Nonparametric curve estimation by wavelet thresholding with locally stationary errors

August 28, 1997

Rainer von Sachs
Department of Mathematics
University of Kaiserslautern, Germany
rvs@mathematik.uni-kl.de

Brenda MacGibbon
GERAD and Département de mathématiques
Université du Québec à Montréal
macgibbon.brenda@uqam.ca

Abstract

In the modeling of biological phenomena, in living organisms whether the measurements are of blood pressure, enzyme levels, biomechanical movements or heartbeats, etc., one of the important aspects is time variation in the data. Thus, the recovery of a "smooth" regression or trend function from noisy time–varying sampled data becomes a problem of particular interest. Here we use non–linear wavelet thresholding to estimate a regression or a trend function in the presence of additive noise which, in contrast to most existing models, does not need to be stationary. (Here, nonstationarity means that the spectral behaviour of the noise is allowed to change slowly over time). We develop a procedure to adapt existing threshold rules to such situations, e.g., that of a time–varying variance in the errors. Moreover, in the model of curve estimation for functions belonging to a Besov class with locally stationary errors, we derive a near–optimal rate for the L_2–risk between the unknown function and our soft or hard threshold estimator, which holds in the general case of an error distribution with bounded cumulants. In the case of Gaussian errors, a lower bound on the asymptotic minimax rate in the wavelet coefficient domain is also obtained. Also it is argued that a stronger adaptivity result is possible by the use of a particular location and level dependent threshold obtained by minimizing Stein's unbiased estimate of the risk. In this respect, our work generalizes previous results, which cover the situation of correlated, but stationary errors. A natural application of our approach is the estimation of the trend function of nonstationary time series under the model of local stationarity. The method is illustrated on both an interesting simulated example and a biostatistical data–set, measurements of sheep luteinizing hormone, which exhibits a clear nonstationarity in its variance.

KEYWORDS: non–stationary time series, time–varying covariance, local stationarity, minimax estimation, non–linear wavelet thresholding, threshold choice.

1 Introduction

In the original series of papers by Donoho and Johnstone (Donoho and Johnstone (1992), Donoho and Johnstone (1994), Donoho and Johnstone (1995), Donoho, Johnstone, Kerkyacharian and Picard (1995)) on minimax wavelet shrinkage methods for the reconstruction of an unknown function observed in a white noise model, the noise was always assumed to be i.i.d. Gaussian. They considered data of the form

$$y_i = f(t_i) + \varepsilon_i \quad (i = 1, \ldots, n) \tag{1}$$

where $t_i = \frac{i}{n}$, ε_i are i.i.d $N(0, \sigma^2)$ and f is the unknown function to be recovered. The performance of $\widehat{f}$, an estimator of f, is usually measured by its L_2–risk evaluated at the sample points (t_i), $(i = 1, \ldots, n)$. When considering representations of such functions in orthonormal bases such as wavelets, the Parseval relation establishes an equivalence between the L_2–risk in the function space and the l_2–risk of the wavelet coefficients of the functions, and consequently theorems can be proved in the sequence space of the wavelet coefficients rather than the function space itself.

As is well known in practice, for many applied examples from biological observations over time this model of independent identically distributed observations is no longer valid. Various authors considered wavelet estimation in the case of stationary correlated noise: Brillinger (1994) obtained pointwise results, Neumann and von Sachs (1995) studied the L_2–risk of wavelet threshold estimator for general stationary errors, Wang (1996) considered minimax rates of threshold estimators for fractional Brownian motion; and Johnstone and Silverman (1997), inspired by a neurophysiological problem, carried out a detailed investigation of wavelet threshold methods for Gaussian stationary observations, with short and long range dependence risk. Johnstone (1996) completed this study by proving that a stronger adaptivity result holds for SURE estimates and applying these to inverse problems.

However, in many applied data analyses, the observations can no longer be assumed to come from a stationary process. For example, there has been much recent interest in the analysis of series with pulsatile components. Here the measured variables might be subject to sudden large increases depending on the way the pulses are generated. Such data, although locally stable, may be nonstationary and non–Gaussian. O'Sullivan and O'Sullivan (1988) and Diggle and Zeger (1989), Karsch, Robinson, Woodfill and Brown (1989), Normolle and Brown (1994) used series of luteinizing hormone concentrations to illustrate such phenomena. The last two mentioned papers studied the detection of seasonality using a methodology previously developed by Kitagawa (1987) based on spline estimation. Although there are many other intuitively appealing methods of spatially adaptive function estimation that could be useful for this type of series, a theoretical evaluation of their risk performance seems difficult. Since such series typically exhibit a mean function with both abrupt and gradual changes, we feel that wavelet threshold methods could best handle such local variation in the trend function and the time varying nature of the variance. The question arises how to treat the trend function of these complex series, which with nonparametric wavelet estimation methods has so far only been addressed for stationary errors.

Obviously some assumption on the time varying nature of the coefficients must be made in order to attain our goal of using wavelet threshold methods to estimate the trend and obtain near–optimal asymptotic rates for the L_2–risk over certain smoothness classes of functions. First, we need to guarantee that in order to estimate a single wavelet coefficient at an individual scale and location, an asymptotically growing number of data in the local neighborhood of this scale-location index shows the same statistical behaviour. Second, we must be able to distinguish between signal and noise to perform proper thresholding. Hence there is a need for restricting the departure from stationarity, at least asymptotically.

In recent work on estimating the second–order structure of nonstationary time series a model of local stationarity was used to derive rigorous asymptotic estimation theory for both time–varying autocovariances by Donoho, Mallat and von Sachs (1996), and for time–varying spectral densities by Dahlhaus (1997), by von von Sachs and Schneider (1996) and by Neumann and von Sachs (1997). Here we also choose the model of Dahlhaus (1997) for the error structure. That is, we use non–linear wavelet thresholding to estimate a regression function in the presence of additive noise which, in contrast to most existing models, is not necessarily stationary; its spectral behaviour is allowed to change slowly over time according to the model of local stationarity. In the context of time series analysis, which is a natural application of our quite general approach, this amounts to estimation of the trend of a locally stationary process, and in this respect we complete the above–mentioned papers on second order estimation of such processes.

Intuitively, the asymptotics of this model allow us to consider the data as being split up into "quasi–stationary" segments. Within these asymptotically increasing segments the noise behaves in a more and more stationary way; that is, the noise spectrum becomes locally less and less time dependent. In the wavelet coefficient space the locally stationary model hence guarantees that if we were to fit, in each level of the coefficient domain, segments of locations with quasi–homoskedastic variances, then these segments get larger and larger asymptotically. This allows a consistent variance estimation over these segements which is the key to a good performance of threshold estimators, with the thresholds being proportional to the variation of the wavelet coefficients. So with this approach, asymptotically, noise can be separated from signal, and thresholding will work if it is adapted to a semi–location and level dependent approach which can serve as a practical compromise between totally individual thresholds (which would call for plug-in estimators) and merely level–dependent thresholds which work without bias for the stationary situation but lead to some considerable oversmoothing here.

The outline of the paper is as follows. In section 2, we introduce the appropriate notation for wavelet thresholding methods, and elaborate the model of locally stationary errors. Then we derive the asymptotic formulae for bias and variance of the empirical wavelet coefficients. The main part of this section is that, inspired by the work of Neumann and von Sachs (1995), we establish an upper bound for the uniform L_2–risk of the wavelet threshold estimator over certain smoothness classes for not necessarily Gaussian locally stationary errors. This is done by showing that in the corresponding sequence space of empirical wavelet coefficients the considered L_2–risk is asymptotically equivalent to the one of an accompanying Gaussian noise model. For this a strong form of asymptotic normality of the empirical wavelet coefficients is needed, which puts emphasis on moderate and large deviations. To achieve this, uniformly bounded cumulant conditions on the errors in the curve estimation model are sufficient. Finally we discuss how to translate these theoretical results into practical terms, i.e. we give a practical threshold rule which depends either on a plug–in estimator of the asymptotic variance of the wavelet coefficients or, preferably, an estimator of their variability which is based on homoskedastic approximations along certain segments of quasi–stationarity.

