MODELLING AND ESTIMATION OF THE TIME-VARYING
STRUCTURE OF NONSTATIONARY TIME SERIES

BY
RAINER VON SACHS

TECHNICAL REPORT NO. 503
SEPTEMBER 1996

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS95 05151

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Modelling and Estimation of the Time-Varying Structure of Nonstationary Time Series

By
Rainer von Sachs

Technical Report No. 503
September 1996

Prepared Under the Auspices

Of
National Science Foundation Grant DMS95-05151

Department of Statistics
Stanford University
Stanford, California
Modelling and Estimation of the Time–varying Structure of Nonstationary Time Series

Rainer von Sachs
August 31, 1996

Statistics, Stanford University

on leave from: FB Mathematik, Universität Kaiserslautern
D-67653 Kaiserslautern, Germany
e-mail: rvs@mathematik.uni-kl.de

Abstract

Recently, quite a few new statistical models have been added to the variety of existing ones which try to capture the time–varying structure of nonstationary stochastic processes. Examples are numerous and can be found, e.g., in economics, speech and sound processing and the biomedical sciences.

This paper gives an overview of different modelling approaches, e.g. locally stationary time–frequency representations or local time–scale decompositions. It also outlines some of the statistical procedures that estimate characteristic second-order quantities of these stochastic processes, e.g. local autocovariances or local spectral densities. Among those, we will find spectograms, Wigner–Ville estimators, localized periodograms as estimates of evolutionary spectra, methods based on local cosine packets and wavelet spectrum estimates. We discuss consistent estimation and adaption of the estimators to the actual time–frequency or time–scale content of the time series and also present some data examples.

1 Introduction

Many phenomena in the applied science (e.g., biomedicine, electrical and acoustical engineering, geophysics, and economics) show an instationary behavior over time. Hence, time series data collected from these signals, should not be modeled by a stationary random process but rather by some approach which tries to capture the time–varying second order structure of these data, i.e. an autocovariance function which is no more simply a function of time–lag, or a time–dependent spectrum, respectively.

There have been several attempts in the last couple of decades to introduce various concepts of time–dependent power spectra, or more general, time–frequency representations of an instationary signal. Among those the Wigner–Ville spectrum plays a prominent role (see, e.g., [MaFl]), which more generally can be considered as one particular member of a general class of time–frequency distributions, the so–called Cohen’s class. For a short good overview we refer to [Adak]. Different approaches try to either generalize from parametric models, like AR– or ARMA–processes, see e.g. [DNvS], and introduce a time dependency
on the parameters of the model. Or they directly generalize from the spectral representation of a second–order stationary time series, as, e.g. [Pr] and [Da].

We will use the latter one, and related concepts of quasi or locally stationary processes, in this article to discuss various models of this kind which will provide a framework for rigorous statistical theory of how to estimate the time–dependent second–order structure of a locally stationary time series. For this some concept of restricting the departure from stationarity of the time series is needed, for which basically the notion of a slow change of its spectral characteristics is introduced. We like to formalize this a bit now.

1.1 The problem of nonstationary spectral analysis

Let \( X_t, 1 \leq t \leq T \), denote an observed segment of a zero mean process \( \{X_t\} \). Let \( \Gamma = (\gamma_{t,s}) \) denote the covariance matrix of \( X_t \), i.e. \( \gamma_{t,s} = EX_tX_s \).

If the process \( \{X_t\} \) is stationary, then we can estimate this covariance \( \gamma_{t,s} = \gamma(t – s) \), or its Fourier transform, the spectrum, from one single realizations of \( X \), a vector of length \( T \). In general, however, without further assumption on the time–varying spectrum of a nonstationary time series this is no more possible, hence \( n > 1 \) replications of the set of observed data \( X_1^{(i)}, \ldots, X_T^{(i)} \), \( 1 \leq i \leq n \), are needed. Then one would use the empirical covariance estimate

\[
C = n^{-1} \sum_{i=1}^{n} X^{(i)}X^{(i)'} ,
\]

which is fairly good, if \( n \gg T \) (for asymptotics, \( T \) would be fixed and \( n \) tend to infinity).

In the setting of the applications described above, however, one has often \( T \) very large – think of realizations of a voice signal or a digital communications signal where \( T \) can be thousands or millions – while \( n \) may be quite small – dozens or hundreds, or even just one. In this setting, the empirical covariance is useless, and one has to assume that the process, though nonstationary, is slowly changing in its spectral characteristics. This restriction to local or quasi–stationary time series allows the estimation of the time–dependent spectrum from a single record of the process. A slowly changing process is, roughly speaking, a time series for which at each time point, an interval of stationarity can be defined, within which the covariance is stationary or at least approximately so.

1.2 Outline of the paper

In Section 2 we will discuss this in more detail by focusing onto one of these approaches, the model of locally stationary processes with evolutionary spectra, as introduced by Dahlhaus, see [Da]. Whereas in general, there is no notion of a unique time–frequency representation, this latter approach allows for some (asymptotically) uniquely defined time–dependent spectrum. Hence a framework for rigorous estimation theory is provided.

We will also shortly present alternative models by Mallat, see [MPZ], by [MHK] and in the work of [DoVsS], where in particular the latter one is kind of a reformulation of the Dahlhaus' model by imposing local stationarity on the sequence of covariances \( (\gamma_{t,s}) \) instead.

In Section 3 we will discuss various estimation procedures which, under this (or a related) model of local stationarity, have some consistency property, often even some optimality, like a (near–) optimal rate of \( L_2 \)–risk between the estimator of a time–dependent (i.e. evolutionary) spectrum and this theoretical quantity (see Section 3.1 and 3.3). Hence, with this we provide some criteria of how to measure and compare the performance of these
estimation procedures which certainly are related to what has been existing since longer in
the engineering literature as spectograms (i.e. short-time or windowed Fourier transforms)
or (smoothed) Wigner Estimates. Also, some (simulated) data examples are shown to
emphasize the usefulness of the estimation algorithms.

