Wavelet smoothing of evolutionary spectra by non–linear thresholding

Rainer von Sachs

Fachbereich Mathematik, Universität Kaiserslautern
Postfach 3049, D–67653 Kaiserslautern, Germany
internet: rvs@mathematik.uni–kl.de

Kai Schneider

Fachbereich Chemie, Technische Chemie, Universität Kaiserslautern
Postfach 3049, D–67653 Kaiserslautern, Germany
internet: kschneid@rhrk.uni–kl.de

22. December 1995

Abstract

We consider wavelet estimation of the time–dependent (*evolutionary*) power spectrum of a *locally stationary* time series. Hereby, wavelets are used to provide an adaptive local smoothing of a short–time periodogram in the time–frequency plane. For this, in contrast to classical nonparametric (*linear*) approaches we use *non–linear* thresholding of the empirical wavelet coefficients. We show how these techniques allow for both adaptively reconstructing the local structure in the time–frequency plane and for denoising the resulting estimates. To this end a threshold choice is derived which results into a near–optimal L_2–minimax rate for the resulting spectral estimator. Our approach is based on a 2–d orthogonal wavelet transform modified by using a cardinal Lagrange interpolation function on the finest scale. As an example, we apply our procedure to a time–varying spectrum motivated from mobile radio propagation.

1 Introduction

Estimating power spectra which (slowly) change over time is an important problem for describing and analysing many physical phenomena which exhibit an instationary behavior over time (quasi–oscillating behavior, transients, etc.). Examples are numerous and can be found, e.g., in the field of speech and sound analysis. There are a lot of approaches trying to model this by introducing time–frequency–spectra, among which the Wigner–Ville (WV)–spectrum plays a prominent role (see, e.g. [11]).
In our approach we start from a *locally stationary* process as underlying model and use an extension of the definition of the WV–spectrum introduced by Dahlhaus [4], the so–called *evolutionary spectrum*. It allows for both the unique *definition* and the consistent *estimation* of what might be called the spectrum of a non–stationary time series in a single time point. In contrast to the (parametric) approach in [4], here, we address the problem of nonparametrically estimating the underlying time–varying spectrum by a short–time periodogram. Then, immediately the question arises how to smooth this eratic estimator as a function of both time and frequency. Here, we suggest the use of 2–d wavelet thresholding. This approach can be interpreted in a two–fold context: On one hand, it serves as a starting point in considerung the general problem in time–frequency analysis and estimation: How can one adapt to the underlying time–frequency content of a non–stationary time series? On the other hand, the original smoothing problem can be considered as a special example in the more general context of *two–dimensional* adaptive smoothing problems. Here, for time–varying spectrum estimation, further difficulties arise from the non–Gaussian and 2–d dependent error structure of the considered random variables.
More specifically, in order to solve this smoothing problem, we use a two–dimensional periodic wavelet expansion of the short–time periodogram over segments of length N of the data $X_1, \ldots, X_T$. Thereby we choose as orthonormal wavelet basis, in the resulting two–dimensional multiresolution analysis, periodized splines of the orthogonal Battle–Lemarié family (see, e.g. [8]). In contrast to classical wavelet transformation, by use of a cardinal Lagrange interpolation function on the finest scale we prevent from losing information of our sampled periodogram data on this finest scale. Moreover, both numerical approximation and statistical properties of our resulting wavelet coefficients are improved by this "collocation" wavelet transform. By *non–linear* thresholding (see, e.g., [6]) of the empirical wavelet coefficients, analogously to the work of [9] for stationary time series, we provide a *local* smoothing procedure, i.e. a denoised estimator which adapts to the local structure of the non–stationary time series. Moreover, we show that our estimator attains a near–optimal minimax L_2–convergence rate uniformly over a whole function class. Hence, its asymptotic properties parallel what can be found in a variety of seminal papers by Donoho et al. (cf. [6], e.g., and [7] for a short overview): Here we only like to mention the superiority of *non–linear* threshold estimators over traditional *linear* schemes, e.g. *global* kernel smoothers, for functions of inhomogeneously distributed low regularity (like those of bounded variation).
The content of the present paper is organized as follows: A two–dimensional periodic multiresolution and a collocation wavelet transform is recalled in Section 2. Then, in Section 3 we review the concept of evolutionary spectra of a locally stationary process, as desribed in [4]. Moreover, we introduce the short–time periodogram as a localized periodogram over segments of the original time series. Section 4, the main part, treats wavelet estimation of the evolutionary spectrum by non–linear thresholding of the empirical co-

efficients, i.e. the coefficients of the localized periodogram estimates. We first investigate their statistical behavior and estimate their tail probability to end up with an appropriate threshold. Our main theorem shows that the resulting threshold estimator achieves a near–optimal rate of convergence of the L_2–risk, i.e. the integrated mean–squared error between estimate and true spectrum. At this place, we restrict to rigorously derive the results for Hölder function classes only, mainly in order to simplify the proofs. But by adopting techniques similar to those in [15] we were able to generalize our main results to more interesting functions of lower inhomogeneously distributed regularity. With this, basically continuous functions of bounded variation are included. Finally, Section 5 deals with applications and simulations. We give an example from mobile radio propagation [10] and study the performance of our non–linear wavelet estimate by simulating this example of a time–dependent spectrum. Also, we give a comparison with a linear estimate, a globally smoothed wavelet estimator. It turns out that with local smoothing we are able to suppress most of the noise in the estimate without losing relevant structure of the spectrum.

2 Periodic orthonormal wavelet bases

2.1 Periodic multiresolution of $L_2(\mathbb{T}^2)$

This section serves to recall the construction of a multiresolution analysis (MRA) of $L_2(\mathbb{T})$, the space of real–valued square integrable 1–periodic functions living on the torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ [16], [12]. It is equipped with the inner product $\langle f, g\rangle_{\mathbb{T}} = \int_{\mathbb{T}} f(x)\overline{g(x)}dx$ and the corresponding norm $\|f\|_2^2 = \langle f, f\rangle_{\mathbb{T}}$. A periodic MRA of $L_2(\mathbb{T})$, which is a sequence of embedded subspaces $V_j \subset V_{j+1},\ j \geq 0$ of $L_2(\mathbb{T})$, can be obtained through periodization of the scaling functions $\tilde{\varphi}$ and the wavelets $\tilde{\psi}$ constituting a MRA of $L_2(\mathbb{R})$ by the relation

$$\varphi_{j,k}(x) = \sum_{n\in\mathbb{Z}} \tilde{\varphi}_{j,k}(x-n) = 2^{j/2} \sum_{n\in\mathbb{Z}} \tilde{\varphi}(2^j(x-n)-k).$$

The definition of a MRA in the periodic case carries over from the nonperiodic case with only slight technical modifications, see [16]. In the classical case the scaling functions $\{\varphi_{j,k},\ k = 0, \ldots, 2^j - 1\}$ constitute an orthonormal basis of V_j, so that $\langle \varphi_{j,i},\ \varphi_{j,k}\rangle_{\mathbb{T}} = \delta_{i,k}$. The orthogonal complement space W_j of V_j in V_{j+1} is given by

$$W_j = \overline{span}\,\{\psi_{j,k}(x),\ k = 0, \ldots, 2^j - 1\}$$

with the orthogonal wavelets $\psi_{j,k}(x) = \psi_{j,0}(x - \frac{k}{2^j})$.

We obtain a two-dimensional MRA of $L_2(\mathbb{T}^2)$ through the tensor product of two one-dimensional MRAs of $L_2(\mathbb{T})$ [12]. We define the bivariate scaling function via $\Phi(x, y) = \varphi(x)\varphi(y)$ and the corresponding wavelets via $\Psi^h(x, y) = \varphi(x)\psi(y)$, $\Psi^v(x, y) = \psi(x)\varphi(y)$, $\Psi^d(x, y) = \psi(x)\psi(y)$. Similarly defining the space $\mathbf{V}_j = V_j \otimes V_j$, we get for its orthogonal complement in $\mathbf{V}_{j+1}$ the space $\mathbf{W}_j$, $j \geq 0$,

$$\mathbf{W}_j = \overline{span}\ \{\Psi^\mu_{j\mathbf{k}}(x, y) :\ \mathbf{k} = (k_1, k_2),\ k_i = 0, \ldots, 2^j - 1, \mu = h, v, d\}\ ,$$

i.e. consisting of three different wavelets (horizontal, vertical and diagonal). Thereby, we obtain a decomposition of $L_2(\mathbb{T}^2)$ into mutually orthogonal subspaces

$$L_2(\mathbb{T}^2) = \mathbf{V}_0 \oplus_{j\geq 0} \mathbf{W}_j$$

and a decomposition of a function $f \in L_2(\mathbb{T}^2)$ by

$$f(x,y) = c_{0\mathbf{0}} + \sum_{j=0}^{\infty} \sum_{\mathbf{k}=0}^{2^j-1} \sum_{\mu=h,v,d} d_{j\mathbf{k}}^{\mu}\, \Psi_{j\mathbf{k}}^{\mu}(x,y) \tag{2.1}$$

with the wavelet coefficients

$$d_{j\mathbf{k}}^{\mu} = \langle f(x,y),\, \Psi_{j,\mathbf{k}}^{\mu}(x,y)\rangle_{\mathbb{T}^2}\,, \quad c_{0\mathbf{0}} = \int_{\mathbb{T}^2} f(x,y)\, dx\, dy. \tag{2.2}$$

Due to the periodicity, the index range in (2.1) is of finite length 2^j. Furthermore, there exists a coarsest scale $j = 0$, because $\varphi_{j,k} = 1$ for $j \leq 0$ as $\sum_k \tilde{\varphi}_{j,k} = 1$.