In the third section we look in detail at a simulated data example and a set of sheep luteinizing hormone data. For the simulated example of the by now famous Doppler test function the type of locally stationary noise (e.g. a time–varying AR(2)) is chosen such that the application of the methods previously developed by Johnstone and Silverman (1997) for stationary noise will fail. Hence an adaptive procedure, as prescribed in section 2 which finds segments of "almost" homoskedastic noise for each level is needed.
In the fourth section, in the case of Gaussian errors, inspired by Johnstone and Silverman (1997), and Johnstone (1996), we investigate the minimax properties of a soft threshold estimator by comparing its behaviour to that of an ideal but unattainable benchmark obtained from an "oracle" that provides the optimum diagonal projection estimate. We also indicate how an estimator that minimizes Stein's unbiased estimate of the risk for the case of heteroskedastic variance and different from the usual one for homoskedastic variances (cf. Donoho and Johnstone (1995), Johnstone and Silverman (1997), Johnstone (1996)) could be built. Finally, we conclude with a section on further comments, a discussion and an Appendix containing remaining proofs.

2 Wavelet estimation for regression with locally stationary errors

2.1 The locally stationary error model in curve estimation

The model in curve estimation of interest to us can be written as follows:

$$X_{t,T} \;=\; \mu(t/T) \;+\; \varepsilon_{t,T}\,,\; t = 1, ..., T, \tag{2}$$

where $\mu(u)$ is a function on $[0,1]$ belonging to some general smoothness class. In order to derive our asymptotic results, here we choose to consider Besov classes $B^m_{p,q}$, $m, p, q \geq 1$. For an exact definition of Besov classes see, e.g, Frazier, Jawerth and Weiss (1991). Here we note that m is a smoothness parameter which corresponds to the number of derivatives that $\mu(u)$ possesses in L_p. Further, Besov spaces can be seen as generalizations of Sobolev classes W^m_p which fulfill $B^m_{p,p} \subseteq W^m_p \subseteq B^m_{p,2}$, if $1 < p \leq 2$, and $B^m_{p,2} \subseteq W^m_p \subseteq B^m_{p,p}$ for $2 \leq p < \infty$, with $B^m_{2,2} = W^m_2$. Also, the case $p = q = \infty$ corresponds to Hölder smoothness. They also include functions of bounded variation BV with $B^1_{1,1} \subset BV \subset B^1_{1,\infty}$. For later convenience (see, e.g., proof of Lemma 2.2 (i)), we assume that $\mu(u) \in BV_{[0,1]}$ (and hence we do not need to impose the additional usual restriction $m > 1/p$ on the parameters).

We would like to consider as general a model as possible for the errors while still being able to attain our goal of using wavelet threshold methods to estimate $\mu(u)$, with an near–optimal rate for the L_2–risk over Besov classes in the non–Gaussian situation and to obtain asymptotic "minimax" results for Gaussian noise. The main contribution of our work is to show that this is possible even for nonstationary errors, where, quite naturally, the departure from stationarity needs to be controlled, at least asymptotically. Thus we choose to model the mean–zero errors $\varepsilon_{t,T}$, as a doubly–indexed sequence (array), which fulfills a *local stationarity* assumption as introduced in Dahlhaus (1997). In this class of models, as T tends to infinity, the $\varepsilon_{t,T}$ are assumed to have an *uniquely* defined time–varying (evolutionary) spectrum with a certain prescribed smoothness. A slightly more general notion of "quasi" or local stationarity has been given in Neumann and von Sachs (1997), in a continuation of the work of Dahlhaus (1997). Other possibilities that express this idea in statistical terms can be found, e.g. in Donoho et al. (1996).

For completeness we now give the precise definition of local stationarity of the doubly–indexed sequence $\{\varepsilon_{t,T}\}_{t=1,\dots,T}$. Assume that the following representation holds.

Definition 2.1 *A sequence $\{\varepsilon_{t,T}\}_{t=1,\dots,T}$ is called locally stationary if*

$$\varepsilon_{t,T} \;=\; (2\pi)^{-1/2} \int_{-\pi}^{\pi} A^o_{t,T}(\omega) \exp i\omega t \, d\xi(\omega) \;, \qquad t = 1, \dots T, \tag{3}$$

where

(a) $\{d\xi(\omega)\}_\omega$ is a mean–zero orthonormal increment process on $[-\pi, \pi]$;

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

$$\sup_{t,\omega} |A^o_{t,T}(\omega) - A(t/T, \omega)| \le K T^{-1} \;. \tag{4}$$

The exact regularity assumptions (A3) and (A4) on A are given below. With these, the sequence $\{\varepsilon_{t,T}\}$ (and hence $\{X_{t,T}\}$) has a uniquely defined *evolutionary spectrum*

$$f(u,\omega) \;=\; |A(u,\omega)|^2 \tag{5}$$

as the limit as $T \to \infty$ of

$$f_T(u,\omega) \;=\; (2\pi)^{-1} \sum_{s=-\infty}^{\infty} \operatorname{cov}\{\varepsilon_{[uT-s/2],T}; \varepsilon_{[uT+s/2],T}\} \; \exp(-i\omega s) \;. \tag{6}$$

In general this holds in some mean–square sense as shown in Neumann and von Sachs (1997), Theorem 3.1:

$$\int_0^1 \int_{-\pi}^{\pi} |\; f_T(u,\omega) - f(u,\omega) \;|^2 \, d\omega \; du \;=\; o(1) \;. \tag{7}$$

In this approach the asymptotics are based on rescaling in time–location which allows asymptotic inference starting from a single realization of the data. This is possible as the smoothness of A in u controls the variation of $A^o_{t,T}(\omega)$ as a function which is continuous in t. Here the underlying idea is that for each time point t implicitly there exists a local interval of stationarity which determines this variation. This neighbourhood becomes asymptotically arbitrarily small in the rescaled time u, or, respectively, in actual time t it gets asymptotically larger but at a slower rate than the length T of the whole time series. Estimation is then possible due to the idea of an asymptotically denser and denser design on $(0,1)$.

To give an example of a locally stationary (error) process consider $\varepsilon_{t,T} = \sigma^0_{t,T} \cdot Y_t$, where $\{Y_t\}$ is a stationary process, and $\sup_t |\; \sigma^0_{t,T} - \sigma(t/T) \;| = O(T^{-1})$, 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 $\{Y_t\}$. Other examples are time–varying autogressive–moving average processes (for details, see again Dahlhaus (1997).) In particular, truly stationary processes are automatically included, for which simply $f(u,\omega) = f(\omega)$, i.e. constant in time.
Note that in a time series context (as in Dahlhaus (1997)) estimation of the regression function μ in our model (2) is nothing but estimation of the trend of a second–order nonstationary time series being modeled as a locally stationary process $\{X_{t,T}\}$.

2.2 The non–linear wavelet threshold estimator

Our aim is to estimate $\mu(u), u \in (0,1)$ by non–linear wavelet thresholding. As usual we start from an orthonormal wavelet basis of $L_2([0,1])$, as, e.g. in Cohen, Daubechies and Vial (1993), which we call $\{\phi_{j_0k}(u)\} \cup \{\psi_{jk}(u)\}_{j \geq j_0, k}$. Here, for the interior wavelets, $\psi_{jk}(u) = 2^{j/2}\psi(2^j u - k)$, for $j \geq j_0$, $k = 0, \ldots, 2^j - 1$, where j_0 is the coarsest scale in the scheme. For the boundary wavelets, i.e. those functions that have a support beyond the interval $[0,1]$, appropriate modifications which preserve orthonormality apply, see again Cohen et al. (1993). For more details on the use of these bases, compare, e.g, what is written in Section 2 of Dahlhaus, Neumann and von Sachs (1995).

