} \right) \int d\lambda \, \tilde{\psi}^2_{j_2 k_2}(\lambda) f^2\left( \frac{s}{2T}, \lambda \right)$$
$$= 2\pi T^{-1} \int_0^1 du \, \psi^2_{j_1 k_1}(u) \int_{-\pi}^{\pi} d\lambda \, \tilde{\psi}^2_{j_2 k_2}(\lambda) f^2(u, \lambda) + O(T^{-2} 2^{(j_1+j_2)}),$$

due to the bounded total variation of all occurring functions. The proof of (ii) ends by applying the same techniques to the remaining two terms of the sum in (4.13).

(iii) This can be shown simply by using Lemma 3.1 with, by Assumption 5,

$$\begin{aligned}\lambda_{\max}(M_I)\,\lambda_{\max}(\mathrm{Cov}(\mathbf{X})) &= O\big(T^{-1}2^{(j_1+j_2)/2}\log(T)\big)\sup_{1\le t\le T}\Big\{\sum_s \mathrm{cov}(X_s,X_t)\Big\}\\ &= O\big(T^{-1}2^{(j_1+j_2)/2}\log(T)\big)\end{aligned}$$

and the estimates for $\max_{u,v}\{|(M_I)_{uv}|\}$ and $\|M_I\|_\infty$ derived in the proof of part (ii). □

PROOF OF PROPOSITION 3.1. By Lemma 3.2(ii) we get, in conjunction with Assumption 2(c) and (d), that $\sigma_I \asymp T^{-1/2}$ for $T^\rho \le 2^{j_2}$. Hence, we obtain by Lemma 3.2(iii), for appropriate $\mu > 0$,

$$\begin{aligned}\big|\mathrm{cum}_n\big(\tilde\theta_I/\sigma_I\big)\big| &\le (n!)^{2+2\gamma}\big(CT^{-1/2}2^{(j_1+j_2)/2}\log(T)\big)^{n-2}\\ &\le (n!)^{2+2\gamma}\big(CT^{\mu}\big)^{-(n-2)}\end{aligned}\tag{4.14}$$

for all $n \ge 3$, which implies by Lemma 1 in Rudzkis, Saulis and Statulevicius (1978) that

$$P\Big(\pm\big(\tilde\theta_I - \mathbb{E}\tilde\theta_I\big)/\sigma_I \ge x\Big) = (1-\Phi(x))(1+o(1))\tag{4.15}$$

holds uniformly in $0 \le x \le T^{\vartheta}$ for some $\vartheta > 0$.

With $\Delta_I := (\mathbb{E}\tilde\theta_I - \theta_I)/\sigma_I = o(1)$, we get

$$P\Big(\pm\big(\tilde\theta_I - \theta_I\big)/\sigma_I \ge x\Big) = (1-\Phi(x))(1+o(1)) + O(|\Phi(x) - \Phi(x+\Delta_I)|).$$

Fix any $c > 1$. For $x \le c$, obviously

$$\Phi(x) - \Phi(x+\Delta_I) = o(1-\Phi(x)).\tag{4.16}$$

Let $\Delta_I \ge 0$ w.l.o.g. Using the formula $(1 - 1/x^2)\varphi(x)/x \le (1-\Phi(x))$, we obtain for $x > c$ that

$$|\Phi(x) - \Phi(x+\Delta_I)| = \Delta_I\varphi(x) = o(1-\Phi(x)),\tag{4.17}$$

which, in conjunction with (4.15) and (4.16), completes the proof. □

PROOF OF PROPOSITION 3.2. First, let $I \in \mathcal{J}_T \cap \{I | 2^{j_2} > T^\rho\}$. Since $\delta^{(\cdot)}$ is monotonic in its first argument, there exists some $\gamma_I$ such that

$$\begin{aligned}\delta^{(\cdot)}\big(\tilde\theta_I, \lambda_{I,T}\big) &\ge \theta_I \quad \text{if } \tilde\theta_I - \theta_I > \gamma_I,\\ \delta^{(\cdot)}\big(\tilde\theta_I, \lambda_{I,T}\big) &\le \theta_I \quad \text{if } \tilde\theta_I - \theta_I < \gamma_I.\end{aligned}$$

Without loss of generality, we assume that $\delta^{(\cdot)}(\tilde\theta_I, \lambda_{I,T}) \ge \theta_I$ if $\tilde\theta_I - \theta_I = \gamma_I$.