2.2 Collocation wavelet transform on $L_2(\mathbb{T}^2)$

The projection $f_J = P_J f$ of $f \in L_2(\mathbb{T}^2)$ onto the 2^{2J}-dimensional subspace $\mathbf{V}_J \subset L_2(\mathbb{T}^2)$is somewhat arbitrary. Identification of the sampled values of $f\left(\frac{k_1}{2^J}, \frac{k_2}{2^J}\right)$ with the scaling coefficients c_{k_1,k_2} results in an approximation of low order $O(2^{-J})$, only. Therefore we choose a collocation projection f_J, defined by

$$f_J(x,y) = \sum_{\mathbf{k}_1=0}^{2^J-1} \sum_{\mathbf{k}_2=0}^{2^J-1} f\left(\frac{k_1}{2^J}, \frac{k_2}{2^J}\right) S_J\left(x - \frac{k_1}{2^J},\, y - \frac{k_2}{2^J}\right) \tag{2.3}$$

which prevents from losing information on the finest scale J. In (2.3) $S_J(x,y)$ denotes the cardinal Lagrange function of the space $\mathbf{V}_J$ defined in [20].
The calculation of the scaling coefficients $c_{J\mathbf{k}}$ starting with samples of f at grid points $\{\frac{k_1}{2^J}, \frac{k_2}{2^J}\}$ is done with a discrete 2^J–periodic interpolation filter $L_J(\mathbf{k}) = \langle S_J(x,y),\, \Phi_{J,\mathbf{k}}(x,y)\rangle_{\mathbb{T}^2}$ [16], which is an optimal quadrature formula in $\mathbf{V}_J$.

The decomposition of $f_J \in \mathbf{V}_J$ into contributions from $\mathbf{V}_0$ and $\mathbf{W}_j$, $j = J-1, \ldots, 0$, is calculated subsequently with projections onto $\mathbf{V}_{j-1}$ and $\mathbf{W}_{j-1}$ by recursive application of discrete periodic filters G and H [16]. The filters G and H are defined as follows [12]

$$\begin{aligned} G_j^{\mu}(\mathbf{n}) &= \langle \Psi_{j-1,\mathbf{0}}^{\mu}(x,y),\, \Phi_{j,\mathbf{n}}(x,y)\rangle_{\mathbb{T}^2} \\ H_j(\mathbf{n}) &= \langle \Phi_{j-1,\mathbf{0}}(x,y),\, \Phi_{j,\mathbf{n}}(x,y)\rangle_{\mathbb{T}^2} \end{aligned} \tag{2.4}$$

The efficient realization of the periodic wavelet transform is based on fast convolution techniques [16] employing discrete Fast Fourier transforms (FFT) and downsampling in Fourier space [8]. The analytic expressions in Fourier space for the wavelets, scaling functions, the cardinal Lagrange function and the discrete filters for spline wavelets of the Battle–Lemarié family, used in the present work, are given in [16], [8]. With the choice $m = 6$, where m denotes the order of the splines, we experienced good results, and this

seems to be a good compromise between localization in Fourier and physical space. Furthermore the resulting wavelets are in $C^4(I\!R^2)$ and are symmetric.

This set–up will be used in section 4 for 2–d wavelet smoothing of an estimator of the evolutionary spectrum. We will introduce this particular class of time-varying spectra for so–called locally stationary processes in the following section.

3 Evolutionary spectra of locally stationary processes

We start with a review of the model of a locally stationary process as given by Dahlhaus in [4], Section 2, which generalizes the Cramér representation of a stationary stochastic process.

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

$$X_{t,T} = \mu(\frac{t}{T}) + \int_{-\frac{1}{2}}^{\frac{1}{2}} A^o_{t,T}(\lambda)\, \exp(2\pi i\lambda t)\, d\xi(\lambda)\,, \tag{3.1}$$

where

(i) $\xi(\lambda)$ is a stochastic process on $[-\frac{1}{2}, \frac{1}{2}]$ with $\overline{\xi(\lambda)} = \xi(-\lambda)$, $\mathsf{E}\,\xi(\lambda) = 0$ and orthonormal increments, i.e. $cov(d\xi(\lambda),\ d\xi(\lambda')) = \delta(\lambda - \lambda')\, d\lambda$, $cum\{d\xi(\lambda_1), ..., d\xi(\lambda_k)\} = \eta(\sum_{j=1}^k \lambda_j)\ h_k(\lambda_1, ..., \lambda_{k-1})\ d\lambda_1\ ...\ d\lambda_k$,
where $cum\{\ldots\}$ denotes the cumulant of order k, $|h_k(\lambda_1, ..., \lambda_{k-1})| \le const_k$ for all k (with $h_1 = 0, h_2(\lambda) = 1$) and $\eta(\lambda) = \sum_{j=-\infty}^{\infty} \delta(\lambda + j)$, and where

(ii) there exists a positive constant K and a smooth fucntion $A(u, \lambda)$ on $[0, 1] \times [-\frac{1}{2}, \frac{1}{2}]$ which is 1–periodic in λ, with $A(u, -\lambda) = \overline{A(u, \lambda)}$ such that for all T,

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

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

This class of locally stationary processes includes ARMA processes with time–varying coefficients (see [5], Theorem 2.3), and, of course, if A and μ do not depend on t and T, ordinary stationary processes. For simplicity we assume that $\mu(u) = 0$, i.e. we do not treat the problem of estimating the mean of the time series. Also, here we restrict ourselves to *Gaussian* processes – mainly for reasons of using proof techniques which are somewhat similar to those of [9]. But we like to note that using techniques as in [14] will enable us to derive similar results also for the non–Gaussian situation.

In this model the smoothness of A in u restricts the departure from stationarity and ensures the locally stationary behavior of the process. It also allows to define a unique underlying time–varying spectrum of $\{X_{t,T}\}$, as follows.

Consider first the so–called *Wigner–Ville spectrum* (see, e.g., [11]), which for fixed T in this situation is

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

where $X_{t,T}$ is defined by (3.1), with $A^o_{t,T}(\lambda) = A(0,\lambda)$ for $t < 1$ and $A^o_{t,T}(\lambda) = A(1,\lambda)$ for $t > T$.
Then, by the following definition, $f_T(u,\lambda)$ will be related to the smooth amplitude function $A(u,\lambda)$:

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

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

This $f(u,\lambda)$ is, in general in some mean–square sense, the limit of $f_T(u,\lambda)$ as $T \to \infty$:

By [5], Theorem 2.2, if $A(u,\lambda)$ is uniformly Lipschitz in u and λ with index $\alpha > 1/2$, then, for all $u \in (0,1)$,

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \mid f_T(u,\lambda) - f(u,\lambda) \mid^2 d\lambda \;=\; o(1) \;. \tag{3.5}$$

Remark 3.3 The representation (3.1) is based on a sequence of functions $A^o_{t,T}(\lambda)$ instead of the smooth function $A(t/T,\lambda)$ itself. For some simple examples, like time–dependent moving average processes, $A^o_{t,T}(\lambda) = A(t/T,\lambda)$. 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.
If we restrict to a sub–class of locally stationary processes with a representation in (3.1) which is based directly on a smooth $A(t/T,\lambda)$ with uniformly bounded first partial derivatives, then equation (3.5) holds in an almost everywhere pointwise sense. If we were to consider less regular $A(u,\lambda)$ being only continuous and of bounded 2–d total variation, then equation (3.5) will continue to hold in some L_2–sense (see [15]). This is the least regular class of functions which we can allow for the generalization of our main Theorem 4.10 to hold (see Remark 4.9 in the next section).