Accordingly, the wavelet expansion of $\mu(u)$ can be written as

$$\mu(u) = \sum_k \alpha_k \, \phi_{j_0k}(u) \; + \; \sum_{jk} \beta_{jk} \, \psi_{jk}(u) \,, \tag{8}$$

where the "true" scaling and wavelet coefficients are

$$\alpha_k = \int_0^1 \mu(u) \; \phi_{j_0k}(u) \; du \qquad \beta_{jk} = \int_0^1 \mu(u) \; \psi_{jk}(u) \; du \; .$$

Given the observations $X_{1,T}, \ldots, X_{T,T}$, empirical analogs of these coefficients can be written as

$$\widehat{\alpha}_k = T^{-1} \sum_t X_{t,T} \; \phi_{j_0k}(t/T) \qquad \widehat{\beta}_{jk} = T^{-1} \sum_t X_{t,T} \; \psi_{jk}(t/T) \; . \tag{9}$$

However, as usual, in practice these are calculated by some fast wavelet transform algorithm.

Note that in order to avoid boundary problems the aforementioned orthonormal wavelet basis adapted to $[0,1]$ should be used. However, not to obscure our investigations with additional technicalities, in the sequel we proceed as if we were using wavelets on the real line (as it is done in most work on wavelet regression). Some more rigorous comments on the equivalence of both approaches, as, e.g, for Besov norm equivalences, can be found in Dahlhaus et al. (1995).

Using the concepts of hard and soft thresholding of empirical wavelet coefficients originally proposed by Donoho and Johnstone (1992), the thresholded empirical wavelet coefficients are written as follows, for hard thresholding:

$$\delta^{(h)}(\widehat{\beta}_{jk}, \lambda_{jk}) \;\; = \;\; \widehat{\beta}_{jk} \, \mathrm{I}(|\widehat{\beta}_{\mathrm{jk}}| \geq \lambda_{\mathrm{jk}}) \tag{10}$$

and for soft thresholding:

$$\delta^{(s)}(\widehat{\beta}_{jk}, \lambda_{jk}) \;\; = \;\; \mathrm{sgn}(\widehat{\beta}_{\mathrm{jk}}) \cdot (|\widehat{\beta}_{\mathrm{jk}}| - \lambda_{\mathrm{jk}})_{+}. \tag{11}$$

In the sequel we will use $\delta^{(\cdot)}$ to denote either soft or hard thresholding.

The resulting non–linear threshold estimator is the empirical analog to the true wavelet expansion (8):

$$\widehat{\mu}(u) = \sum_k \widehat{\alpha}_k \; \phi_{j_0k}(u) \; + \; \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \widetilde{\beta}_{jk} \; \psi_{jk}(u) \,, \tag{12}$$

where $J = \log_2(T)$ and $\widetilde{\beta}_{jk} = \delta^{(\cdot)}(\widehat{\beta}_{jk}, \lambda_{jk})$. The problem of appropriately choosing the thresholds λ_{jk} and moreover, of how to determine them from the data, is discussed below.

2.3 Asymptotic properties of the empirical wavelet coefficients

In order to do asymptotics on the empirical wavelet coefficients for deriving both (the necessary strong form of) asymptotic normality and also a suitable choice of threshold, we first need to calculate bias, variance and higher cumulants of the empirical wavelet coefficients. Note that analogous results under similar conditions hold for the scaling coefficients $\widehat{\alpha}$, as well.

For this we need the following conditions, where in accordance with the regularity m of the considered Besov class $\mathcal{B}^m_{p,q}$ for our function μ we choose compactly supported wavelet functions of regularity $r > m$, as given by (A1):

(A1) For some $r > m$, we assume that

(i) ϕ and ψ are $C^r[0,1]$ and have compact support,

(ii) $\int \phi(t)\,dt = 1, \quad \int \psi(t) t^k\,dt = 0 \quad \text{for} \quad 0 \leq k \leq r.$

Additional further assumptions are:

(A2) $\sup_{t_1} \sum_{t_2,\ldots,t_p} |\operatorname{cum}(\varepsilon_{t_1,T},\ldots,\varepsilon_{t_p,T})| \leq C^p (p!)^{1+\gamma}$ for all $p = 2,3,\ldots$, for all $T \geq 1$, where $\gamma \geq 0$ and C denotes some positive constant.

(A3) a) $\sup_{u,\omega} |A(u,\omega)| < \infty$.

b) $\inf_{u,\omega} |A(u,\omega)| \geq \kappa$ for some $\kappa > 0$.

(A4) Let $\widehat{A}(u,s) := 1/(2\pi) \int A(u,\omega)\, \exp(i\omega s)\, d\omega$, $s \in \mathbf{Z}, u \in [0,1]$. Then:

a) $\sum_s |s| \, \sup_u |\widehat{A}(u,s)| < \infty$.

b) $\sum_s |s| \, TV_{[0,1]}(\widehat{A}(\cdot,s)) < \infty$,
where $TV_{[0,1]}(\widehat{A}(.,\ell))$ denotes the total variation of the Fourier transform $\widehat{A}(.,\ell)$ of $A(.,\omega)$ as a function in the first argument $u \in [0,1]$.

Now condition (A2) is a type of uniform mixing condition and allows us to derive asymptotic normality of the $\widehat{\beta}_{jk}$ in a uniform way for an increasing number of coefficients by the use of a technique found in Neumann (1996), Lemma 3.1, and also in Neumann and von Sachs (1995) and Dahlhaus et al. (1995). Alternatively, appropriate moment conditions and mixing could be used to derive our results. However, (A2) is a very convenient conditions which works uniformly in order to deal with a wide range of correlated, non–Gaussian, and even nonstationary data. The case $\gamma > 0$ would allow us to include heavier–tailed distributions other than the Gaussian.
The conditions in (A4) essentially express that $A(u,\omega)$ is of bounded total variation in $u \in [0,1]$ and continuously differentiable in $\omega \in [-\pi,\pi]$, uniformly in u. For technical reasons some slightly stronger assumptions are needed to facilitate proofs. Note that (A4)(a) implies (A3)(a) (which is given for the sake of completeness).
Note also that (A3)(b) is fulfilled in the special situation of Section 4, under the condition (33) on a lower bound on the eigenvalues of the covariance matrix of the errors $\{\varepsilon_{t,T}\}$.

For the following lemma, only minimal smoothness of the wavelets (bounded variation on $[0,1]$) is needed.

In view of (the specific form of) the asymptotic normality of the empirical wavelet coefficients considered in Section 2.4 (see, e.g. equation (16)) we introduce here the following index set $\mathcal{J}_T$. It basically formalizes the fact that asymptotic normality can only hold on scales which are asymptotically bounded away from the finest scale $J = \log_2(T)$, i.e. for scales j with $2^j = o(T)$. A convenient possibility is to introduce

$$\mathcal{J}_T = \{(j,k) \mid \; 2^j \leq C \; T^{1-\delta}\}$$

for some (small enough) $\delta > 0$ where δ will be further specified in equation (19).

Lemma 2.2 *Under assumptions (A2)–(A4), with $\psi \in BV_{[0,1]}$, the following holds uniformly in $\mathcal{J}_T$:*

(i)

$$E\;(\widehat{\beta}_{jk}) \;-\; \beta_{jk} \;=\; O\;(2^{j/2} \cdot T^{-1})\;. \tag{13}$$

(ii)

$$\sigma^2_{jk} \;:=\; \mathrm{var}\{\widehat{\beta}_{jk}\} \;=\; T^{-1} \int_0^1 \psi^2_{jk}(u)\; f(u,0)\; du \;+\; o\;(T^{-1}) +\; O\;(2^j \cdot T^{-2})\;, \tag{14}$$

(iii) If $\sigma^2_{jk} \geq C \cdot T^{-1}$ for some positive C, then

$$|\,\mathrm{cum}_p(\widehat{\beta}_{jk}/\sigma_{jk})| \;\leq\; (p!)^{1+\gamma}\;(\tilde{C}T^{\nu})^{-(p-2)}\;, \tag{15}$$

for all $p \geq 3$, uniformly in $\mathcal{J}_T$, with γ as in Assumption (A2) and appropriate $\tilde{C}, \nu > 0$.