In Section 4, then, we will address some ideas of time–scale models as useful alternatives
to the time–frequency ideas previously discussed. These express a different idea of a local
autocorrelation analysis where it might be more appropriate to think of a decomposition
of the nonstationary time series into blocks which are both localized in time and in scale
(or “wave”) length. For this a (stochastic) wavelet representation is some obvious example,
which was introduced by the work of [vSNK]. Again ideas of local stationarity leading to
an evolutionary wavelet spectrum are used to provide some (asymptotic) estimation theory
([vSNK2]). Finite sample performance of the estimator is demonstrated by some biomedical
data set which shows a particular nonstationary behavior.

In order to provide the necessary background on wavelets and their usefulness with adaptive
smoothing of the considered time–frequency and time–scale estimators, we now include a
short section on denoising by non–linear wavelet thresholding.

1.3 Denoising by non–linear wavelet thresholding

The basic idea of using wavelets for statistical denoising is to exploit the adaptation
properties in an expansion of a function into a series of localized basis functions. Those
are chosen to be an orthonormal basis of $L_2([a, b])$, for some interval $[a, b]$ with $-\infty \leq a < b \leq \infty$. Generalizations to higher dimensional situations, as in Section 3, are
possible. Moreover, under certain additional assumptions on the wavelets (like vanishing
moments), the resulting representation is very sparse, in particular even for curves with
inhomogeneously distributed regularity, and adapts to the local behavior of the underlying
function, i.e. to discontinuities, like cusps or jumps, e.g. This is in contrast to, e.g., a
Fourier basis representation. 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 spectrum show in Section 3, Figure 4. To benefit from this property also on the
empirical side of estimating in a noisy environment the function $f(x)$, which is assumed to
be member of a certain function class (Hölder, Sobolev, Besov,...), but unknown otherwise,
one has to non–linearly manipulate the empirical coefficients, build from the observed data
of sample size $N$.

Here, we use the following terminology: $\{\varphi_{j_0 k}\}_k \cup \{\psi_{j k}\}_{j \geq j_0, k}$ denotes the wavelet basis,
where the wavelets $\psi_{j k}(x) = 2^{j/2} \psi(2^j x – k), j, k \in \mathbb{Z}$, are dilated and translated versions
of a given localized $\psi(x)$ (“mother wavelet”). So are the scaling functions $\varphi_{j_0 k}$ on the lowest
resolution scale $j_0$, where $\varphi(x)$ measures the overall trend (“low frequency”) behavior of the
function $f(x)$. The wavelet coefficients, $d_{j k} = \int f(x) \psi_{j k}(x) \, dx$ reflect the details (“band
pass behavior”) of $f(x)$, which then writes as

$$f(x) = \sum_k c_{j_0 k} \varphi_{j_0 k}(x) + \sum_j \sum_k d_{j k} \psi_{j k}(x) .$$

On the empirical side, e.g. for the signal–in–noise model $Y_i = f(x_i) + \epsilon_i, i = 1, ..., N$,

$$\hat{d}_{j k} \sim N^{-1} \sum_{i=1}^N Y_i \psi_{j k}(x_i) ,$$

3
where in practice however, fast filter algorithms (Fast Wavelet Transform) are used to calculate these empirical coefficients $\tilde{c}_{jk}$ on a dyadic hierarchy. For more details on this, we refer to [Dau] and [Do], e.g.

Denoising of curve estimates in general can be performed by applying techniques which were introduced by Donoho et al. (see, e.g., [Do], [DJKP]) and which non-linearly threshold the empirical coefficients $\tilde{c}_{jk}$. First theoretical investigations in the context of spectral density estimation for stationary time series can be found in [Gao] for Gaussian time series and in [Neu] for more general stationary processes. 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 $T$ denotes the observed sample size. For hard thresholding, apply $\delta^\lambda_T(x) = x \cdot 1_{(|x| > \lambda)}$ to the empirical wavelet coefficients, for soft thresholding choose $\delta^\lambda_S(x) = \text{sgn}(x) \cdot (|x| - \lambda)_+$. With this, only those coefficients remain which are supposed to carry significant signal information, and these are used for the reconstruction by inverse wavelet transformation. The right level of significance has to be delivered by an appropriate choice of the threshold $\lambda_T$, 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 $\lambda_T$ 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 $\sqrt{\text{const.} \log T}$, for periodograms of Gaussian stationary time series a choice of the form const. $\log T$ takes the heavier tails of the $\chi^2$-distribution into account. Also, in this context, a level-dependent threshold is often needed, i.e. the constant varies from scale to scale, as in Section 3.3. The resulting non-linear threshold estimators do not only provide local smoothers, but in very many situations (even for heteroscedastic correlated noise) achieve the near-minimax $L_2$-rate for the risk of estimation.

2 Locally stationary time-frequency models

For later purpose of estimation theory (see Section 3.1 and 3.3), here we somehow put more emphasis on the approach of [Da], which we explain in the first subsection. However, in the second subsection, we also shortly mention alternative models, i.e. the ones of [MPZ], [MHK] and [DoMvS], and also the models of [Adak], motivated by piecewise stationary processes, and the time-varying autoregressive model of [DNvS], which is closely related to (actually a special case of) the Dahlhaus model of locally stationary processes.