Remark 3.4 In (3.1) – (3.3) t and s denote time points in the interval $[1,T]$ while u denotes a time point in the rescaled interval $[0,1]$, i.e. $u = \frac{t}{T}$. Note that (3.1) does <u>not</u> define a finer and finer discretized continuous time process as T tends to infinity. It rather means that more and more data of the same local structure, given by $A(\frac{t}{T},\lambda)$, are observed with increasing T. As illustration we cite an example given in [4], which is

$$A(\frac{t}{T},\lambda) \;= \mathbb{1}_{[\frac{1}{4},\frac{1}{2}]}(\frac{\mathrm{t}}{\mathrm{T}})\,\delta(\lambda-\lambda_1) \;+\; \mathbb{1}_{(\frac{1}{2},\frac{3}{4}]}(\frac{\mathrm{t}}{\mathrm{T}})\,\delta(\lambda-\lambda_2) \tag{3.6}$$

With increasing T more and more periods of the two harmonics $\exp(2\pi i\lambda_1 t)$ and $\exp(2\pi i\lambda_2 t)$ are observed.

We can interprete this approach of rescaling in time as trying to bound the complexity of the spectrum as the object which defines the underlying model: Without any rescaling (or related further assumptions) it is per se not possible to make statistical inference on a spectrum $f(t, \lambda)$ which depends on t in the naive way, i.e. $t = 1, ..., T$, even with T tending to infinity. Consider for a moment, however, non–parametric regression, where with growing sample size we get more and more statistical information due to a denser and denser regression design. There the object which defines the underlying model is a fixed regression function (on some unit interval). Hence, back to considering the spectrum, by rescaling $f(t/T, \lambda)$ and assuming smoothness the complexity of this new object now increases slower as the amount of statistical information when the length T of the series grows.
Moreover, Definition 3.2, under the smoothness assumptions of (3.5), turns out to be unique (cf. the uniqueness of the Wigner–Ville spectrum, pointed out in [11], Section 2, B.7). This is an inherent advantage of the approach given in [4] when trying to define what is meant by "the spectrum of a non–stationary process $X_1, \ldots, X_T$ at a fixed time point t_0": Due to the non–stationarity only a few points around t_0 may have the same spectral structure. It is clear that the probability structure of these few points does not specify a spectral density uniquely. But as pointed out before and as can be observed by (3.3), the number of observations in a neighborhood around any fixed t_0 is growing with T. Hence, we can think of an inherent "length of stationarity", which in rescaled time $u_0 = t_0/T$ asymptotically shrinks but in actual time is allowed to grow slowlier than the length T of the series. By this, it is possible to end up with a *unique* spectral density $f(u_0, \lambda) = |A(u_0, \lambda)|^2$ at a fixed time point based on infinitely many observations of the same kind at this fixed time. (More details on that can be found in [4], Section 2, and in [5], also).

Having defined the evolutionary spectrum as our underlying object of interest, we now turn to what we like to call a localized periodogram estimator of this object. It is a local version of the classical periodogram over a segment of length N of the tapered data $X_{t,T}, 1 \le t \le T$, with midpoint $[uT]$:

$$I_N(u, \lambda) = H_{2,N}^{-1} \mid \sum_{s=0}^{N-1} h(\frac{s}{N})\, X_{[uT-\frac{N}{2}+s+1],T}\, \exp(-2\pi i \lambda s)|^2, \tag{3.7}$$

where $h : [0, 1] \to [0, 1]$ is a sufficiently smooth taper–function ("window") and $H_{2,N} = \sum_{s=0}^{N-1} h^2(s/N)$ the appropriate norming factor with $H_{2,N} \sim N$ (see again [4]).
Using data–tapers, e.g. a Hanning–window which is of cosine form

$$h(u) = \left\{ \begin{array}{l} \frac{1}{2}(1 - (\cos 2\pi u)) \\ h(1-u) \end{array} \right\} \left\{ \begin{array}{l} u \in [0, \frac{1}{2}] \\ u \in [\frac{1}{2}, 1] \end{array} \right. , \tag{3.8}$$

is a well–known remedy in spectral estimation to reduce leakage effects, which occur in particular for spectra with a high dynamic range.
In practice, $I_N(u, \lambda)$ is calculated on possibly overlapping segments of $X_{t,T}$ of length $N = 2^J$: The shift from segment to segment is denoted by S, with $1 \le S \le N$, and the resulting number of segments M controls smoothing in u–direction. That is, we calculate $I_N(u, \lambda)$ at M timepoints

$$u_i = t_i/T, \quad \text{where } \; t_i = S \cdot i + N/2, \; 0 \le i \le M - 1,$$

with

$$T = S \cdot (M-1) + N\,.$$

This principle of segmentation is illustrated in Figure 1.

With respect to asymptotics, as will be found in the next section, we have to impose some additional assumptions which adapt the estimation to the asymptotically shrinking inherent "length of stationarity", as in Remark 3.4. For an asymptotically unbiased estimate of the evolutionary spectrum we basically need that $N^2/T \to 0$, as $T \to \infty$, but still has N to grow sufficiently fast. The precise rates can be found along with Assumption (A3) in the next Section 4.
Of course, as for classical periodograms of stationary time series, now the question arises how to smooth this wildly fluctuating estimate, moreover, how to do it *adaptively* over frequency *and* time. This is to be found in the next section on wavelet thresholding the localized periodograms.

4 Wavelet threshold estimates of the evolutionary spectrum

In this section we present our approach of how to use wavelet thresholding to adaptively smooth localized periodograms. We give asymptotic properties of the empirical wavelet coefficients and use these to show the basic Theorem 4.10 on the minimax L_2–risk of our resulting estimator of the evolutionary spectrum.

4.1 Denoising by non–linear thresholding

Denoising of curve estimates in general, and the periodogram as a spectrum estimator in particular, can be performed by applying *non–linear* thresholding techniques which were introduced by Donoho et al. (see [6], [7], e.g.). First theoretical investigations in the context of spectral density estimation for stationary time series can be found in [9] for *Gaussian* time series and in [14] for more general stationary processes.
Basically, these non–linear techniques are important to benefit also on the empirical side of estimation from a particular nice property of wavelets: They deliver sparse representations for curves with inhomogeneously distributed regularity. This can often be observed by spatially varying local structure of the curve. For spectra this would typically mean regions with sharp peaks followed by regimes of widely spread mode–like structure. As an example we like to refer to the simulations in Section 5.2
The general idea of non–linear thresholding is to set to zero, by the now common rules of soft or hard thresholders, those empirical wavelet coefficients which do not exceed a suitable chosen threshold $\lambda = \lambda_T$, where again T denotes the observed sample size. For hard thresholding, apply $\delta_\lambda^H(x) = x \cdot \mathbb{1}_{\{|x|>\lambda\}}$ to the empirical wavelet coefficients, for soft thresholding choose $\delta_\lambda^S(x) = sgn(x) \cdot (|x| - \lambda)_+$. With this, only those coefficients remain which are supposed to carry significant signal information. The right level of significance has to be delivered by an appropriate choice of the threshold λ_T, which in general can also depend on the resolution scale and location of the wavelet coefficients. In many approaches, its choice is motivated by ending up with a smooth estimator. The then

called *universal threshold* λ_T basically has to be proportional to the standard deviation of the empirical coefficients, plus some extra factor which protects against large deviations in their tail distribution. Whereas for data with a Gaussian noise structure this extra factor is of the form $\sqrt{const.\ \log T}$, for periodograms of Gaussian stationary time series Gao [9] showed that a choice of the form $const.\ \log T$ takes the heavier tails of the χ^2–distribution into account.
In the following we generalize the existing approaches to the 2–d situation of the time–varying spectrum and derive similar results.

4.2 Wavelet expansion of evolutionary spectrum and local periodogram

First, we give the (theoretical) wavelet decomposition of the projection $f_J(u,\lambda)$ of the spectrum $f(u,\lambda)$ onto the space $\mathbf{V}_J$ (cf. equation (2.1)):

$$f_J(u,\lambda) = c_{0\mathbf{0}} + \sum_{j=0}^{J-1} \sum_{\mathbf{k}=0}^{2^j-1} \sum_{\mu=h,v,d} d^{\mu}_{j\mathbf{k}} \Psi^{\mu}_{j\mathbf{k}}(u,\lambda) \tag{4.1}$$

with the coefficients

$$c_{0\mathbf{0}} \;=\; \int_0^1 \int_{-\frac{1}{2}}^{\frac{1}{2}} f(u,\lambda)\, du\ d\lambda \tag{4.2}$$

$$d^{\mu}_{j\mathbf{k}} \;=\; \int_0^1 \int_{-\frac{1}{2}}^{\frac{1}{2}} f(u,\lambda) \Psi^{\mu}_{j\mathbf{k}}(u,\lambda)\, du\ d\lambda \tag{4.3}$$

Here, we will identify $[0,1] \times [-\frac{1}{2},\frac{1}{2}]$ with $\mathbb{T}^2$, using the notation of Section 2, to avoid notational inconvenience. Note that in the simulation examples of Section 5 we consider only spectra which are periodic in time, too. Hence, there, we use the same periodic wavelet transform for both dimensions. However, in general, one would use a transform in time which is based on some wavelets adapted to a compact interval, see e.g. [3].