In the leading term of the asymptotic variance σ^2_{jk}, $f(u,0)$ denotes the evolutionary spectrum of the $\{\varepsilon_{t,T}\}$ at frequency $\omega = 0$. This is, under Assumption (A4), similar to the stationary error case, e.g. Brillinger (1994) (cf. the proof of (14)).

Note also that $\sigma^2_{jk} \geq C \cdot T^{-1}$ is fulfilled asymptotically as by (A3)b) we assume that $\inf_{u\in[0,1]} f(u,0)$ is uniformly bounded away from zero. It is, however, no problem to deal with those coefficients which violate that assumption, as is indicated in Section 2.2 of Neumann and von Sachs (1995) and is covered in detail in Neumann (1996), Section 4.

Lemma 2.2(ii) indicates how to (theoretically) choose the threshold in order for the main theorem (in Section 2.4) to hold. Some ideas, though still preliminary, on automatic threshold choice, more adapted to the situation of heteroskedasticity in the variance of the empirical wavelet coefficients, are discussed in Section 2.5.

2.4 Upper bound on the minimax risk in function space for locally stationary non–Gaussian errors

We prove our main theorem here by techniques developed, e.g. in Neumann and von Sachs (1995). The idea is to show that in the corresponding sequence space of empirical wavelet coefficients the considered L_2–risk between the unknown function $\mu(u)$ and our soft or hard threshold estimator $\widehat{\mu}(u)$ is asymptotically equivalent to the one of an accompanying Gaussian noise model, which is stated in equation (18) below. This can be shown by a strong form of asymptotic normality of the empirical wavelet coefficients, which puts emphasis on moderate and large deviations (see equation (16) below). For this, uniformly bounded cumulants of the errors in the function space model (2), as given by (A2), are sufficient. Once this equivalence has been established, we can apply the (theoretical) thresholding methods developed for the case of Gaussian noise.

As mentioned previously, we observe that, as in many analogous situations in curve estimation, the empirical coefficients at the scales j with $2^j = o\,(T)$ are asymptotically Gaussian. But, as a simple central limit theorem would not be sufficient for proving the desired risk equivalence to the case of Gaussian noise, the following stronger form of asymptotic normality needs to be derived.
Using Lemma 2.2 (ii) and (iii), by Lemma 1 in Rudzkis, Saulis and Statulevicius (1978), we can show that, with $\Phi(x)$ denoting the cumulative function of the standard normal distribution,

$$P\left(\pm(\widetilde{\beta}_{jk} - \beta_{jk})/\sigma_{jk} \geq x\right) = (1 - \Phi(x))(1 + o(1)) \tag{16}$$

holds uniformly in $(j,k) \in \mathcal{J}_T$ with $\sigma_{jk}^2 \geq CT^{-1}$, and uniformly on some interval $-\infty < x \leq \Delta_T$, $\Delta_T \asymp T^\eta$ for some $\eta > 0$. We omit the proof which runs completely analogous to the one of Neumann (1996), Theorem 4.1.

Now we can define an accompanying Gaussian model as

$$\xi_{jk} = \beta_{jk} + \varepsilon_{jk}, \quad (j,k) \in \mathcal{J}_T, \tag{17}$$

where $\varepsilon_{jk} \sim N(0, \sigma_{jk}^2)$. Essentially by integration by parts, it can be shown quite similarly to the situation of regression with stationary dependent errors as in Neumann and von Sachs (1995), Section 2.2, that

$$\sum_{(j,k)\in\mathcal{J}_T} E\left(\delta^{(\cdot)}(\widehat{\beta}_{jk}, \lambda_{jk}) - \beta_{jk}\right)^2 = (1+o(1)) \sum_{(j,k)\in\mathcal{J}_T} E\left(\delta^{(\cdot)}(\xi_{jk}, \lambda_{jk}) - \beta_{jk}\right)^2 + O\,(T^{-1}). \tag{18}$$

Here we need that $\mathcal{J}_T$ is such that

$$\sum_{(j,k)\notin\mathcal{J}_T} \beta_{jk}^2 = O\,(T^{-2m/(2m+1)})\ , \tag{19}$$

uniformly over $\mu \in \mathcal{F}$ where $\mathcal{F}$ is a ball in the Besov space $B^m_{p,q}$ under consideration. But this is automatically fulfilled if in $\mathcal{J}_T = \{(j,k) \mid 2^j \leq C\ T^{1-\delta}\}$ the $\delta > 0$ is not too large; e.g. $\delta \leq 1/3$ to achieve the optimal rate of L_2–convergence in $B^m_{p,q}$. Compare the discussion in Neumann (1996) along equation (4.1). A detailed proof of (18) is completely analogous to the proof of Neumann (1996), Theorem 5.1, which also covers the case where the additional assumption $\sigma_{jk}^2 \geq CT^{-1}$ is not fulfilled.
Although possible to do, we will ignore this case here. Thus, we have established the asymptotic risk equivalence to the situation of Gaussian noise, and obtain the following theorem from known results in Gaussian regression (see, e.g., Donoho and Johnstone (1992), Donoho et al. (1995)).

For this we distinguish between two different situations. We call

$$\widehat{\mu}^0(u) = \sum_k \widehat{\alpha}_k\ \phi_{j_0k}(u) + \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \delta^{(\cdot)}(\widehat{\beta}_{jk}, \lambda^0_{jk})\ \psi_{jk}(u) \tag{20}$$

the estimator that is based on an optimal (non–random) threshold $\lambda^0_{jk} = \lambda(T,j,k;\mathcal{F})$, whereas

$$\widetilde{\mu}(u) = \sum_k \widehat{\alpha}_k\ \phi_{j_0k}(u) + \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \delta^{(\cdot)}(\widehat{\beta}_{jk}, \lambda_{jk})\ \psi_{jk}(u) \tag{21}$$

is an estimator with (random) individual thresholds λ_{jk} which fulfill $\sigma_{jk}\ \sqrt{2\log(\#\mathcal{J}_T)} \leq \lambda_{jk} \leq C\ T^{-1/2}\sqrt{\log(T)}$ for a positive constant C.

Theorem 2.3 *Let $\mathcal{F}$ be a ball in a Besov space $B^m_{p,q}$, $m,p,q \geq 1$ Let $\widehat{\mu}^0$ and $\widetilde{\mu}$ be the wavelet threshold estimators defined by (20) and (21), respectively. Then,*

(i) for an optimal choice of the thresholds $\lambda^0_{jk} = \lambda(T,j,k;\mathcal{F})$,

$$\sup_{\mu\in\mathcal{F}}\left\{E\|\widehat{\mu}^0 - \mu\|^2_{L_2}\right\} = O\left(T^{-2m/(2m+1)}\right),$$

(ii) for thresholds λ_{jk} satisfying $\sigma_{jk}\ \sqrt{2\log(\#\mathcal{J}_T)} \leq \lambda_{jk} \leq C\ T^{-1/2}\sqrt{\log(T)}$ for any positive constant C,

$$\sup_{\mu\in\mathcal{F}}\left\{E\|\widetilde{\mu} - \mu\|^2_{L_2}\right\} = O\left((\log(T)/T)^{2m/(2m+1)}\right).$$

I.e., we have attained the "classical" rate $T^{-2m/(2m+1)}$ for the L_2–risk by exactly the same treatment of the empirical coefficients as in the Gaussian case.