2.1 Dahlhaus locally stationary processes and evolutionary spectra

The approach of [Da] is a straightforward generalization of the Cramér representation of a stationary stochastic process (see [Pr2], e.g.). This approach defines a sequence of doubly-indexed processes as follows:

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

$$X_{t,T} = \mu(t) + \int_{-\pi}^{\pi} A^0_{t,T}(\omega) \exp(i\omega t) \, d\xi(\omega),$$

(1)
where

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

(ii) there exists 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 $T$,

$$\sup_{t, \omega} |A^0_{t,T}(\omega) - A(t/T, \omega)| \leq KT^{-1}. \quad (2)$$

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

The asymptotics are based on rescaling in time–location which allows asymptotic inference starting from a single realization of $\{X_{t,T}\}$. This is because the smoothness of $A$ in $u$ controls the variation of $A^0_{t,T}(\omega)$ as a function of $t$ so that it can only vary slowly enough. Basically the idea is that one implicitly assumes some local interval of stationarity which determines this variation. This neighbourhood becomes asymptotically arbitrarily small in the rescaled time $u$, or, to say it the other way round, in actual time $t$ it gets asymptotically larger but on a slower rate than the length $T$ of the whole time series (cf. Section 3.1 below). Estimation theory then parallels the one of nonparametric regression with an asymptotically denser and denser design on $(0,1)$.

There are a variety of possibilities for exact smoothness assumptions on $A(u, \omega)$: Basically bounded partial derivatives in both time and frequency are needed, but without too much effort these can be relaxed quite a bit, to some slightly stronger than continuity and being of bounded variation, see [NvS].

This model also allows to define a unique underlying time–varying spectrum of $\{X_{t,T}\}$, as follows.

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

$$f_T(u, \omega) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} \text{Cov}\{X_{[uT-s/2],T}; X_{[uT+s/2],T}\} \exp(-i\omega s), \quad (3)$$

where the $X_{t,T}$'s are given by (1), with $A^0_{t,T}(\omega) = A(0, \omega)$ for $t < 1$ and $A^0_{t,T}(\omega) = A(1, \omega)$ for $t > T$. This quantity, for fixed $T$, is similar to the so-called Wigner–Ville spectrum (see, e.g., [MaFl]).

Then, by the following (4) and (5), $f_T(u, \omega)$ will be related to the smooth amplitude function $A(u, \omega)$, which defines the "evolutionary spectrum":

**Definition 2.2** The evolutionary spectrum of $\{X_{t,T}\}$ given in (1) is defined, for $u \in (0,1)$, by

$$f(u, \omega) := |A(u, \omega)|^2. \quad (4)$$

This $f(u, \omega)$ is the limit of $f_T(u, \omega)$ as $T \to \infty$, in general in some mean–square sense as shown in [NvS], Theorem 3.1:

$$\int_{0}^{1} \int_{-\pi}^{\pi} |f_T(u, \omega) - f(u, \omega)|^2 d\omega du = o(1). \quad (5)$$

Some examples for locally stationary processes are the following:
• Modulated processes: \( X_{t,T} = \sigma_{t,T}^0 \cdot Y_t \), where \( \{Y_t\} \) is a stationary process, and
\[ | \sigma_{t,T}^0 - \sigma(t/T) | = O(T^{-1}), \] for some smooth function \( \sigma(u) \) on \((0,1)\). The resulting evolutionary spectrum is simply a one dimensional function in \( u \) times a one dimensional function of frequency, the spectrum of \( \{Y_t\} \).

• Time-varying moving average processes: Here, the sequence \( A_{t,T}^0(\omega) \) coincides with the smooth function \( A(t/T, \omega) \) which is a rational function in the frequency \( \omega \) (as it is for stationary MA-processes). This does not hold any more for the next example, due to the additional autoregressive part.

• Time-varying ARMA processes: These can again be best characterized by the form of their evolutionary spectrum, which as a function of frequency is exactly the same ratio of rational AR and MA parts as it is for the stationary case. The time dependency simply shows in the dependency of the AR- and MA-parameters, being functions of rescaled time \( u \), now.

In all of these examples one easily observes that the theory of locally stationary processes is a real generalization of classical stationarity: For stationary processes the evolutionary spectrum coincides with the classical spectral density, as should be expected from a concise theory.

2.2 Alternative or related models

An alternative idea to describe the deviation from stationarity is the approach of [MPZ] which tries to explicitly control first and higher order derivatives of a continuum symbol describing a generalized Fourier spectrum of the nonstationary process. It relates to the Dahlhaus approach in as much as it works with explicit bounds on the variation in time of the spectrum and with explicit local intervals of stationarity in which the covariance approximately behaves like a stationary one, instead of defining a sequence of implicitly by the asymptotics over \( T \) controlled intervals. So, the [MPZ] approach might be more appropriate for approximation of a given time series of finite sample size by quasi-stationary models (if one gets sufficient knowledge on the explicit bounds). It does not allow, however, for an asymptotic estimation theory. This is in the same spirit as is the concept of underspread processes of [MHK] where the deviation from stationarity is modeled by the so-called spread. This is accompanied by some definition of and analysis by an evolutionary Weyl spectrum of nonstationary random processes.

The concept of [DoMvS] is kind of a reformulation of the Dahlhaus' model by imposing local stationarity on the sequence of covariances \( \{\gamma_{s,t}\} \):

Assume a triangular array of processes \( \{X_{t,T}\}_{1,T} \), in which the \( T \)-th process obeys a uniform decay of the correlation
\[
\sum_u (1 + 2|u|)^2 \sup_{1 \leq t \leq T} \gamma_{t+t+u}^2 \leq Q_1
\]
and a uniform quasi-stationarity of the covariance
\[
\sup_{s,t} |\gamma_{s,t} - \gamma_{s+u,t+u}| \leq Q_{2,T} |u| \quad \forall u.
\]
Here \( Q_1 \) is independent of \( T \), but \( Q_{2,T} = Q_{2,1}/T \) depends on \( T \), which basically is the necessary smoothness assumption on the time-varying behavior of the covariances and
which parallels equation (2) in the previous subsection. With this, an explicit bound on the deviation of the given covariance $\gamma_{n,t}$ from a stationary one is given which asymptotically tends to zero. There is an additional promising aspect of this approach in that, though it is still uniform, it possibly allows for a generalization to some sort of inhomogeneous nonstationary where one needs different width of segmentation in different stretches of the series. This is currently addressed in work in progress.