Now, we replace the unknown $f(u,\lambda)$ by a nonparametric estimate, the localized periodogram $I_N(u_i,\lambda), u_i = t_i/T, i = 0,...,M-1$, as introduced in the previous section. Note that in practice we have to choose M to equal N to be able to use a traditionally quadratic 2–d wavelet scheme.
The resulting empirical coefficients are

$$\check{c}_{0\mathbf{0}} = \frac{1}{M} \sum_{i=0}^{M-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} I_N(u_i,\lambda)\, d\lambda \tag{4.4}$$

and

$$\check{d}^{\mu}_{j\mathbf{k}} = \frac{1}{M} \sum_{i=0}^{M-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} I_N(u_i,\lambda) \Psi^{\mu}_{j\mathbf{k}}(u_i,\lambda)\, d\lambda\ . \tag{4.5}$$

Remark 4.1 Note, that in practice we do not calculate the coefficients according to (4.4) and (4.5). We start from a sample of N^2 data points $I_N(u_i,\lambda_n)$, sampled on an equally spaced grid $(u_i,\lambda_n), 0 \le i,n \le N-1$, with Fourier frequencies $\lambda_n = \frac{n}{N}$, and with $N = 2^J$, which determines the finest scale J. Then, we use the collocation projection described in

Section 2 by using the Lagrange function $S_J(u,\lambda)$ for both dimensions to calculate the empirical wavelet coefficients on the finest scale J.

$$\begin{aligned}
\tilde{c}_{J\mathbf{k}} &= \int_0^1 \int_{-\frac{1}{2}}^{\frac{1}{2}} \tilde{f}_J(u,\lambda)\,\varphi_{Jk_1}(u)\,\varphi_{Jk_2}(\lambda)\,du\;d\lambda \\
&= \int_0^1 \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=0}^{N-1}\sum_{n=0}^{N-1} I_N(u_i,\lambda_n)\,S_J(u-\frac{i}{N})\,S_J(\lambda-\frac{n}{N})\,\varphi_{Jk_1}(u)\,\varphi_{Jk_2}(\lambda)\,du\;d\lambda \\
&= \sum_{i,n} I_n(u_i,\lambda_n) L_J(k_1-i) L_J(k_2-n)\ . \qquad (4.6)
\end{aligned}$$

As an analog to equation (2.3) $\tilde{f}_J(u,\lambda)$ can be considered as an interpolated periodogram. The empirical coefficients $\check{d}_{j\mathbf{k}}$ and $\check{c}_{0\mathbf{0}}$ then are calculated from $\check{c}_{J\mathbf{k}}$ by the corresponding filters G and H (see Section 2). The form of the coefficients is similar to the one of equation (4.5), and hence they share the same statistical properties. However, the projection of the empirical sample onto $\mathbf{V}_J$ by an interpolatory spline quadrature is numerically more accurate than approximation of the inner product by simply a sum over equally spaced grid points (sometimes also called "Fourier interpolation"). This is important, in particular, for scales j close to the finest scale J. Here using Fourier interpolation would result into remainders which are of the same order as the leading terms. On the other hand, equation (4.6) does provide, however, the same degrees of freedom as do (4.4) and (4.5), because L_J is of order $O(2^{-J}) = O(N^{-1})$. Hence, in particular, bias and variance are of the same order as in equations (4.4) and (4.5). So, we use (4.4) and (4.5) merely for reasons of notational convenience to demonstrate the statistical principles in more clarity.

Figure 1 shows the following principle: The time–dependent covariance of the non–stationary series $\{X_{t,T}\}$ is transformed into a two–dimensional sample $I_N(u_i,\lambda_n)$ in the time–frequency plane and, subsequently, into the corresponding empirical wavelet coefficients. With this, the information in the time–dependent frequency content of $X_{t,T}$ is localized.

Remark 4.2 Obviously, the adaption properties of this estimate depend on the choice of the segment length N and the shift S, respectively: Once S or N are chosen, the maximal resolution of the periodogram estimate w.r.t. both time and frequency is fixed: The smaller N the better is the resolution in time, but the worse is the one in frequency, and vice versa. So, it is of further interest to study how this choice, too, can be done adaptively with wavelet–related methods. One possibility, which avoids a preliminary choice of a fixed N, has meanwhile been explored by the first author and can be found in [15]. Also, in meantime, an interesting work of Sudeshna Adak [1] came to our attention which addresses the problem of adaptively finding the appropriate segmentation by using a tree–structured search algorithm.
In practice, for our examples considered in Section 5, the choice of an intermediate S seems appropriate.

4.3 Statistical behavior of two–dimensional empirical coefficients

It is now our goal to investigate the statistical behavior of the empirical wavelet coefficients $\check{d}^{\mu}_{j\mathbf{k}}$. First, we give the precise regularity assumptions on both the spectrum $f(u,\lambda)$,

i.e. on $A(u,\lambda)$, and on the wavelet basis functions used.

Assumptions:
(A1) Let $A(u,\lambda)$ be differentiable in u and λ with uniformly bounded first partial derivatives.

(A2) Let $\Psi^u_{j\mathbf{k}}(u,\lambda)$ be differentiable in u and λ with uniformly bounded first partial derivatives.

(A3) The parameters N, S and T fulfill the relations (with the notation $'' \ll ''$ meaning "asymptotically smaller")

$$T^{\frac{1}{4}} \ll N \ll T^{\frac{1}{2}}/\ln T \quad \text{and} \;\; S = N \quad \text{or} \;\; S/N \to 0, \text{ as } T \to \infty.$$

(A4) The data-taper $h(x)$ is continuous on [0,1] and twice differentiable at all $x \notin P$ where P is a finite set and $\sup_{x \notin P} |h''(x)| < \infty$.

We give a discussion of these assumptions in Remark 4.5 further below.

Secondly, we state a central limit theorem which is one of the two keypoints for deriving the statistical behavior of the $\check{d}_{j\mathbf{k}}$ in *general*, i.e. non–Gaussian, situations whereas for a Gaussian time series, it is sufficient to investigate the asymptotics of the mean and the variance. The other keypoint will be the examination of the tail probability of the empirical wavelet coefficients which is to be found in the next subsection.

Theorem 4.3: *Suppose $\{X_{t,T}\}_{t=1,\ldots,T}$ is given by (3.1), and let Assumptions (A1) – (A4) be fulfilled. Then, as $T \to \infty$, uniformly over j,k with $2^j = o(N)$, the empirical coefficients $\check{d}^{\mu}_{j\mathbf{k}}$ are asymptotically normally distributed, i.e.*

$$\sqrt{T}(\check{d}^{\mu}_{j\mathbf{k}} - d^{\mu}_{j\mathbf{k}}) \xrightarrow[\mathcal{L}]{} \mathcal{N}(0, A^{\mu}_{j\mathbf{k}}), \;\; \mu = h, v, d,$$

where $d^{\mu}_{j\mathbf{k}}$ and $\check{d}^{\mu}_{j\mathbf{k}}$ are given by (4.3) and (4.5), respectively, and where

$$A^{\mu}_{j\mathbf{k}} = 2\, C_h \cdot \int_{I\!R^2} \{f(u,\lambda)\}^2 \, \Psi^{\mu}_{j\mathbf{k}}(u,\lambda) \left[\Psi^{\mu}_{j\mathbf{k}}(u,\lambda) \; + \Psi^{\mu}_{j\mathbf{k}}(u,-\lambda)\right] du \; d\lambda \qquad (4.7)$$

with $C_h = \int_0^1 h^4(x)dx / (\int_0^1 h^2(x)dx)^2$ for $S = N$ and $C_h = 1$ if $S/N \to 0$.

Proof: Theorem 4.3 is shown by straightforward application of [4], Theorem A.2, which holds regardless to the Gaussian assumption. For reasons of clarity (dealing with weight functions which depend on j and k, with $2^j = o(N)$), we give the detailed results, holding uniformly in $\mathbf{k}$, for bias, variance and the cumulants of higher order needed in the proof to make use of the method of cumulants:

(i) $$E\; \check{d}^{\mu}_{j\mathbf{k}} - d^{\mu}_{j\mathbf{k}} = O(2^{-j}N^{-1}) = o(T^{-\frac{1}{2}}) \quad \forall \mu = h, v, d$$

(ii) $$var\{\check{d}^{\mu}_{j\mathbf{k}}\} = \frac{A^{\mu}_{j\mathbf{k}}}{T} + O(\frac{2^j N}{T^2}) + O(2^{-j}T^{-1}) \quad \forall \mu = h, v, d$$

where the last term, which is the second part of the asymptotic variance as given in [4], Theorem A.2, is exactly zero for Gaussian time series.