This rate is attained for the optimal threshold (not known in practice, however), whereas the "price" to pay for some threshold rules which can be replaced by appropriate data–driven ones is the additional log term in the otherwise unchanged rate of convergence. Of course, a threshold that comes quite close is the one based on an estimator $\widehat{\sigma}^2_{jk}$ of the unknown variance σ^2_{jk}, e.g., $\widehat{\lambda}_{jk} = \widehat{\sigma}_{jk}\ \sqrt{2\log(\#\mathcal{J}_T)}$. An extensive discussion on appropriate data–driven threshold choice follows now.

2.5 Data–driven threshold choice

Lemma 2.2(ii) which emphasizes the localization of the wavelets in time, tells us how to (theoretically) choose the threshold in order for Theorem 2.3 to hold. Any threshold λ_{jk} satisfying $\sigma_{jk}\ \sqrt{2\log(\#\mathcal{J}_T)} \leq \lambda_{jk} \leq C\ T^{-1/2}\sqrt{\log(T)}$ for a positive constant C, will do; thus, the upper bound yields a *universal* threshold which is known to be very conservative, in particular, for a problem with heteroskedasticity. To avoid the resulting oversmoothing, a better rule would depend both on the scale j (as for stationary but correlated errors) and also on the location k. One possibility, of course, would be to determine the constant in the leading term of the asymptotic variance in (14) for each j and k, based, e.g., on a plug–in estimator for the unknown spectrum $f(u,0)$. In principle, all that is needed would be any consistent estimator which fulfills some type of uniform convergence criteria, without the need for certain rates (cf. Neumann (1996), Section 6); for example, a kernel estimator with an appropriate (global) bandwidth. However, in practice, this is certainly cumbersome, and moreover, second order effects are potentially neglected. Nevertheless, it should be emphasized that this is a possible rigorous approach.

However, here we suggest the use of another option which is based on some ideas that are easier to apply. This technique should be found to be quite generally useful whenever the asymptotic variance of the empirical wavelet coefficients is a fairly complicated functional (of both the incorporated wavelets and possibly an additional unknown quantity which needs to be estimated). The idea is to estimate the (finite sample) variance of the empirical wavelet coefficients for each fixed scale j directly from the sequence of $\{\widehat{\beta}_{jk}\}_k$.

Of course, as there are no replications of the $\widehat{\beta}_{jk}$ we need to appropriately pool or average over those with adjacent locations k. This is essentially a curve estimation problem for a local variance (as a curve), and the smoothing is done over neighboring values of squared empirical coefficients, the coefficients themselves being asymptotically Gaussian. This is possible since asymptotically, for each j, the unknown variance, σ_{jk}^2, behaves as a curve $\sigma_0^2(k/2^j)$, due to the model of local stationarity for the evolutionary spectrum $f(u,\omega)$, as a function of u. To give an example we study again, as at the end of Section 2.1, a modulated locally stationary process $\varepsilon_{t,T} = \sigma_{t,T}^0 \cdot Y_t$, where now the Y_t are i.i.d., and with $\sup_t \mid \sigma_{t,T}^0 - \sigma_0(t/T) \mid = O(T^{-1})$. Here we observe by comparison with (14) that for T large enough, $\sigma_{jk}^2 \sim \int \sigma_0^2(u)\ \psi_{jk}^2(u)\ du \sim \sigma_0^2(k/2^j)$, as the wavelets are concentrated around $k/2^j$. In other words, by (4), local stationarity of $f(u,\omega) = |A(u,\omega)|^2$ transfers to the variance, with the existence of a smooth $\sigma_0^2(u)$, such that

$$\sup_{jk} |\sigma_{jk}^2 - \sigma_0^2(k/2^j)| = o\ (1)\,, \text{ as } T \to \infty\ .$$

For this, an asymptotically increasing scale $j = j(T)$ is needed in order to be able to consistently estimate $\sigma_0^2(u_0)$ through a growing number of adjacent values of the variance estimator, with $k/2^j \to u_0$ as $j = j(T) \to \infty$. We could consistently estimate the truly level and location dependent variance functions by means of various smoothing procedures such as kernel smoothing with appropriately chosen bandwidths or, again, by wavelet thresholding applied on the sequence of the $\{\widehat{\beta}_{jk}\}_k$, for each level j. (This approach can be compared to both wavelet smoothing of periodograms, as in Neumann (1996), and of so–called "wavelet periodograms", as in von Sachs, Nason and Kroisandt (1996)). We believe that for our specific purposes a fit of locally constant segments would be sufficient, i.e. simple averaging of (the appropriate number of) adjacent values, which, e.g., can (and will be done in Section 3) by the use of Haar wavelets for fitting constant lines.

Some additional remarks on the dependency of the leading term of σ_{jk}^2 might be useful here. We observe that the asymptotics of Lemma 2.2(ii) puts emphasis of the localization of the considered wavelet coefficient and its variance in *time*, because it is an asymptotic expression holding for scales j coupled to the coarse scales ($2^j = o(T)$). Hence, it seeems that in formula (14) the frequency localization is completely lost, which, in particular, for the stationary error case would lead us to assume that the asymptotic variance is no longer dependent on the scale j (cf. the respective formula in Brillinger (1994)). This can be explained by the asymptotics of the classical curve estimation model which implicitly assumes an asymptotically decreasing correlation between two data points originally of fixed distance apart, a fact which seems not to be in accordance with the need for levelwise thresholding. As we assume that this is of general interest, not only for the locally stationary situation, we will discuss different model asymptotics where we picture our observations X_t no longer coupled to functional observations in rescaled time $t/T \in [0,1]$, but as genuine time series values of a stochastic process with a given correlation structure which is *not* modeled to change with asymptotically growing T. We feel that many data analysts would actually prefer this model to the widely used curve estimation model, and we describe the asymptotics of this alternative model (as used, e.g., also in von Sachs et al. (1996)) more completely in our discussion section 5.

Now, in our suggested approach, we need some practical guidelines for a proper segmentation within each scale j. First we observe that we must asymptotically keep away both from the coarsest and the finest scale because we need a growing number of locations k within each scale j in order to estimate, and also we need $2^j = o\ (T)$ for the asymptotic normality of the empirical wavelet coefficients. We suggest for each j with $2^j = o\ (T)$ to split the range of the sequence $\{\widehat{\beta}^j_k\}_k$ of length 2^j into $i = 1, ..., M_j$ segments $S_j^{(i)}$. We denote the length of each segment $S_j^{(i)}$ by $N_j^{(i)}$ which needs to be appropriately determined from the data. It must satisfy, however, $M_j \cdot N_j^{(i)} = 2^j$ with both M_j and $N_j^{(i)}$ growing to infinity as T (and $j = j(T)$) tend to infinity, but also such that $N_j^{(i)}/2^j \to 0$. Then the variance could be estimated by estimating the function $\sigma_0^2(u_i), i = 1, ..., M_j$ by using all the $N_j^{(i)}$ elements $\{\widehat{\beta}_{jk}\}_{k \in S_j^{(i)}}$ in the i–th segment, with $k/2^j \to u_i$. It is clear by the quasi–stationarity within the segments that we will get a consistent estimator of $\sigma_0^2(u_i)$. The actual determination of the appropriate segmentation for finite sample size T can be done, e.g., by fitting piecewise constant lines to the sequence of squared $\{\widehat{\beta}_{jk}\}_k$. For this a global smoothing parameter seems to be sufficient (e.g., a kernel bandwidth proportional to $(2^j)^{-1/(2m+1)}$, if the evolutionary spectrum f is of regularity m in u.).