Finally, we like to mention the work of [Adak] which uses a slightly modified version of the Dahlhaus model. This leads to a class of locally stationary processes for which simple piecewise stationary models are a good (mean-square) approximation. Last not least, in [DNvS], time-varying autoregressive models are used which are a special case of the general Dahlhaus class of locally stationary processes. I.e. in this model of again a doubly-indexed sequence of AR-processes, the autoregressive parameters are now functions of rescaled time $u$ and allow estimation theory as in traditional nonparametric regression.

3 (Semi-) Adaptive time–frequency estimation

In this section, we now like to present and compare some of the recent estimation procedures which all non–parametrically address the problem of either estimating the time–varying second–order structure of a locally stationary time series or of finding the proper segmentation of the series. Obviously, these two goals are closely connected, and often by trying to come up with an adaptive time–frequency estimate, a suitable time segmentation is given implicitly. This applies, e.g., to the approach of [NvS], which uses the Fourier transform of a very local autocovariance estimate (i.e., a Wigner–Ville estimate) as starting point. In this respect, it can be considered as an "indirect approach" (because the goal of achieving a suitable time–frequency representation is used to yield a proper adaptive segmentation). This is opposed to the more "direct approach" of [DoMvS], which is actually a continuation of the work of [MPZ]. Here, after starting from an explicitly given (overcomplete) set of segmentations by local cosine windows, a suitable algorithm is run to pick up the one best adapted to the time–changing content of the signal. Finally, the less recent approaches of [vSS] and of [Adak] stick to more traditional window Fourier techniques (segmented periodograms), where the segmentation has either to be chosen in advance or can be found by some additional search algorithm, as e.g. presented by the work of [Adak]. We like to discuss these first.

In the following we choose our notation to be already adapted to the situation of a doubly-indexed process $\{X_{t,T}\}$ following a model of local stationarity, as in Definition 2.1. That is, in the following the time variable of the estimators will be the rescaled time $u = t/T \in [0, 1]$. However, as long as we are not concerned about asymptotics, our presented procedures should be understood as algorithms which provide estimators as functions of time $t$ (and frequency $\omega$).

3.1 Segmented periodogram estimation of evolutionary spectra

Given the stretch of data $X_1, ..., X_T$ of the observed time series, we consider a short–time (or windowed) periodogram localized at the midpoints $t_i$ of segments of length $N_i$. Here, for simplicity, we restrict to $N_i = N$ for each $i$. For asymptotic considerations, in a model of local stationarity (1), $X_1, ..., X_T$ is identified with $X_{1,T}, ..., X_{T,T}$, and $t = [uT]$. If necessary a data–taper is used to reduce leakage effects, which occur in particular for spectra with a
high dynamic range. Hence, the segmented periodogram writes as follows:

\[ I_N(u, \omega) = H_{2,N}^{-1} \sum_{s=0}^{N-1} \left( \frac{h(s)}{N} \right) X_{[uT-s/N+s+1],T} \exp(-i\omega s) \]  

(8)

where \( h : [0, 1] \rightarrow [0, 1] \) is a sufficiently smooth taper and \( H_{2,N} = \sum_{s=0}^{N-1} h^2(s/N) \) the appropriate norming factor with \( H_{2,N} \sim N \) (see [Da]).

In practice, \( I_N(u, \omega) \) is calculated on possibly overlapping segments of \( X_{i,T} \) of length \( N \) with, for asymptotic considerations, \( N \rightarrow \infty \), as \( T \rightarrow \infty \), but \( N^2/T \rightarrow 0 \), in order to be able to estimate the correct time dependency of the evolutionary spectrum \( f(u, \omega) \). For details we refer to [Da].

Of course, as for classical periodograms of stationary time series, now the question arises how to smooth this inconsistent and fluctuating estimate, moreover, how to do it adaptively over frequency and time. For this, [vSS] use a 2-d wavelet thresholding scheme, instead of using classical (2-d) kernel estimation (with possibly local bandwidth selection), see e.g. [Ri], [Da2]. Basically, given a periodized orthonormal wavelet basis \( \{ \phi_{00}(\omega) \} \cup \{ \psi_{j2,k2}(\omega) \}_{j2 \geq 0, k2} \) in frequency and an orthonormal basis \( \{ \phi_{00}(u) \} \cup \{ \psi_{j1,k1}(u) \}_{j1 \geq 0, k1} \) of \( L_2([0,1]) \) in time, they apply traditional wavelet algorithms (cf. Section 1.3) of depth \( J = \log_2(N) \) to represent \( I_N(u, \omega) \) as a wavelet series both with respect to frequency \( \omega \) and (rescaled) time \( u \). Then, as briefly described in Section 1.3, by soft or hard thresholding of the 2-d wavelet coefficients, \( N^2 \) in number, they achieve a denoised reconstruction which, using the framework of the model of local stationarity, is proven to be an asymptotically unbiased and consistent estimator with a near-optimal rate of convergence of the \( L_2 \)-risk.

As a side remark, we like to mention that in many situations, the use of a 2-d tensor product basis proves to be superior to the use of a one-scale MRA basis, widely used in image processing applications. It has better compression abilities and is more flexible to adapt to different degree of smoothness in different directions (for details on this discussion, see [NvS]). Hence, in this overview, in our examples shown below, we restrict to a presentation of the 2-d tensor product wavelet basis which is simply application of two (not necessarily identical) 1-d wavelet schemes for each of the two considered directions, time and frequency.