(iii) $$T^{L/2} cum_L\{\check{d}^{\mu}_{j\mathbf{k}}\} = o(1) \quad \forall L \geq 3 \quad \text{if } 2^j = o(N).$$

■

Note that in the general, i.e. non–Gaussian, case we have, in addition, to assume finiteness of all cumulants of the underlying process, as given by Definition 3.1, (i).

We like to mention that in [14], Proposition 3.1, results analogous to the rates in (i) through (iii) are derived for the situation of a stationary, not necessarily Gaussian process.

However, we also show a result which is slightly stronger than (i) in the proof of Theorem 4.3:

Lemma 4.4: *Let $f(u, \lambda) \in C^{\sigma}(\mathbb{T}^2)$ with $\sigma \geq 1$. Then,*

$$E\ \check{d}^{\mu}_{j\mathbf{k}} - d_{j\mathbf{k}} = O(2^{-j}N^{-\sigma}) \quad \forall \mu = h, v, d \quad \textit{uniformly in } \mathbf{k}. \tag{4.8}$$

Proof: The proof runs analogously to the one of [4], Lemma (A.8) (with M = N), for a sufficiently smooth taper function ($\in C^{\sigma}$). Using estimates like

$$\int_{\mathbb{T}^2} |\Psi^{\mu}_{j\mathbf{k}}(u, \lambda)|\, du\ d\lambda\ = O(2^{-j})\,,$$

we conclude:

$$E\, \check{d}^{\mu}_{j\mathbf{k}} - d_{j\mathbf{k}} = O(2^{-j}\ \ln N\ N/T^{\sigma})\ +\ O(2^{-j}\, N^{-\sigma}).$$

■

Note that it is not straightforward to generalize this lemma on the decay of the bias for functions in $C^{\sigma}(\mathbb{T}^2)$ with $\sigma < 1$. However, adopting techniques used in [15] would allow to derive sufficiently fast rates for bias, and variance, also: We would get results for functions of lower regularity, basically being continuous and of bounded variation over $\mathbb{T}^2$ (cf. Remark 3.3).

Remark 4.5 We give the following heuristics for the motivation of (A3): With the periodogram over the first segment we estimate f at time $\frac{N}{2T}$. To conclude from this to f at the ends of the first segment and end up with a $\sqrt{T}$–consistent estimator, we need that $\frac{N}{\sqrt{T}} \to 0$. On the other hand the bias of the periodogram, with a data–taper, is of order $O(N^{-2})$ which leads to the condition $\frac{\sqrt{T}}{N^2} \to 0$. Note that without taper we have a bias of order $O(N^{-1})$, such that $\frac{N}{\sqrt{T}} \to 0$ cannot be fulfilled. Note also that the variance is <u>not</u> increased by the use of the taper if $S/N \to 0$. This seems to be heuristically clear since then, asymptotically, there are no more observations which are downweighted or totally excluded by tapering.

Corollary 4.6: *As a consequence of Theorem 4.3*

$$T \cdot var\ \check{d}^{\mu}_{j\mathbf{k}} \to A^{\mu}_{j\mathbf{k}}. \tag{4.9}$$

In particular, as $f(u,\lambda)$ is bounded from below and above by some positive constants K_1 and K_2,

$$\frac{K_1}{T} \leq var\ \check{d}^{\mu}_{j\mathbf{k}} \leq \frac{K_2}{T}\ , \quad uniformly\, in\, j\, and\, k. \tag{4.10}$$

Compare the similarity of the asymptotic variance of $\check{d}^{\mu}_{j\mathbf{k}}$ with the one–dimensional situation as in [9], Lemma 5.
Though, in this section we were quite general in treating also non–Gaussian processes, for the derivation of what follows for Gaussian processes, we only make use of equations (4.8) and (4.10).

4.4 Estimating the tail probability of the empirical wavelet coefficients

The goal of this section is to estimate the tail probability of $\check{d}^{\mu}_{j\mathbf{k}}$ in order to bound the maximum in probability of $\sup_{j,\mathbf{k}}(\check{d}^{\mu}_{j\mathbf{k}} - E\check{d}^{\mu}_{j\mathbf{k}})/(var\ \check{d}^{\mu}_{j\mathbf{k}})^{\frac{1}{2}}$. This will be used for proving our main Theorem 4.10, i.e. for an appropriate choice of the threshold λ_T in thresholding the empirical coefficients $\check{d}^{\mu}_{j\mathbf{k}}$.

In the sequel, we consider $\check{d}_{j\mathbf{k}} = \check{d}^{\mu}_{j\mathbf{k}}$ for a fixed $\mu \in \{h, v, d\}$. For simplicity, let us first restrict to non overlapping segments $S = N$, i.e. $T = NM$.
Let

$$\Gamma_T \ = \ Cov\ \{X_{t,T};\ X_{s,T}\}\,, \quad t,s = 1,\ldots,T$$

denote the $T \times T$ - covariance matrix of the Gaussian vector $\underline{X}_T = (X_{1,T},\ldots,X_{T,T})$.
Let further $\underline{V}^t_T = \Gamma^{\frac{1}{2}}_T$ and let $\underline{\xi}_T = \underline{V}^{-1}_T \underline{X}_T$. With $E\ \underline{\xi}^t_T \underline{\xi}_T = I_{T\times T}$, as $E\ \underline{X}^t_T \underline{X}_T = \underline{V}^t_T \underline{V}_T$, the elements of $\underline{\xi}_T$ are i.i.d. standard Gaussian random variables.

We use Parseval's relation for the inner product w.r.t. to frequency λ and write the empirical coefficients as a quadratic form in the observations $\underline{X}_T$:

$$\begin{aligned}
\check{d}_{j\mathbf{k}} &= \frac{1}{M}\sum_{i=0}^{M-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} I_N(u_i,\lambda)\, \Psi_{j\mathbf{k}}(u_i,\lambda)\, d\lambda \\
&= \frac{1}{M}\sum_{i=0}^{M-1} \psi^{(1)}_{jk_1}(u_i) \int_{-\frac{1}{2}}^{\frac{1}{2}} I_N(u_i,\lambda)\, \psi^{(2)}_{jk_2}(\lambda)\, d\lambda \\
&= \underline{X}^t_T\, B_T\, \underline{X}_T\,,
\end{aligned}$$

where B_T is a $T \times T$ - block–diagonal matrix with M blocks each of size N:

$$B_T \;=\; \Big[(\frac{1}{M}\,\psi^{(1)}_{jk_1}(u_i))\;(H_{2,N})^{-1}\widehat{\psi^{(2)}_{jk_2}}(t-t'))_{t,t'=0,\ldots,N-1}\Big]_{i=0,\ldots,M-1}\,.$$

Observe now, that the eigenvectors $\underline{\eta}_T = P_T\underline{\xi}_T$ of the Hermitian matrix $\underline{V}_T^t\,B_T\,\underline{V}_T$ are again i.i.d. standard Gaussian. Let κ_ν denote the corresponding eigenvalues (cf. [9], proof of Lemma 5). Hence,

$$\check{d}_{j\mathbf{k}} \;=\; \underline{X}_T^t\,B_T\,\underline{X}_T \;=\; \underline{\xi}_T^t\,\underline{V}_T^t\,B_T\,\underline{V}_T\,\underline{\xi}_T \;=\; \sum_{\nu=1}^{T}\kappa_\nu|\eta_\nu|^2\,. \tag{4.11}$$

Remark 4.7 For overlapping segments, i.e. $S < N$, hence $T < NM$, the rank T of the weight matrix B_T is smaller than NM, as the blocks are overlapping. Hence, one has to slightly modify this principle of orthogonal transformation.

As a consequence of (4.11), and by (4.10), the boundedness of $var\;\check{d}^{\mu}_{j\mathbf{k}}$, an assertion similar to that of [9], Lemma 6, on the tail probability of the empirical coefficients $\check{d}_{j\mathbf{k}}$ holds. Its proof runs completely analogously to the one of [9], Lemma 6.