Now, in order to circumvent the problem that the empirical coefficients need also to be centered around their correct (local) means in order to estimate the local variance, we suggest a slightly modified variant of the above general method, namely the use of a median absolute deviation (MAD) estimate as a measure of variability. Obviously the MAD estimator has to be approximately scaled in order to consistently estimate the standard deviation; that is we use MAD/.6745 as did Johnstone and Silverman (1997). Within each of the fitted segments we calculate the MAD of the included empirical wavelet coefficients, where in practice by the robustness of this estimator it is not really important to distinguish whether it is the deviation from the (unknown) mean, median or simply from zero. As long as the (local) signal–to–noise ratio in the wavelet coefficient domain is not too small, we expect a small sensitivity in the variance estimator to this problem of mean estimation. Here we aim to benefit from the same philosophy which makes wavelet thresholding (asymptotically) work. The level of the noise within each segment is assumed to be clearly smaller than the *sparse* signal contribution. (Of course, there are obvious limitations like, e.g, in nonparametric spectral density estimation for regions with high small peaks in the spectrum). Certainly for additive noise models and function classes for which wavelets do give sparse representations, this should be an adequate method to consistently estimate the square root of the leading term in the asymptotical variance σ_{jk}^2.

So the use of a (local) MAD should be superior to the use of a (local) empirical variance, which seems to be confirmed by our simulation examples of the next section. In practice, we use the MAD over these segments, and find the segments by fitting box–cars, i.e. perform Haar wavelet smoothing on the $\widehat{\beta}_{jk}^2$ (with a global smoothing parameter, i.e. a linear wavelet thresholder).

3 Numerical examples: Application to simulated and real data sets

In this section we demonstrate the performance of our wavelet threshold estimator for trend functions in the presence of nonstationary noise, both in a simulated and a real data example.

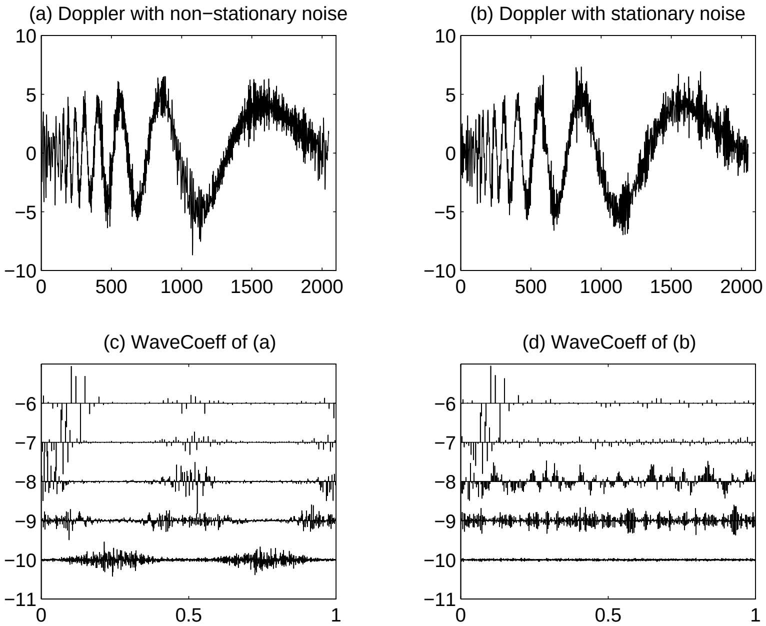

Figure 1: (a) Doppler signal ($T = 2048$) plus non–stationary noise, (b) Doppler signal plus stationary noise, (c) Wavelet coefficients of (a), (d) Wavelet coefficients of (b).

For the simulation we have chosen an example from Johnstone and Silverman (1997) in order to compare our level and location dependent threshold rule, introduced in the previous section, with the threshold rule used in Johnstone and Silverman (1997), Section 3 for for stationary noise. A levelwise MAD threshold estimator (see equation (22) below), which is a common robustified rule in practice is also used here. The simulated example is the by now famous Doppler signal, an artificial function of spatially varying frequency used by Donoho and Johnstone (1994) in the presence of i.i.d. Gaussian noise, which is sampled at $T = 2048$ equally spaced grid points of $[0, 1]$. In Figure 1 we first show the simulated signal (a) with nonstationary noise and, for comparison, (b) with the same stationary noise as used by Johnstone and Silverman (1997). The first noise process is a time–varying autoregressive process of order 2, i.e. a locally stationary process following model (2), as given in an example by Dahlhaus (1997), i.e. with parameter functions $a_1(u) = -1.8 \cdot \cos(1.5 - \cos(4\pi \cdot u))$, and $a_2(u) = 0.81$, with $0 \leq u \leq 1$. The second one is a stationary AR(2)–process $X_t + a_1 X_{t-1} + a_2 X_{t-2} = z_t$ with parameters $a_1 = -4/3, a_2 = 8/9$. Both AR–models are driven by Gaussian i.i.d. noise $\{z_t\}$. The variance of $\{z_t\}$ is scaled such that in both situations the resulting noise processes $\{X_t\}$ has the same variance, chosen in order to match the signal-to-noise ratio of the corresponding example in Johnstone and Silverman (1997).

The subplots (c) and (d) then clearly show the difference of the nature of the noise on the higher levels in the wavelet coefficient space. In particular on levels 8–10 in the first example, the nonstationary nature of the noise is clearly exhibited. In Figure 2 and Figure 3 the performance of various threshold rules for these two examples is shown. In the left column can be seen the remaining wavelet coefficients, on the interesting levels 6–10, after

thresholding, in the right columns the corresponding reconstructions (all based on (hard) thresholding the levels 6–10, only, as done in Johnstone and Silverman (1997)).
Three different types of estimators are used: In (a) and (b), called "Local MAD", we applied precisely the threshold procedure described in the previous section: After fitting piecewise constant lines, by linear Haar smoothers, to the squared empirical wavelet coefficients $\{\widehat{\beta}_{jk}^2\}$ on each fixed level j, we pooled together all those which correspond to one fitted segment, the $i-$th segment $S_j^{(i)}$, say, $i = 1, \ldots, M_j$, and applied Local MAD $\widehat{m}_j^{(i)}$ to find the appropriate threshold for those coefficients. As always, and as in Johnstone and Silverman (1997) equation (7), this MAD has to be rescaled by a constant to match calibration with the Gaussian distribution, i.e.

$$\widehat{m}_j^{(i)} = \mathrm{MAD}\{\widehat{\beta}_{jk}, k \in S_j^{(i)}\}/.6745 \qquad i = 1, \ldots, M_j \ , \tag{22}$$

where MAD denotes the median absolute deviation from zero, and where we use the notation introduced at the end of Section 2.5.
Then a threshold $\lambda_j^{(i)} = \widehat{m}_j^{(i)} \sqrt{2\log(2^j)}$ is applied to all coefficients in segment $S_j^{(i)}$; it can be seen as a locally universal threshold for the segment $S_j^{(i)}$. In other words, we use semi–individual thresholds $\lambda_{jk} = \lambda_j^{(i)}$ for all $k \in S_j^{(i)}$. This approach is justified by the considerations of Section 2.5, where we observed that asymptotically $\widehat{m}_j^{(i)}$ tends to $\sigma_0(u_i)$ for all k with $k/2^j \to u_i$, i.e. for all $k \in S_j^{(i)}$, and such that $\#(S_j^{(i)})/2^j = N_j^{(i)}/2^j \to 0$ as $T \to \infty$.
In (c) and (d) the same MAD is used levelwise, i.e. globally over each level j, which results in exactly the same estimator as in Johnstone and Silverman (1997). And finally, in (e) and (f) the levelwise MAD (labeled "Level MAD") is replaced by a levelwise empirical variance estimator (labeled "Level VAR") $\widehat{\sigma}_j^2 = 2^{-j} \sum_k \left(\widehat{\beta}_{jk} \ - \ 2^{-j} \sum_k \widehat{\beta}_{jk}\right)^2$.
For all wavelet transforms Daubechies Symmlets of order 8 were used, as in Johnstone and Silverman (1997).