More practically speaking, non-linear wavelet thresholding can be seen as a simple methodology for providing a locally smoothed estimator, which can be observed by the following example:

Figure 1 shows the grey-scale image plot of a time-varying Doppler spectrum, the “true underlying” evolutionary spectrum. Figure 2 shows the corresponding time series of length 2048, called “Mojuti2048”, a locally stationary process as in (1) generated to have this evolutionary spectrum as in (4). Its particular analytic expression can be found in [vSS], Section 5.2.

Based on this realization, Figure 3 provides the raw segmented (log-) periodogram, with \( N = 128 \), and the denoised version of it, with the use of tensor wavelets (i.e. periodized symmlets of order 6) and hard thresholding: As we used the logarithm of the periodograms in order to benefit from its stabilized variance, the threshold was chosen to be a uniform one of the form \( \hat{\sigma} \cdot \sqrt{2 \log(T)} \) (cf. Section 1.3), with \( \hat{\sigma} \) being a standard variance estimator of the wavelet coefficients of the finest scale. The outcome of this algorithm is reproducible by using the function “TFEstWavethresh” of the matlab software WaveLab, (forthcoming) version 0.800, see [WaveLab], available via ftp from playfair.stanford.edu/pub/wavelab or
Figure 1: Grey–scale image of true evolutionary spectrum of process "Mojuti2048"

Figure 2: Realization of length $T = 2048$ of process "Mojuti"
Figure 3: Segmented logperiodogram and wavelet denoised version. Top: Grey-scale images in t–f–plane. Bottom: corresponding wavelet coefficients before and after hard thresholding


Finally, for enhanced comparison, in Figure 4 some cuts in frequency direction, at time point \( t = 64 \) are shown, now including a comparison with the use of the MRA wavelet basis.

For more pictures with some better visual quality we also refer to [vSS].

Obviously, the adaptation properties of this estimate depend on the choice of the segment length \( N \). Once 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. Its optimal choice depends on the relation between the (unknown) smoothness of the evolutionary spectrum in time and frequency direction, and has to be adapted to rate of the change of the spectrum. So the segmented periodogram approach possibly suffers from the drawback how to choose the segment length \( N \) appropriately, best in a data-driven way.

3.2 Choice of segment length

One possibility to come up with an automated choice of the appropriate segmentation is given by [Adak], who addresses this problem by using a tree-structured search algorithm: This is done by partitioning the time-domain into a hierarchy of segments of dyadic lengths, halving each segment recursively into shorter, possibly overlapping, segments. After estimation of the spectrum with a segmented periodogram (8) in each of the blocks of this tree, an optimal pruning algorithm is used to recombine adjacent segments for which the spectra are the same. That is, the algorithm is supposed to find the most stationary segmentation with the smallest number of segments, which is achieved by minimizing a
Figure 4: Cuts in frequency direction, at time \(t=64\): true spectrum, raw segmented periodogram, wavelet denoised, with MRA and Tensor 2-d basis.

sum (over the segments) of distances between the estimated spectra from the left and the right half of the segment respectively. As distance, a measure of discrepancy between the two spectra is proposed, e.g. a Kolmogorov–Smirnov distance. The optimal pruning algorithm is similar to Wickerhauser’s best basis algorithm ([CW]). It incorporates the choice of a penalty parameter penalizing the number of segments. This is mainly to avoid a segmentation with very small segments resulting into a loss of frequency resolution. For the choice of the penalty parameter Adak suggests the use of cross validation procedures.

We feel that this approach of [Adak] is certainly very useful for signals which clearly show different pieces of stationarity (like in speech processing), but not exclusively for those.

Another possibility, which totally avoids a preliminary choice of the segmentation and which combines this adaptation problem with the appropriate smoothing in time and frequency, will be discussed in the next subsection.

3.3 Adaptive Wigner estimation of evolutionary spectra

In contrast to using the short–time periodogram, in order to avoid the \textit{a priori} choice of segmentation, the approach of [NvS] starts from a different estimator for a time–dependent spectrum. It is basically a Wigner estimate, and is constructed as the Fourier transform of the most possible local estimate of the local autocovariance \(c(t, s) = \text{Cov}\{X_{t-s/2}, X_{t+s/2}\}\) occurring in equation (3) of the definition of the evolutionary spectrum (again with \(t = \lfloor uT \rfloor\)):

\[
I_{t,T}(\omega) = \frac{1}{2\pi} \sum_{s: 1 \leq [t-s/2], [t+s/2] \leq T} X_{[t-s/2],T}X_{[t+s/2],T} \exp(-i\omega s), \quad t = 1, ..., T. \quad (9)
\]
We observe that it is simply \( X_{[t-s/2],T} \cdot X_{[t+s/2],T} \) which is used to estimate the local autocovariance \( c(t, s) \). No additional summation over \( t \), i.e. no averaging over time, is performed as is in the (segmented) periodogram (8). This is a meaningful estimator of the evolutionary spectrum (4) as the expected value of \( I_{[u,T]^\prime} \cdot T(\omega) \) tends to \( f(u, \omega) \) as \( T \) tends to infinity. However, it is even more fluctuating then the segmented periodogram, shows cross interference terms (as typical for Wigner estimates) and is not necessarily positive. So a sophisticated adaptive smoothing procedure needs to be applied to the set of \( T \) Wigner estimates in time, for each of the \( T \) grid (Fourier) frequencies. It is also used to reduce the redundancy in the original set of \( T^2 \) resulting wavelet coefficients by ruling out most of them.