Proposition 4.8:

$$\sup_{\{\kappa_\nu:\sum_{\nu=1}^{T}\kappa_\nu^2=1\}} P\,\{|\sum_{\nu=1}^{T}\kappa_\nu\cdot(|\eta_\nu|^2-1)|\geq x\} \;\leq\; const.\;\;\exp\{-\frac{x}{2}\}\,.$$

That is,

$$\sup_{0\leq j<J,\;0\leq k<2^j} P\left\{\left|\frac{\check{d}_{j\mathbf{k}}-E\;\check{d}_{j\mathbf{k}}}{(var\;\check{d}_{j\mathbf{k}})^{\frac{1}{2}}}\right|\geq x\right\} \;\leq\; const.\;\;\exp\{-\frac{x}{\sqrt{2}}\}\,,$$

where, by Theorem 4.3 and equation (4.9),

$$T\cdot var\;\check{d}_{j\mathbf{k}} \;\longrightarrow\; A^{\mu}_{j\mathbf{k}}.$$

Proposition 4.8 tells us that the tail probability of $\check{d}_{j\mathbf{k}}$ can be estimated from above by a χ^2_2–distribution (which is proportional to the exponential). Hence, an appropriate choice of the threshold λ_T should incorporate the fact that the noise in the empirical wavelet coefficients of the periodogram ordinates – for finite T – has a non–Gaussian character: it reaches high levels somewhat more frequently than a (white) Gaussian noise. Consequently, a higher threshold of order $\log T$ is needed which guarantees a sufficiently fast decay of the tail probability as $T\to\infty$:

$$P\left\{\left|\frac{\check{d}_{j\mathbf{k}}-E\;\check{d}_{j\mathbf{k}}}{(var\;\check{d}_{j\mathbf{k}})^{\frac{1}{2}}}\right|\geq \log T\right\} \;=\; o\;(T^{-1/2}).$$

The result of Proposition 4.8 is, somewhat surprisingly, independent of the assumption of stationarity as it totally parallels the one of the stationary situation (as in [9]).

4.5 The resulting wavelet threshold estimator

Now we use the asymptotic properties of the empirical wavelet coefficients, in particular those established by equations (4.8) and (4.10) and by Proposition 4.8, to derive the asymptotic properties of the non–linear wavelet estimation scheme.
The results of Sections 4.3 and 4.4 motivate the following threshold choice for the $\check{d}^{\mu}_{j\mathbf{k}}$:

$$\lambda_{T;j,\mathbf{k}} \;=\; \lambda_T \;=\; K \cdot T^{-\frac{1}{2}} \;\log T \;. \tag{4.12}$$

As preliminary upper estimate $\hat{K}$ for the constant $K > 0$ we propose

$$\hat{K} = 2\, C_h \cdot |f|_\infty = 2\, C_h \cdot \max_{0<u<1} \max_{-\frac{1}{2}\le\lambda\le\frac{1}{2}} |f(u,\lambda)| \;.$$

In practice, for a (slightly suboptimal) threshold, one needs an additional estimate for $|f|_\infty$, which can be estimated by replacing $f(u,\lambda)$ by a consistent estimate, e.g. a smoothed version of the periodogram, or even a *robust* version of that, as developped in [17]. This parallels the suggested use of a robust estimate of the noise level as in the classical situation of Gaussian i.i.d. data.

We consider the resulting *non–linear* estimate

$$\tilde{f}_T(u,\lambda) = \check{c}_{0\mathbf{0}} + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \sum_{\mu=h,v,d} \tilde{d}^{S\mu}_{j\mathbf{k}}\, \Psi^{\mu}_{j\mathbf{k}}(u,\lambda) \tag{4.13}$$

with soft–thresholding $\tilde{d}^{S\mu}_{j\mathbf{k}} = \delta^S_{\lambda_T}(\check{d}^{\mu}_{j\mathbf{k}})$, i.e. we choose the same threshold for all $\mu \in \{h,v,d\}$. For this estimate $\tilde{f}_T(u,\lambda)$ it is possible to show a result analogously to [9], Theorem 1, which treats the one–dimensional stationary situation. We cite a slightly more general version of [9], Theorem 1:

Let $\mathcal{F} = B^{\sigma}_{p,q}(C)$ be some ball in the Besov space $B^{\sigma}_{p,q}$ with either $p \ge 1$ and $\sigma > 1/p$, or $\sigma, p \ge 1$ for $f \in \mathcal{F}$ being of bounded variation. Let $\widehat{f}$ be the wavelet estimator based on thresholds $\lambda_T = 2\, \log(T)\, T^{-1/2}$. Then

$$\sup_{f\in\mathcal{F}} \{\mathsf{E}\, \|\widehat{f} - f\|^2_{L_2([-\pi,\pi])}\} \;=\; O\left(T^{-2\sigma/(2\sigma+1)}\, (\log(T))^2\right). \tag{4.14}$$

Observe that $r := 2\sigma/(2\sigma+d)$ is the power of the rate of convergence T^{-r}, which is optimal for d–dimensional non–parametric estimation. If $p < 2$, that is in the case of spatial variability (e.g. for functions of bounded variation), $r > r'$, where r' is the respective optimal rate for linear procedures (cf. [6], [7]). I.e. for $\mathcal{F}$ being a class of functions with inhomogeneous regularity, the *non–linear* estimate $\tilde{f}_T$ of *local* smoothing with threshold λ_T as in (4.12) is able to outperform *linear* ones with *global* smoothing at the level of the *rate of convergence*.

Here, we shall prove a result which, by some straightforward effort, can be generalized to more interesting function classes (including the ones of spatial inhomogeniety). In the following remark, we like to give a formulation of this more general result, which completely parallels [9], Theorem 1, now of course with $d = 2$ in $r = 2\sigma/(2\sigma+d)$.

Remark 4.9 *Let $\mathcal{F}$ be an appropriate smoothness class for functions in $L_2(\mathbb{T}^2)$ with smoothness σ. Let $\widehat{f}$ be the wavelet estimator based on thresholds $\lambda_T = K \log(T)\, T^{-1/2}$, with some appropriate positive constant K. Then*

$$\sup_{f\in\mathcal{F}} \{\mathsf{E} \,\|\widehat{f} - f\|^2_{L_2(\mathbb{T}^2)}\} = O\left(T^{-2\sigma/(2\sigma+2)} \,(\log(T))^2\right). \tag{4.15}$$

Appropriate function classes can be 2–d Besov classes with some additional smoothness assumptions like continuity and being of bounded variation over $\mathbb{T}^2$, or simply Sobolev or Hölder classes.

Note that this result fits into what might be expected for generalizing the one–dimensional situation to dimension d. For a proof of (4.15) the precise assumptions, generalizing our (A1) and (A2), and also the techniques of proving, can be found in [15]. Basically we have to derive new bias and variance expansions, parallelizing the ones of Lemma 3.1 in [15]. However, in order not to obscure the basic ideas with lengthy technical proofs, for the reader's convenience we restrict to derive the following version of our result.

Let $\Lambda^\sigma(C)$, $\sigma > 0$ being no integer, be the set of σ-Hölder functions on $\mathbb{T}^2$, with uniform Hölder bound C (if σ is an integer, we use instead Zygmund's definition, see [12].). For $\sigma = 1$ our following main theorem holds by Assumptions (A1), (A2) and (A4). For $\sigma > 1$ we assume in addition that

(A5) $\Psi^u_{j\mathbf{k}}(u,\lambda)$ are in $C^\sigma(\mathbb{T}^2)$ and have vanishing moments up to order $\sigma - 1$.

Theorem 4.10: *Assume (A1)–(A5) and let in addition $f(u,\lambda) \in \Lambda^\sigma(C)$ with $\sigma \geq 1$. Let $\lambda_T = K \cdot T^{-1/2}\log(T)$ as in (4.12), and let $\tilde{f}_T(u,\lambda)$ be based on soft thresholding with this threshold λ_T. Then*

$$\sup_{f\in\Lambda^\sigma(C)} E \,\|\, \tilde{f}_T - f \,\|^2_{L^2(\mathbb{T}^2)} \leq \tilde{C}\cdot T^{-r}(\log T)^2$$

with $r = 2\sigma/(2\sigma+2)$ and a constant $\tilde{C} = \tilde{C}(C,\sigma)$ which depends on C and σ only.