Let $\eta_T = CT^{-1/2}\sqrt{\log(T)}$ for some appropriate $C$. Then

$$\begin{aligned}\mathbb{E}\Big(\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,T}\big) - \theta_I\Big)^2 &= \mathbb{E}I\big(\gamma_I \le \tilde{\theta}_I - \theta_I < \eta_T\big)\Big(\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,T}\big) - \theta_I\Big)^2 \\ &\quad + \mathbb{E}I\big(-\eta_T < \tilde{\theta}_I - \theta_I < \gamma_I\big)\Big(\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,T}\big) - \theta_I\Big)^2 \\ &\quad + \mathbb{E}I\big(|\tilde{\theta}_I - \theta_I| \ge \eta_T\big)\Big(\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,T}\big) - \theta_I\Big)^2 \\ &= S_1 + S_2 + S_3.\end{aligned}$$

Applying integration by parts w.r.t. $x$, we obtain by Proposition 3.1 that

$$\begin{aligned}S_1 &= -\int\Big[I(\gamma_I \le \sigma_I x < \eta_T)\big(\delta^{(\cdot)}(\theta_I + \sigma_I x, \lambda_{I,T}) - \theta_I\big)^2\Big] \\ &\qquad \times d\Big\{P\Big(\big(\tilde{\theta}_I - \theta_I\big)/\sigma_I \ge x\Big)\Big\} \\ &= \int\Big\{P\Big(\big(\tilde{\theta}_I - \theta_I\big)/\sigma_I \ge x\Big)\Big\} \\ &\qquad \times d\Big[I(\gamma_I \le \sigma_I x < \eta_T)\big(\delta^{(\cdot)}(\theta_I + \sigma_I x, \lambda_{I,T}) - \theta_I\big)^2\Big] \\ &\qquad + P\Big(\big(\tilde{\theta}_I - \theta_I\big)/\sigma_I \ge \gamma_I\Big)\big(\delta^{(\cdot)}(\theta_I + \gamma_I, \lambda_{I,T}) - \theta_I\big)^2 \qquad (4.18) \\ &\le C_T\bigg\{\int\{1 - \Phi(x)\} \\ &\qquad \times d\Big[I(\gamma_I \le \sigma_I x < \eta_T)\big(\delta^{(\cdot)}(\theta_I + \sigma_I x, \lambda_{I,T}) - \theta_I\big)^2\Big] \\ &\qquad + P\Big(\big(\tilde{\xi}_I - \theta_I\big)/\sigma_I \ge \gamma_I\Big)\big(\delta^{(\cdot)}(\theta_I + \gamma_I, \lambda_{I,T}) - \theta_I\big)^2\bigg\} \\ &= C_T\mathbb{E}I\big(\gamma_I \le \tilde{\xi}_I - \theta_I < \eta_T\big)\Big(\delta^{(\cdot)}\big(\tilde{\xi}_I, \lambda_{I,T}\big) - \theta_I\Big)^2\end{aligned}$$

holds uniformly in $I \in \mathscr{I}_T \cap \{I | 2^{j_2} > T^\rho\}$ for some $C_T \to 1$. The term $S_2$ can be estimated analogously.

Using Lemma 3.2, we obtain, for arbitrary even $n$, that

$$\mathbb{E}\big(\tilde{\theta}_I - \theta_I\big)^n = O\Bigg(\sum_{r=1}^{n} \prod_{i_1,\dots,i_r:\, i_1+\cdots+i_r=n,\, i_j \ge 1} \Big|\mathrm{cum}_{i_j}\big(\tilde{\theta}_I\big)\Big|\Bigg) = O(T^{-n/2}),$$

which implies, by the Cauchy–Schwarz inequality, that

$$\begin{aligned}S_3 &\le \sqrt{P\big(|\tilde{\theta}_I - \theta_I| \ge \eta_T\big)}\sqrt{\mathbb{E}\Big(\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,T}\big) - \theta_I\Big)^4} \qquad (4.19) \\ &= O(T^{-2})\end{aligned}$$

if $C$ is chosen large enough.