This can be achieved by applying the same principle of 2-d tensor product wavelet thresholding as before, now with two 1-d wavelet transforms of depth \( J = \log_2(T) \). However, in this particular situation, in addition to the smoothing (denoising) aspect, this algorithm will adaptively pick up the right time–frequency segmentation for the given signal: Only at most \( T \) out of originally \( T^2 \) 2-d wavelet coefficients of the Wigner estimate carry significant non–redundant signal information, and if the underlying evolutionary spectrum is smooth then possibly a lot fewer coefficients. Hence, appropriate non–linear wavelet thresholding does a two-fold task: It provides the right time–frequency segmentation, i.e. in particular the \( a \ postriori \) right windowing of the time series in time, and it is used to denoise and smooth the resulting estimator of the time–dependent spectrum (as in the previous approach). In particular, if the time series is actually stationary then all wavelet coefficients (i.e. the ones of the wavelet series of the underlying spectrum) in time will be zero (only a scaling coefficient remains). Appropriate thresholding should set also (almost) all the empiricals to zero then.

We apply this algorithm to the simulated example of Section 3.1: In Figure 5 we show the raw Wigner Estimate, with \( T = 2048 \), on a logarithmic scale of its modulus, to enhance the effects of the cross interference terms. The denoised version was achieved by hard thresholding of the scales \( j = 0, \ldots, 7 \) in each direction. Here, level–dependent thresholding is needed, with thresholds of the form \( \hat{\sigma}_j \cdot \log(T) \), with \( \hat{\sigma}_j \) being a median absolute deviation estimate of the coefficients on each scale \( j \). Again, symmlets of order 6 were used in the WaveLab function TFEstWaveThresh (cf. Section 3.1). Note that the cross interference terms are almost completely suppressed by the denoised estimator.

Summarizing this approach of [NvS] is able to treat situations with a sudden change in the covariance structure of a nonstationary time series. Also, by using a tensor wavelet approach with its separable time and frequency windows, it provides immediately an automatic choice of the right degree of smoothing in both time and frequency direction. The resulting non–linear threshold estimator achieves the near–optimal rate for the \( L_2 \)–risk in function classes with \textit{anisotropic} smoothness, i.e. different degree of regularity into different directions.

### 3.4 Adaptive covariance estimation by Best Basis methods

Quite a different approach to a related question, namely to find the proper time segmentation in order to adaptively estimate the time–varying autocovariance of a non–stationary process, can be found in the work of [MPZ] and [DoMvS]. In the first work, in the model shortly described at the beginning of Section 2.2, [MPZ] suggested a method
for approximating the covariance of a locally stationary process by a covariance which is diagonal in an ideally constructed Coifman–Meyer basis of cosine packets [CM]. Hence, as a priori overcomplete set of time window functions a family of local cosine functions is used for segmentation.

Their discussion is primarily based on inference in the population statistics of the process, i.e. they investigate the theoretical quantities of variance and covariance. As criterion to find the ideal best local cosine basis they use maximizing the sum of squares of the theoretical variances of the local cosine transform of the data over all elements in the library of local cosine bases.

A natural question arising from their work is to translate these approximation results into estimation results, i.e. to treat the situation where only sample statistics are available, with the attendant noise, i.e. small sample fluctuations. This is considered in [DoMvS], where the problem of estimation of the covariance from sampled data is discussed. There it is shown that it is possible to obtain an empirical basis from sampled data which is nearly as good as the ideal theoretical basis.

The abstract point of view behind these ideas is that for stationary process there is a certain basis – the Fourier basis – which essentially diagonalizes the covariance matrix $\Gamma$ of the stationary process (which is exactly true for periodic processes with a discrete cyclic spectrum).

Now, for a non-stationary process, suppose an orthonormal basis $B$ in which the covariance is essentially diagonalized were given. Then one could attempt to estimate the covariance by first rotating into the basis $B$, forming the empirical covariance estimate in the new basis, killing all terms off diagonal (because one knows that the theoretical covariance
is essentially diagonal in this basis), resulting in the diagonal matrix $\Delta_B$ and then rotating back into the original basis. The resulting covariance estimate $C_B = B\Delta_B B'$ is mean-square converging (in an appropriate norm) with a rate of order $O(1/n)$ independently of $T$, and so can work well if one can obtain an increasing number $n$ of realizations of $X_1^{(i)}, \ldots, X_T^{(i)}$, $1 \leq i \leq n$. If one knows that the entries on the diagonal in the basis $B$ are smooth, then by appropriately smoothing these diagonal entries and proceeding as before, as $T$ tends to infinity, one can get a consistent covariance estimate even for $n = 1$ single realization. This is exactly the same paradigm as is for the slowly time-varying spectrum of a locally stationary process, see above.

Now it remains to estimate the unknown diagonalizing basis from the sampled data which is done by a fast search algorithm, similar to the Best Basis search of [CW], to maximize the squared sum of appropriately smoothed empirical variances of the local cosine transform of the data within a certain window over all bases in the library of cosine packets. The resulting best adapted empirical basis proves to be nearly as good as the ideal theoretical basis. For the appropriate smoothing again a wavelet soft-thresholding procedure is used, which exactly parallels wavelet thresholding of periodogram estimates for stationary processes (because within each block of varying size the process is assumed to be stationary). Under the following model, even when $n = 1$ and $T \to \infty$ this approach leads to a consistent estimate of the covariance which is nearly-diagonal in the estimated basis, and to some rates on the convergence of a mean-square loss measuring the overall deviation of the estimated from the true covariance matrix.

What proves useful in order to accomplish this is the description of local stationarity formulated directly on the sequence of covariances, as described by equations (6) and (7) in Section 2.2.