In Figure 2 we observe that both the traditional levelwise MAD and levelwise variance estimator cannot cope with the nonstationary noise, whereas Local MAD does quite well. We also observe its limitations for finite sample size performance: on levels where the local signal-to-noise ratio deteriorates, signal contribution will be killed by (any kind of) thresholding. This happens in this example in local regions on levels 6 and 7 leading to a slightly unsatisfactory reconstruction at the beginning of the Doppler signal. Increasing the sample size T would, of course, help to cope with this defect as it improves the (local) signal-to-noise ratio for the region of high signal oscillation, in a model of locally stationary noise (i.e. for nonstationary noise with an asymptotically fixed spectral behaviour).
Another interesting aspect is the performance of the local MAD estimator in situations with truly *stationary* noise where it performs as well as both levelwise estimators, which can be observed in Figure 3.

In our second numerical example, we now apply exactly the same estimators studied above, to a real data set, which is an example for a typical biomedical time series with non–homoskedastic noise. In Figure 4(a) 512 data points of a set of sheep luteinizing hormone (LH) concentrations in blood samples from sheep collected twice weekly for a period of 256 consecutive weeks are shown. These data have been collected and investigated by (Karsch et al. (1989)), and were also used in (Normolle and Brown (1994)) to identify aperiodic seasonalities in possibly nonstationary, non–Gaussian time series. For reasons of space,

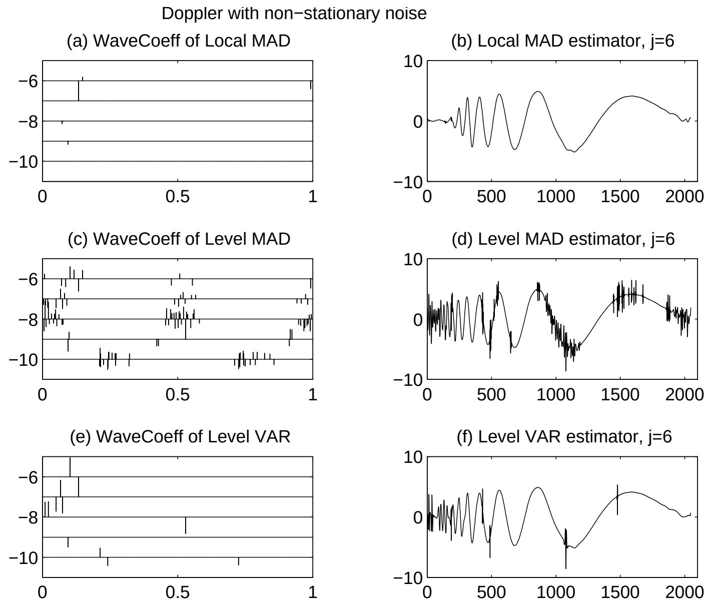

Figure 2: Doppler signal ($T = 2048$) with non–stationary noise. Wavelet coefficients and reconstruction for (a), (b): Local MAD; (c), (d): Levelwise MAD; (e), (f): Levelwise VAR (all estimators based on hard thresholding of levels 6–10).

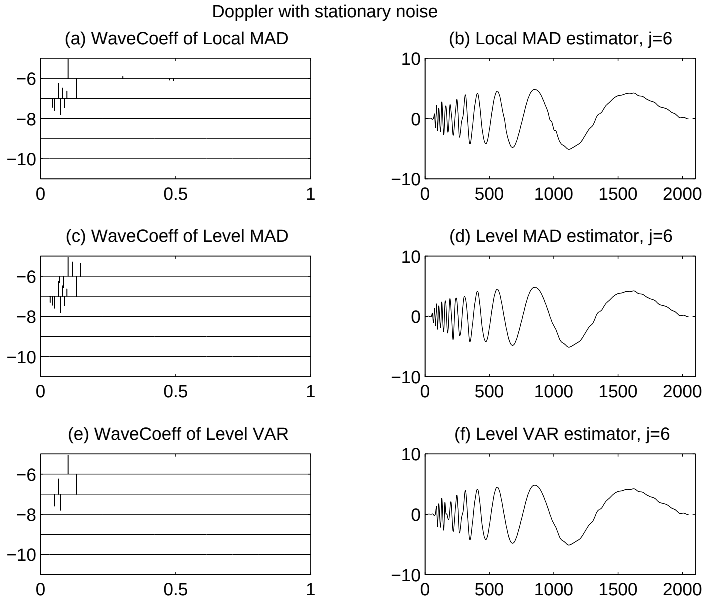

Figure 3: Doppler signal ($T = 2048$) with stationary noise. Wavelet coefficients and reconstruction for (a), (b): Local MAD; (c), (d): Levelwise MAD; (e), (f): Levelwise VAR (all estimators based on hard thresholding of levels 6–10).

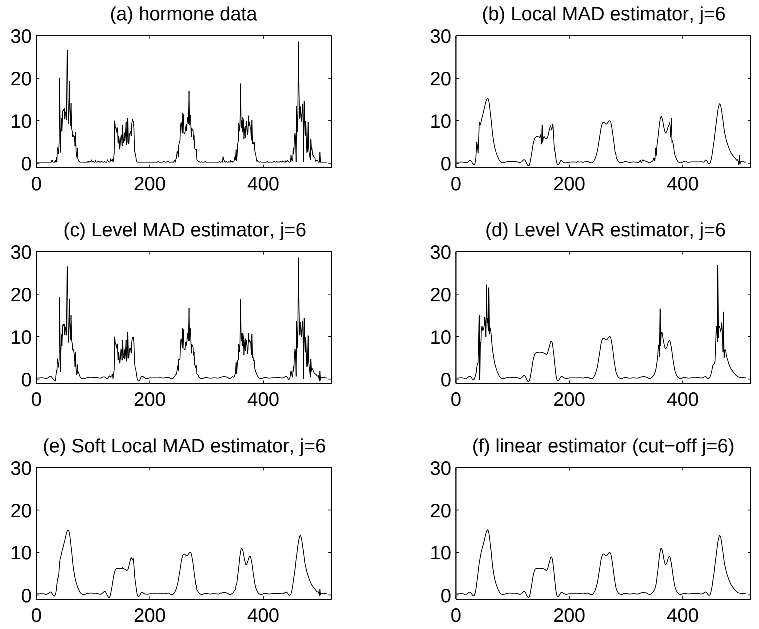

Figure 4: (a) hormone data ($T = 512$), (b) Local MAD, (c) Level MAD, (d) Level VAR (all estimators based on hard thresholding), (e) Soft Local MAD, (f) Linear wavelet estimator (cut–off scale j=6).

here, we have chosen to analyze only one set of data, those, displayed also in Figure 1 (top) of Normolle and Brown (1994), that were taken from a control group of sheep that was held outdoors. Typically, the goal of those studies is to compare variation in the cyclic behavior in the hormone level of animals under different external conditions. Quite naturally, the regularity of these variations may be both disrupted by experimental interventions and also be subject to sudden changes in the generating process, such as pulses in the series of hormone measurements. Hence, it is quite a challenge, in particular for a purely non–parametric procedure, to distinguish between time–variation due to nonstationary noise and significant changes (here cyclic seasonalities) in the trend function modeling the signal behavior itself. It is exactly this problem which lead (Normolle and Brown (1994)) to use some clustering algorithm designed specifically for this application rather than to rely on the use of traditional function fitting procedures. However, we believe, that the use of *localized* and adaptive non–parametric estimators like orthogonal wavelet series may indeed compete with more specifically designed procedures.