Proof: Following the lines of the proof of [9] Theorem 1, with

$$E \,\|\, \tilde{f}_T - f \,\|^2 = E(\check{c}_{0\mathbf{0}} - c_{0\mathbf{0}})^2 + \sum_\mu \sum_{j=0}^{J-1} \sum_{\mathbf{k}=0}^{2^j-1} E(\tilde{d}^{S\mu}_{j\mathbf{k}} - d^\mu_{j\mathbf{k}})^2 + \sum_{j\geq J}^{\infty} \sum_{\mathbf{k}=0}^{2^j-1} (d^\mu_{j\mathbf{k}})^2,$$

we have to bound the three terms of this sum, making use of equations (4.8) and (4.10):
a)

$$E(\check{c}_{0\mathbf{0}} - c_{0\mathbf{0}})^2 = Var\,\check{c}_{0\mathbf{0}} + (E\check{c}_{0\mathbf{0}} - c_{0\mathbf{0}})^2 \leq \frac{C_1}{T} + \frac{C_2}{N^{2\sigma}} \leq \tilde{C}\cdot T^{-r},$$

where C_1 and C_2 depend on C and σ only.
(Note that the results of Theorem 4.3 and Lemma 4.4 do hold for $\check{c}_{0\mathbf{0}}$, also).
b)

$$d^\mu_{j\mathbf{k}} \leq C\cdot 2^{-j/2}2^{-(\sigma+\frac{1}{2})j} \leq C\cdot 2^{-(\sigma+1)j} \;\; \forall\mu, \text{ uniformly in } \mathbf{k},$$

by standard estimation of the decay of the wavelet coefficients (note that the "second dimension" only contributes by a factor of order $O(2^{-j/2})$ if one of the basis functions is a scaling function φ).
Hence,

$$\sum_{j\geq J}\sum_{\mathbf{k}=0}^{2^j-1}\left(d^{\mu}_{j\mathbf{k}}\right)^2 = O(2^{-2J\sigma}) = O(N^{-2\sigma}) = O(T^{-r}).$$

c)

$$\sum_{j=0}^{J-1}\sum_{\mathbf{k}=0}^{2^j-1} E(\tilde{d}^{S\mu}_{j\mathbf{k}} - d^{\mu}_{j\mathbf{k}})^2 \leq 2\sum_j\sum_{\mathbf{k}} E(\delta_{\lambda_T}(\check{d}^{\mu}_{j\mathbf{k}}) - E\check{d}^{\mu}_{j\mathbf{k}})^2 + 2\sum_j\sum_{\mathbf{k}}\left(E\check{d}^{\mu}_{j\mathbf{k}} - d^{\mu}_{j\mathbf{k}}\right)^2.$$

With Lemma 4.4, for $\sigma \geq 1$,

$$E\ \check{d}^{\mu}_{j\mathbf{k}} - d^{\mu}_{j\mathbf{k}} = O(2^{-j}N^{-\sigma})\ \ \forall\mu,\ \text{uniformly in}\ \mathbf{k}.$$

Hence, by this explicit bound on each of its terms, the second sum is bounded by some $O(N^{-2\sigma}\log N) = o(T^{-r})$.

As the bias $E\ \check{d}^{\mu}_{j\mathbf{k}} - d^{\mu}_{j\mathbf{k}}$, $j < J = \log(N)$, tends faster to zero than the coefficient $d^{\mu}_{j\mathbf{k}}$ we can proceed quite analogously to the proof of [9], Theorem 1. In particular, we use the same scheme of upper estimates leading to [9], equation (2.41).

$$\sum_j\sum_{\mathbf{k}} E(\delta_{\lambda_T}(\check{d}^{\mu}_{j\mathbf{k}}) - E\check{d}^{\mu}_{j\mathbf{k}})^2$$

$$\leq \tilde{C}\cdot(\log T)^2\sum_j\sum_{\mathbf{k}}((E\check{d}^{\mu}_{j\mathbf{k}})^2 \wedge \frac{C'}{T})\,(1+o(1))$$

$$\leq \tilde{C}\cdot(\log T)^2\sum_j\sum_{k}((d^{\mu}_{jk})^2 \wedge \frac{C'}{T})\,(1+o(1))$$

$$\leq \tilde{C}\cdot T^{-r}\ (\log T)^2\,(1+o(1))\ .$$

In this final step Proposition 4.8 is used: The first inequality, analogously to equation (2.30) in Theorem 3 of [9], is based on an exponential inequality for the tail probability of the standardized empirical wavelet coefficients, as given by Proposition 4.8. We like to refer to Section 5.5 of [9] (Lemmas 1–3 and Theorem 3) for details of this technique, which transfers results of Donoho and Johnstone, [6], [7], to the periodogram situation. ■

Remark 4.11 An alternative approach to non–linear smoothing of a (localized) *periodogram* estimate would be to apply the very same techniques to the *log–periodogram* instead. It is well known that the logarithmic transformation stabilizes the variance. Moreover, it can be shown by transferring results of von Sachs [18] for the tapered log–periodogram, that, for the one–dimensional stationary situation, the resulting empirical wavelet coefficients asymptotically follow a normal law, for $2^j = o(N)$:

$$\sqrt{N}\int_{-\frac{1}{2}}^{\frac{1}{2}}\{\log I_N(\lambda) + \gamma - \log f(\lambda)\}\,\Psi_{jk}(\lambda)d\lambda\ \overrightarrow{\mathcal{L}}\ N(0,\frac{\pi^2}{6})$$

where $\gamma = 0.57721$ ... (i.e. Euler's constant).
That is, the asymptotic variance does no longer depend on the unknown $f(\lambda)$, and it is proportional to $O(N^{-1})$, with a constant not depending on j and k.

One could use this result for our two–dimensional problem, too, as we observe that, using the (one–dimensional) collocation wavelet transform introduced in Section 2, *all* resulting coefficients have a variance proportional to N^{-1} for all scales, and the ones with $2^j = o(N)$ are asymptotically normal. Hence, for our simulation example in the following Section 5 we use the log–periodograms, and it seems that the results confirm our conjectures.

5 Applications and Simulations

First, we describe a typical situation in practice for a time–dependent power spectrum (motivated by [10], Section 1).

5.1 An example from mobile radio propagation

A microwave radio signal transmitted between a fixed base station and a moving vehicle in a typical urban environment (see Figure 2) exhibits extreme variations in both amplitude and apparent frequency:
From the viewpoint of an observer on the mobile unit, the received signal, a plane wave of the form

$$C \cdot \cos(2\pi\, \lambda(\alpha)t \,+\, \phi)$$

(a superposition of many of those plane waves, to be specific), may be represented as a carrier with randomly varying phase ϕ, amplitude C and frequency $\lambda(\alpha)$ (with randomly varying α). Due to the Doppler shift, caused by the movement of the mobile unit with velocity V into direction α w.r.t. the sender station, the frequency

$$\lambda(\alpha) = \lambda_c \,+ \lambda_m \cos\, \alpha,$$

is to be found in a narrow band around the carrier frequency λ_c. This band is (for $\alpha = 0$) of maximum width $2\lambda_m = 2V/\nu$, with ν being the wavelength of the transmitted carrier frequency (see Figure 3).
A suitable model for the three field components of the signal (electric field E_z, magnetic field H_x and H_y) is a Gaussian random process, *stationary* as long as V (and ν) do not depend on time t. Instead of determining the statistical properties of this Gaussian random process from its moments, they are most easily obtained from the power spectrum (as the Fourier transform of the autocorrelation of the signal components).
As simplest model, the probability distribution $p(\alpha)$ of the power over the angle α is assumed to be constant. Hence, for studying the H_x–field component of the signal as typical example only, we end up with a spectrum of the following form, by [10],

$$f_{H_x}(\lambda) \sim \left[1 - \left(\frac{\lambda - \lambda_c}{\lambda_m}\right)^2\right]^{\frac{1}{2}},$$

where we used that $|d\lambda| = \lambda_m| - \sin\alpha||d\alpha| = (\lambda_m^2 - (\lambda - \lambda_c)^2)^{\frac{1}{2}}|d\alpha|$ (for details, see [10], Section 1.2.1).

As long as V is constant in time, $f_{H_x}(\lambda)$ is also. But, in practice, of course, the mobile unit changes its velocity: hence, a more realistic model would be to allow for a time dependent power spectrum

$$f(t,\lambda) = f_{H_x}(\lambda(t)) = \left[1 - \left(\frac{\lambda - \lambda_c}{\lambda_m(t)}\right)^2\right]^{\frac{1}{2}} ,$$

where $\lambda_m(t) = V(t)/\nu$.
Consequently, the model for the underlying Gaussian random process becomes *instationary.*
An additional modification arises if we allow for a changing environment of the transmitting channel, i.e. an explicit variation of $f(t,\lambda)$ in t, too.

5.2 A simulation study

Motivated by Section 5.1, as example for our simulation we choose a modification of f_{H_x}, an evolutionary spectrum of the following form

$$\begin{aligned} f(u,\lambda) &= [1 - (\frac{\lambda}{\lambda_m})^2]^{\frac{1}{2}} \cdot \mathbb{I}_{[0,\lambda_m]}(\lambda) + [1 - (\frac{\lambda - \lambda_0}{\lambda_m})^2]^{\frac{1}{2}} \cdot \mathbb{I}_{[\lambda_0 - \lambda_m, \lambda_0 + \lambda_m]}(\lambda) + \\ &+ P_1 \cdot \exp\{-\frac{(\lambda - \lambda_1)^2}{2\sigma_1^2}\} + R_0 , \end{aligned} \tag{5.1}$$

where w.l.o.g. we assume $\lambda \geq 0$ due to symmetry, and where $\lambda_m = \lambda_m(u) = \sigma_0\ (2 + \cos 2\pi\nu_0 u)$ with $\sigma_0 = 0.03$, $\nu_0 = 4$, $\lambda_0 = 0.3$, $P_1 = 0.2$, $\lambda_1 = 0.45$, $\sigma_1 = 0.001$ and a constant $R_0 = 0.001$ (resulting as the spectral component of a background white noise component in the underlying process). Note that in addition to a time–dependent $\sigma = \sigma(u)$, which is motivated by a changing velocity V of the mobile unit in Section 5.1, $f(u,\lambda)$ has isolated singularities in its first partial derivative w.r.t. λ and a high dynamic frequency range of smoother and sharper components.