4 Time-scale modelling

As an alternative to time-frequency decompositions of non-stationary processes, time-scale models express a different idea of a local autocovariance analysis: here the process is decomposed into atoms which match the typical correlation length of $X_t$ in a local neighborhood of $t$, i.e. this correlation length is matched by the scale (or "wave") length of the building blocks. A simple example (mentioned also in [Wal]) is the superposition of box car functions $\varphi(t-k)$ with (uncorrelated) random amplitudes $\eta_k$,

$$X_t = \sum_k \eta_k \phi(t-k), \quad \phi(t) = 1_{[0,1]}(t),$$

or, to allow for really changing scale length $a_j$ of the atoms,

$$X_t = \sum_{j,k} \eta_{jk} \psi(a_j(t-k)), \quad \psi(t) = 1_{[0,1/2]}(t) - 1_{[1/2,1]}(t),$$

which resembles a stochastic wavelet decomposition with Haar wavelets. In fact, one possibility arises for the specific choice $a_j = 2^j$, which is a discretely indexed family of wavelets $\{\psi_{jk}\}, \psi_{jk}(t) = 2^{j/2}\psi(2^jt - k)$, not necessarily restricted to the Haar family, of course.

This specific set-up and its statistical treatment has been introduced by [vSNK] where a wavelet decomposition of a stochastic process is developed which parallels a time-localized
Cramér (Fourier) spectral representation - again to be understood in the mean squared sense. Instead of thinking as scale in terms of “inverse frequency” they start from genuine time–scale building blocks or “atoms”: Assume \( \psi(t) \) is localized at time 0 then typically the variance (“power”) of a coefficient \( \eta_{j,k} \) will be large if at time \( t = 2^{-j} k \) the correlation length of \( X_t \) is approximately of the order of magnitude of the “wavelength” of the atom \( \psi_{j,k} \) (basically the length of the support of this wavelet), which is proportional to \( 2^{-j} \). The promising aspect of this model is that it allows local (time) variation, in a way that fast “oscillations” are modeled to change quickly and slow “oscillations” to change slowly. Here, the restriction to discretely indexed wavelets is mainly in order to embed this new concept into a statistical (estimation) theory. One can certainly think about more general continuously–indexed models ([vS]) and their usefulness. Also, some related approaches should be mentioned, e.g. the one of [MoCh], which however covers only the stationary situation, and the one of [AGF] which uses a concept similar to the following concept of wavelet spectral analysis, in a different context, i.e. analysis of \( 1/f \) processes.

To do second–order analysis, a wavelet spectrum is introduced which is the expected value of the squared modulus of the coefficients \( \eta_{j,k} \). This quantity is a localized wavelet spectral measure in the time–scale plane. It measures the local power in the variance–covariance decomposition of the process \( \{X_t\} \) at a certain scale \( j \) and a time location \( k \) and delivers a time–scale decomposition in as much as (1) represents a time–frequency decomposition. In comparison with so-called “scalegrams”, where for fixed \( j \) the power in a whole “frequency band” corresponding to the inverse of the scale \( j \) is considered by \( \sum_k |\eta_{j,k}|^2 \), this new approach delivers a real local analysis (cf. Figure 7 for the data example in the next subsection.)

As for the Dahlhaus model in Definition 2.1, again a concept of local stationarity – now in “time” \( k \) – is needed to allow for identification and estimation of the model and its coefficients. What is mainly needed for asymptotics is a coupling of the lowest wavelet scale length (which is called \(-J\) here) to the (asymptotically growing) length of the process itself. Moreover, one needs to collect a growing local information per time–scale. In order to so, the control of slow variation over \( k \) is introduced by the concept of “locally stationary wavelet (LSW) processes”, again a doubly–indexed array of processes \( \{X_{t,T}\}_{t=1,...,T}, T \geq 1 \), which fulfills the following definition:

**Definition 4.1** Define a class of LSW processes, a sequence of doubly-indexed stochastic processes \( \{X_{t,T}\}_{t=1,...,T}, T \geq 1 \) with the following representation in the mean–square sense with respect to a given wavelet basis \( \{\psi_{jk}(t)\}_{j,k} \) of \( L_2(\mathbb{R}) \), and an orthonormal random sequence of increments \( \xi_{jk} \)

\[
X_{t,T} = \sum_{j=-J}^{+1} \sum_k w_{j,k:T}^0 \psi_{jk}(t) \xi_{jk},
\]

where \( 2^J = T \), \( \psi_{jk}(t) = 2^{j/2} \psi(2^j t - k) \), \( j = -1, -2, \ldots, -J(T) = -\log_2(T) \), \( k \in \mathbb{Z} \), and possessing the following properties:

1. \( E\xi_{jk} = 0 \) for all \( j, k \). Hence \( E X_{t,T} = 0 \) for all \( t \) and \( T \).
2. \( \text{Cov}(\xi_{jk}, \xi_{km}) = \delta_{jk} \delta_{km} \).
3. There exists for each \( j \leq -1 \) a Lipschitz-continuous function \( W_j(z) \) on \((0,1)\) such that

\[
\sup_k |w^0_{j,k;T} - W_j \left( \frac{k 2^{-j}}{T} \right)| = O(T^{-1}) \quad \text{for } T \to \infty,
\]

where for each \( j = -1, \ldots, -J(T) = -\log_2(T) \) the sup is over \( k = 1, \ldots, 2^j T \).

4. The real-valued wavelet basis \( \{\psi_{jk}\}_{jk} \) is orthonormal and the wavelets have compact support.

5. \( \sum_{j=-\infty}^{-1} |W_j(z)|^2 < \infty \), for all \( z \in (0,1) \).

With this a theory for the estimation of the “evolutionary wavelet spectrum” is developed. This is a localized density of a wavelet spectral measure in the (rescaled-) time-scale plane:

**Definition 4.2** Let

\[
S_j(z) := |W_j(z)|^2, \quad z \in (0,1).
\]

\( \{S_j(z)\}_{j \leq -1} \) is the “evolutionary wavelet spectrum” of the sequence \( \{X_{t,T}\} \) with respect to \( \{\psi_{jk}\} \).