As $|\delta^{(\cdot)}(\tilde{\theta}_I, \lambda_{I,T}) - \theta_I| \le \lambda_{I,T} + |\tilde{\theta}_I - \theta_I|$, the terms with $2^{j_2} \le T^\rho$ contribute to the risk a term of order $O(T^{\rho-1}\log(T))$, which is $O(T^{-\vartheta(m_1, m_2)})$, if $\rho$ is chosen sufficiently small. □

PROOF OF THEOREM 3.2. Using Parseval's identity, we infer from Proposition 3.2 and by $|\delta^{(\cdot)}(\tilde{\theta}_I, \lambda) - \theta_I| \le \lambda + |\tilde{\theta}_I - \theta_I|$ that

$$
\begin{aligned}
\mathbb{E}\|\hat{f} - f\|^2 &= \sum_{I \in \mathcal{I}_T} \mathbb{E}\Big(\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,T}\big) - \theta_I\Big)^2 + \sum_{I \notin \mathcal{I}_T} \theta_I^2 \\
&= (1 + o(1)) \sum_{I \in \mathcal{I}_T^*} \mathbb{E}\Big(\delta^{(\cdot)}\big(\tilde{\xi}_I, \lambda_{I,T}\big) - \theta_I\Big)^2 + O(T^{-\vartheta(m_1, m_2)}) \\
&\quad + \sum_{I \in \mathcal{I}_T \setminus \mathcal{I}_T^*} \Big(2\lambda_{I,T}^2 + 2\mathbb{E}\big(\tilde{\theta}_I - \theta_I\big)^2\Big) \\
&\quad + O(T^{-(1-\delta)\gamma(m_1, m_2, p_1, p_2)}).
\end{aligned} \tag{4.20}
$$

From (2.10) we see that the first term on the right-hand side of (4.20) can be estimated by

$$
\begin{aligned}
&C \sum_{I \in \mathcal{I}_T^*} \left(\sigma_I^2 \left(\frac{\lambda_{I,T}}{\sigma_I} + 1\right) \varphi\left(\frac{\lambda_{I,T}}{\sigma_I}\right) + \min\{\lambda_{I,T}^2, \theta_I^2\}\right) + O(T^{-\vartheta(m_1, m_2)}) \\
&\quad \le C \sum_{I \in \mathcal{I}_T^*} \min\{\sigma_I^2 \log(T), \theta_I^2\} + O(T^{-\vartheta(m_1, m_2)}) \\
&\quad = O\Big(\Omega_{\max\{\sigma_I\sqrt{\log(T)}\}}\big(\mathcal{B}, \mathcal{F}_{p_1, p_2}^{m_1, m_2}\big) + T^{-\vartheta(m_1, m_2)}\Big) \\
&\quad = O\Big(\big(\log(T)/T\big)^{\vartheta(m_1, m_2)}\Big).
\end{aligned}
$$

The remaining terms on the right-hand side of (4.20) are also of order $O((\log(T)/T)^{\vartheta(m_1, m_2)})$, which completes the proof. □

PROOF OF THEOREM 3.3. Since $\delta^{(\cdot)}$ is monotonic in the second argument, we have for any random threshold $\hat{\lambda}_I$ satisfying $\lambda_{I,1} \le \hat{\lambda}_I \le \lambda_{I,2}$ that

$$
\Big|\delta^{(\cdot)}\big(\tilde{\theta}_I, \hat{\lambda}_I\big) - \theta_I\Big| \le \max\Big\{\Big|\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,1}\big) - \theta_I\Big|, \Big|\delta^{(\cdot)}\big(\tilde{\theta}_I, \lambda_{I,2}\big) - \theta_I\Big|\Big\}.
$$

For

$$
CT^{-1/2}\sqrt{\log(T)} \ge \gamma_T \sigma_I \sqrt{2\log(\#\mathcal{I}_T^0)},
$$

both the nonrandom thresholds $\lambda_{I,T} = \gamma_T \sigma_I \sqrt{2\log(\#\mathcal{I}_T^0)}$ as well as $\lambda_{I,T} = CT^{-1/2}\sqrt{\log(T)}$ provide the desired rate for the risk. Hence we obtain

$$
\begin{aligned}
&\sum_{I \in \mathcal{I}_T^0} \mathbb{E}\Big(\delta^{(\cdot)}\big(\tilde{\theta}_I, \hat{\lambda}_I\big) - \theta_I\Big)^2 \\
&\quad \le \sum_{I \in \mathcal{I}_T^0} \mathbb{E}\Big(\delta^{(\cdot)}\Big(\tilde{\theta}_I, \gamma_T \sigma_I \sqrt{2\log(\#\mathcal{I}_T^0)}\Big) - \theta_I\Big)^2 \\
&\qquad + \sum_{I \in \mathcal{I}_T^0} \mathbb{E}\Big(\delta^{(\cdot)}\Big(\tilde{\theta}_I, CT^{-1/2}\sqrt{\log(T)}\Big) - \theta_I\Big)^2
\end{aligned}
$$