The evolutionary spectrum in equation (5.1) defines a locally stationary process $X_{t,T}$ according to (3.1), with $f(u,\lambda) = |A(u,\lambda)|^2$. For its simulation we generate $T = 2048$ data, using the following discretization (in λ) of the integral in (3.1):

$$X_{t,T} \;=\; 2^{\frac{1}{2}}\ T_s^{-\frac{1}{2}} \sum_{k=0}^{T_s-1} A(\frac{t}{T}, \frac{k}{T_s})\ \exp(2\pi i\ t\ \frac{k}{T_s})\ \xi_k\ , \quad 1 \leq t \leq T,$$

where $T_s = 8192$ and where $\xi_k,\ 0 \leq k \leq T_s - 1$, is a simulated Gaussian white noise $(\sim \mathcal{N}(0,1))$, generated by a standard pseudo random generator.

Further, we calculate the short–time periodogram over $M = 128$ segments of length $N = 128$, with shift $S = 15$, using a data–taper as given by (3.8). Note that, in order to use a quadratic two–dimensional MRA, here we have to choose $M = N$. Finally, we take logarithms, in order to benefit from the variance stabilizing effects of the logarithmic transformation.
For the wavelet basis used we choose the orthogonal (periodized) splines of the Battle–Lemarié–family (as indicated in Section 2) with order $m = 6$, i.e. the functions are

elements of $C^4(I\!R^2)$ and, hence, fulfill assumption (A2).
Figure 4 shows the true log–spectrum, the raw and and the smoothed log–periodogram for example (5.1), as grey scaled image plots, in the time–frequency plane. The local smoothing was performed by hard–thresholding with $\lambda_T = K\, T^{-\frac{1}{2}}\ \log T = 1 \cdot 10^{-2}$, i.e. $K = 4 \cdot 10^{-2}$ as $T = 2048$. For comparison we add an example of global smoothing, where we consider a linear wavelet estimator which incorporates all coefficients below a cut-off scale $J_0 = 5$.
It can be clearly observed that the noise in the periodogram–estimator is suppressed by non–linear thresholding without losing local structure of $f(u, \lambda)$ (e.g. the bump at λ_1), whereas with global smoothing this is not possible simultaneously.
Cuts in λ– and in u–direction (Figure 5) confirm this behavior.

A second time–dependent spectrum arises by simply replacing the cosine function $\lambda_m(u)$ in $f(u, \lambda)$ by some periodically piecewise linear one:

$$\lambda_m(u) \ = \ \sigma_0 \left(2 \ + \ 10/3 \cdot (r(\nu_0 u) - 0.7)\, I\!I_{[0.4,1)}(\mathrm{r}(\nu_0 \mathrm{u})) \ - I\!I_{[0,0.4]}(\mathrm{r}(\nu_0 \mathrm{u}))\right), \qquad (5.2)$$

where $r(x) := \ x - [x]$.
This function describes perhaps a slightly more realistic dependence of the velocity V of the mobile unit of Section 5.1 on time. It introduces an example which is less regular in time, and so it might be interesting to compare the performance of the same estimators as before (with same parameters), which, together with the true spectrum, are shown in Figure 6.

Further examples can be found in a more comprehensive version of this work [19], which is accessible via anonymous ftp at ftp://www.mathematik.uni-kl.de/pub/Math/Papers/AGTM-reports/ or at the web–site htpp://www.mathematik.uni-kl.de/.

ACKNOWLEDGEMENTS

We would like to thank R. Dahlhaus, J. Fröhlich and M. Neumann for encouraging discussions and helpful suggestions, showing a continuous interest and support of our work. Parts of the results of this paper have been presented in a talk titled "Non –linear wavelet methods for locally stationary time series", given by the first author on the IMS/Bernoulli Society meeting, Chapel Hill, on June 20, 1994.

References

[1] S. Adak, Time–dependent spectral analysis of nonstationary time series, Technical report No. 478, Department of Statistics, Stanford University, 1995.

[2] D. R. Brillinger, "Time Series: Data Analysis and Theory", Holden–Day, San Francisco, 1981.

[3] Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the Interval and Fast Wavelet Transform. *Appl. Comp. Harmonic Anal.* **1**, 54-81.

[4] R. Dahlhaus, Fitting time series models to nonstationary processes, Preprint "Beiträge zur Statistik", 4, Universität Heidelberg, 1993.

[5] R. Dahlhaus, On the Kullback–Leibler information divergence of locally stationary processes, Preprint "Beiträge zur Statistik", 27, Universität Heidelberg, 1995, to appear in Stoch. Proc. Appl.

[6] D. Donoho, I. Johnstone, G. Kerkyacharian, D. Picard, Wavelet shrinkage: Asymptopia? (with discussion) *J. Royal Stat. Soc., Ser. B* **57** (1995), 301–369.

[7] D. Donoho, Wavelet shrinkage and Wavelet -Vaguelette Decomposition: a 10–minute tour, *in* "Progress in Wavelet Analysis and Applications", pp. 109–128, Proc. Int. Conf. "Wavelets and Applications", Toulouse, June 1992.

[8] J. Fröhlich, K. Schneider, An adaptive Wavelet Galerkin Algorithm for one and two dimensional flame computations, *Eur. J. Mech., B/Fluids* **13**, 4 (1994), 439–471.

[9] H.–Y. Gao, Wavelet estimation of spectral densities in time series analysis, Ph. D. dissertation, U.C. Berkeley, 1993.

[10] W. C. Jakes, Mobile Radio Propagation, *in* "Microwave Mobile Communications", Ch. 1, Wiley, New York, 1974.

[11] W. Martin, P. Flandrin, Wigner–Ville spectral analysis of nonstationary processes, *IEEE Trans. Signal Processing* **33** (1985), 1461–1470.

[12] Y. Meyer, "Ondelettes et opérateurs I", Herman, Paris, 1990.

[13] P. Moulin, Wavelet thresholding techniques for power spectrum estimation, Preprint.

[14] M. H. Neumann, Spectral density estimation via nonlinear wavelet methods for stationary non–Gaussian time series, Preprint, WIAS Berlin, 1994.

[15] M. H. Neumann, R. von Sachs, Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra, Technical Report "Berichte der AG Technomathematik", **132**, Universität Kaiserslautern (1995).

[16] V. Perrier, C. Basdevant, La décomposition en ondelettes périodiques, un outil pour l'analyse de champs inhomogènes. Théorie et algorithmes, *Rech. Aérosp.*, 53–67, 1989–3.

[17] R. von Sachs, Peak–insensitive nonparametric spectrum estimation, *J. Time Series Analysis* **15**, 4 (1994), 429–452.

[18] R. von Sachs, Estimating non–linear functions of the spectral density, using a data–taper, *Ann. Inst. Stat. Math.* **46** (1994), 453–474.

[19] R. von Sachs, K. Schneider, Wavelet smoothing of evolutionary spectra by non–linear thresholding, Technical Report, Universität Kaiserslautern (1995).

[20] G. Walter, A sampling theorem for wavelet subspaces, *IEEE-IT* **38** (1992), 881–884.

Captions:

Figure 1: scheme of transformation:
$X_{t,T} \longrightarrow I_N(u_i, \lambda)$, local periodogram over segments of length N with shift $S = N/2 - 1$ (demonstration of localization principle, for $i = 0, ..., 4$)

$I_N(u_i, \lambda_n), 0 \leq i, n \leq N - 1$, in the time–frequency plane $\longrightarrow$ two–dimensional wavelet coefficients

Figure 2: scheme of signal transmission between mobile vehicle and fixed base station in the radio propagation example of Section 5.1

Figure 3: Frequency spectrogram of RF signal of Figure 2 at $\lambda_c = 910$ Mhz

Figure 4: Grey scaled image plot for example as in equation (5.1):
$\log f(u, \lambda)$,
$\log I_N(u, \lambda), N = 128, S = 15$,
of globally smoothed log–periodogram ($J_0 = 5$),
of locally smoothed log–periodogram (hard thresholding, $\lambda_T = 1 \cdot 10^{-2}$),
in the time–frequency plane

Figure 5: cuts of Figure 4 in frequency direction at $u = 0.54$
and in time direction at $\lambda = 0.24$

Figure 6: Grey scaled image plot of second example as in equation (5.2)