Using assumption 3 of Definition 4.1,

\[
S_j(z) = \lim_{T \to \infty} |w^0_{j,2^j(zT);T}|^2, \quad \forall z \in (0,1).
\]

Quite naturally this quantity is a constant over time (\( z \)) if the underlying process is actually stationary in the classical sense.

The asymptotics are based on rescaling in time–location which, similarly to the model of Definition 2.1 of time–frequency local stationarity, allows to perform rigorous estimation theory starting from a single stretch of observations of \( \{X_{t,T}\} \). Here, the smoothness assumption on \( W_j(z) \) controls the variation of each coefficient \( w^0_{j,k} \) as a function of \( k \) so that again it can only vary slowly enough.

This model is slightly restrictive in that it sticks to uncorrelated increments \( \xi_{jk} \). It can certainly be generalized to more realistic models, e.g. those which allow for weak dependence within (or even across) scales.

An estimate of the evolutionary wavelet spectrum is given by a wavelet periodogram or scalogram, i.e. the squared coefficients from a discrete or stationary wavelet transform (SWT), as in [NaSi]. However, in order to estimate the spectrum consistently the periodogram needs to be smoothed. In [vSNK2] asymptotic results are developed for smoothing by non-linear wavelet shrinkage of the wavelet periodogram treated as a function of rescaled time–location with respect to another orthogonal wavelet basis (as in analogous approaches in non-parametric curve estimation, e.g. [Do], [NvS]). Basically they parallel the ones for time–frequency estimation, as in Section 3.1 and 3.2

Finally, an inverse transformation of the smoothed wavelet periodogram is suggested that estimates a localized autocovariance of the original stochastic process.
Figure 6: Electrocardiogram recording of 66 day old baby. Series is sampled at 0.25Hz and is recorded from 21:17:59 to 06:27:18.

4.1 An example

As motivating example let us consider the following time series:

Figure 6 shows a subsample of a electrocardiogram (heart rate) recording of an 66 day old baby sampled at 0.25Hz from 21:17:59 to 06:27:18. The original series was sampled at 1Hz however the power at high frequencies was found to be insignificant and so a 0.25Hz subsampled version was derived. It is unlikely that this will be a stationary time series. For one thing, of interest to paediatricians, the “heart rate” varies considerably over time and changes significantly between periods of sleep and waking. On the other hand, application of the time–frequency concepts of the previous sections does not reveal any informative features. So a wavelet–based time–scale approach seems to be an appropriate possibility: Figure 7 shows the raw wavelet periodogram coefficients for scale $j = -3$ for the baby ECG data series shown in Figure 6. Figure 8 shows the results of a non–linear wavelet smoothing of the wavelet periodogram. Here Daubechies’ [Dau] least–asymmetric wavelets $N = 10$ were used to form the SWT. The logarithms of the squared coefficients $I_{k,8192}^{3}$ were then directly subjected to non–linear stationary wavelet smoothing using universal soft thresholding on all scales with a MAD–based estimate of the noise level using Daubechies’ least–asymmetric $N = 10$ wavelets. This resulted in our estimate of $S_{-3}(z)$ given in Figure 8 (solid line). The overlying dotted line in that figure refers to the baby’s sleep state as judged by a trained human observer. The observer classifies the baby’s state of sleep as unknown, quiet sleep, hard to tell, active sleep and awake as indicated in the caption to the figure. It is clear that there is some, if somewhat imperfect, relationship between the estimate of $S_{-3}(z)$ and the sleep state. In particular, periods of active and wakefulness occur whilst the estimate of $S_{-3}(z)$ is large and periods of quiet sleep when it is small. In effect, the $S_{-3}$ quantity is measuring the local stochastic variability of the signal at that scale and it correlates fairly well with the sleep state variable. Objective measures of heart–rate variability are of great interest to paediatricians especially when linked to external
Figure 7: The (stationary) wavelet periodogram $I_{k,8192}^{3}$ for baby heart rate data (solid line) for the particular scale $j = -3$. The dotted line indicates the sleep state for the baby and its values are indicated by the right-hand axis ($0=$unknown, $1=$asleep, $2=$hard to tell, $3=$active sleep, $4=$awake).

Figure 8: Estimate of evolutionary wavelet spectrum $S_{-3}(x)$ for baby heart rate data (solid line). The dotted line indicates the sleep state for the baby and its values are indicated by the right-hand axis ($0=$unknown, $1=$asleep, $2=$hard to tell, $3=$active sleep, $4=$awake).
observables such as sleep state. More on this example can be found in [vSNK2].

5 Acknowledgements

The author likes to greatly acknowledge the hospitable and fruitful atmosphere of the Stanford Statistics Department, where this paper was written during his sabbatical leave from Kaiserslautern, under the funding of the German National Science Foundation (DFG). He would also like to thank all his co-authors, and Thomas P.-Y. Yu and Xiaoming Huo for their support with WaveLab (see [WaveLab]): This is a Matlab software, developed at Stanford by David L. Donoho and collaborators, which can be used to reproduce most of the plots shown in section 3. The plots of section 4 were produced by Guy Nason with his S-Plus software WaveThresh3. The ECG data were kindly supplied by Prof. Peter Fleming, Dr Andrew Sawczenko and Jeanine Young of the Institute of Child Health, Royal Hospital for Sick Children, Bristol, U.K.

This paper was given as a presentation on the 12th SINAPE, at Caxambu, Brazil, July 29 - August 2, 1996. The author likes to thank the organizers for the good organization of and the invitation to this successful meeting in a lovely and very enjoyable environment.

References