$$+ \sum_{I \in \mathscr{I}_T^0} \mathbb{E} I\Big( \hat{\lambda}_I \notin \Big[ \gamma_T \sigma_I \sqrt{2 \log(\# \mathscr{I}_T^0)}\, , C T^{-1/2} \sqrt{\log(T)} \Big] \Big) \big( 2 \tilde{\theta}_I^2 + 2 \theta_I^2 \big)$$

$$= O\Big( (\log(T)/T)^{\vartheta(m_1, m_2)} \Big).$$

From the proof of Theorem 3.2, we know that the risk arising from the estimation of $\theta_I$, $I \notin \mathscr{I}_T^0$, is also of order $O((\log(T)/T)^{\vartheta(m_1, m_2)})$, which completes the proof. □

**Acknowledgments.** We thank D. Donoho and P. Mathé for helpful discussions, and two anonymous referees and an associate editor for their suggestions that helped to improve the presentation of our work.

## REFERENCES


BERKOLAJKO, M. Z. and NOVIKOV, I. YA. (1992). Bases of wavelets in spaces of differentiable functions of anisotropic smoothness. *Russian Acad. Sci. Dokl. Math.* **45** 382–386.

BEYLKIN, G., COIFMAN, R. and ROKHLIN, V. (1993). Fast wavelet transforms and numerical algorithms I. *Comm. Pure Appl. Math.* **44** 141–183.

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densitiés: risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137.

BROWN, D. L. and LOW, M. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

COHEN, A., DAUBECHIES, I. and 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. In *Athens Conference on Applied Probability and Time Series Analysis* (P. M. Robinson and M. Rosenblatt, eds.) **2**. Springer, New York.

DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** 1–37.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmonic Anal.* **3** 215–228.

DONOHO, D. L. (1995). CART and best-ortho-basis: a connection. Technical report, Dept. Statistics, Stanford Univ.

DONOHO, D. L. and JOHNSTONE, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

DONOHO, D. L. and JOHNSTONE, I. M. (1994b). Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sci. Paris Sér. I Math.* **319** 1317–1322.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

EDWARDS, R. E. (1979). *Fourier Series. A Modern Introduction* **1**, 2nd ed. Springer, New York.

GAO, H.-Y. (1993). Wavelet estimation of spectral densities in time series analysis. Ph.D. dissertation, Univ. California, Berkeley.

MARTIN, W. and 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. Iberoamericana* **7** 115–133.

NEUMANN, M. H. (1994). Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. *J. Time Ser. Anal.* To appear.

NEUMANN, M. H. and SPOKOINY, V. G. (1995). On the efficiency of wavelet estimators under arbitrary error distributions. *Math. Methods Statist.* **4** 137–166.

NEUMANN, M. H. and VON SACHS, R. (1995). Wavelet thresholding: beyond the Gaussian i.i.d. situation. *Wavelets and Statistics*. *Lecture Notes in Statist*. 301–329. Springer, New York.

NIKOL'SKII, S. M. (1975). *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, Berlin.

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann*. *Statist*. **24** 2399–2430.

PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series* **2**. Academic Press, London.

RUDZKIS, R., SAULIS, L. and STATULEVICIUS, V. (1978). A general lemma on probabilities of large deviations. *Lithuanian Math*. *J*. **18** 226–238.

TRIBOULEY, K. (1995). Practical estimation of multivariate densities using wavelet methods. *Statist*. *Neerlandica* **49** 41–62.

VON SACHS, R. and SCHNEIDER, K. (1996). Wavelet smoothing of evolutionary spectra by nonlinear thresholding. *Appl*. *Comput*. *Harmonic Anal*. **3** 268–282.

WALSH, J. B. (1986). Martingales with a multidimensional parameter and stochastic integrals in the plane. *Lectures in Probability and Statistics*. *Lecture Notes in Math*. **1215** 329–491. Springer, Berlin.

WEIERSTRASS FÜR INSTITUT
ANGEWANDTE ANALYSIS UND STOCHASTIK
MOHRENSTRASSE 39
D-10117 BERLIN
GERMANY

UNIVERSITÄT KAISERSLAUTERN
FACHBEREICH MATHEMATIK
POSTFACH 3049
D-67653 KAISERSLAUTERN
GERMANY