Again, we use Local MAD (shown in (b)), levelwise MAD (c) and variance estimator (d) based on hard thresholding with exactly the same rules as described above, plus, for extra comparison, a local MAD based on soft thresholding in (e). Again, thresholding was applied to the highest levels 6 through 8, the result of which can be seen in the full wavelet coefficients tableaus of Figure 5, for (a) through (d). For comparison, in Figure 4 (f), a linear wavelet estimator with cut–off level 6 is displayed. This is the wavelet series of the data, with all coefficients set to zero above level 5 (and no shrinkage applied elsewhere). As in the simulated example, we observe that the levelwise MAD (c) hardly shrinks any coefficients, and also the levelwise VAR (d) undersmooths, whereas, in particular, the soft

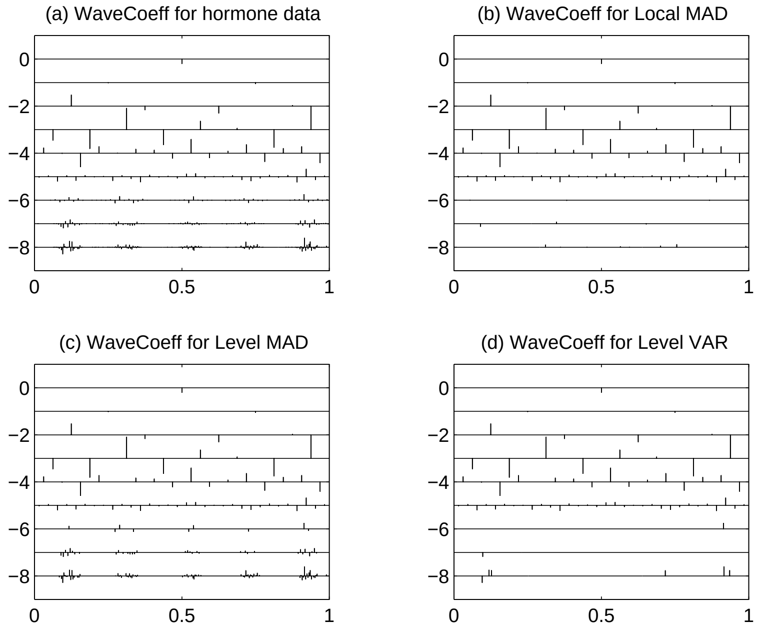

Figure 5: Wavelet coefficients for (a) hormone data ($T = 512$), (b) Local MAD, (c) Level MAD, (d) Level VAR (all estimators based on hard thresholding).

threshold version of the local MAD (e) does almost as well as does the linear estimator (f). Obviously, in this specific example a fit which cuts–off all oscillations on levels 6 and higher might be the most appropriate one in order to display the five seasonal patterns with peaks of the LH level in the fall (which is due to the biology of the sheep being in estrus during in fall). But again, in this example, it can be observed in particular in the coefficient plots of Figure 5, that due to the clear heteroskedasticity of the noise variance it is not sufficient to apply a purely levelwise thresholding.

4 Lower bounds on minimax risk for Gaussian locally stationary errors

4.1 Comparison of the wavelet threshold estimator risk to the ideal risk

In the case of locally stationary errors as shown in Section 2.4, an upper bound on the minimax risk exists for "smooth" function estimation. This was derived by techniques which work independently of the assumption of normality. In this section, we show that in the particular situation of Gaussian noise, inspired by the work of Donoho and Johnstone (1994) in the i.i.d. case, and Johnstone and Silverman (1997) and Johnstone (1996) for stationary errors, it is also possible to achieve a lower bound on the l_2–minimax rate in the wavelet coefficient domain. We prove that a universal location and level–dependent soft threshold for the discrete wavelet transform of the data in the presence of locally stationary errors achieves the lower bound of the minimax risk by comparing its behaviour to that of an ideal but unattainable benchmark obtained from an "oracle" that provides the optimum diagonal projection estimate. As previously explained by Donoho and Johnstone (1994) the benchmark risk can be considered as a measure of the sparsity of the wavelet expansion of the function μ. This ideal or benchmark risk can be obtained as follows.

Following Johnstone and Silverman (1997) we can use the general multivariate normal model to (abstractly) describe the signal–plus–noise model in the wavelet coefficient domain.

We suppose even more generally that the observations $Y_{jk,T}$ are generated according to

$$Y_{jk,T} \;=\; \theta_{jk} + z_{jk,T}\ , \tag{23}$$

where for (j,k) satisfying $j \leq J = \log_2(T)$, $k = 0,\ldots,2^j - 1$, the $z_{jk,T}$ have a T-variate normal distribution with mean zero and covariance matrix V_T. Note that the elements of V_T are allowed to vary with T. In order to simplify notation in this section, henceforth, i, $(i = 1,\ldots,T)$ will now denote the double index (j,k) defined above. The additional subscript T is also used to emphasize the connection to the doubly–indexed locally stationary regression model (2). In our model, the θ_i's and the $Y_{i,T}$'s would represent wavelet coefficients β_{jk} and empirical ones $\widehat{\beta}_{jk}$, respectively, and the matrix V_T, the variance of the empirical wavelet coefficients.

The i-th diagonal element of V_T is denoted by $v^2_{ii,T}$ and equals $\mathrm{var}(Y_{i,T})$ for each $i = 1,\ldots,T$. Furthermore, let

$$\overline{v^2_T} = T^{-1}\sum_{i=1}^{T} v^2_{ii,T} = T^{-1}\ tr(V_T)\ . \tag{24}$$

As the data here is "Gaussian", the inverse of the covariance matrix, V_T^{-1} exists. Let $v^{lm,T}$ denote the (lm)-th element of V_T^{-1}. Write $\tau^2_{ii,T} = 1/v^{ii,T}$, and set

$$\overline{\tau^2_T} = T^{-1}\sum_{i=1}^{T}\tau^2_{ii,T}. \tag{25}$$

Also as observed by Johnstone and Silverman (1997), note that $\tau^2_{ii,T} \leq v^2_{ii,T}$ for all i, as $\tau^2_{ii,T} = \mathrm{var}(Y_i|Y_m, i \neq m)$ for multivariate normally distributed $(Y_1,\ldots,Y_T)$.

In order to obtain the ideal risk it is assumed that there is an oracle (cf. Donoho and Johnstone (1994), Johnstone and Silverman (1997)) which indicates whether to kill or keep each coordinate θ_i for all diagonal projection (DP) estimators: $\widehat{\theta}_i = \delta_i x_i$ with $\delta_i = 0$ or 1. The ideal risk $R(DP;\theta)$ is the risk obtained under the ideal choice of the sequence $\{\delta_i\}$. Clearly here

$$R(DP;\theta) = \sum_{i=1}^{T}(\theta_i^2 \wedge v^2_{ii,T})$$

and this risk is attained if

$$\delta_i = I[\theta_i^2 \geq E(Y_{i,T} - \theta_i)^2] = I[\theta_i^2 \geq v^2_{ii,T}].$$

We now consider the optimal choice of thresholds for soft threshold estimators.
It is well known (e.g. Donoho and Johnstone (1994)) that if $Y \sim N_T(\theta, v^2\mathrm{I})$ and if for each $i = 1,\ldots,T$

$$\widetilde{\theta}_i = \delta^{(s)}(Y_i, \sqrt{2\log T}\,v) \tag{26}$$

$$\text{then}\qquad E\ \|\widetilde{\theta} - \theta\|^2 \leq (1 + 2\log T)\{v^2 + \sum_{i=1}^{T}(\theta_i^2 \wedge v^2)\} \tag{27}$$

Now, clearly, arguing as in Donoho and Johnstone (1994) and Johnstone and Silverman (1997) we have the following result for soft threshold estimators. We shall concentrate on soft thresholding although we conjecture the results remain true for the analogous hard thresholding.

Theorem 4.1 *In the model given in (23) let us assume that the $z_{jk,T}$ have a Gaussian distribution such that the eigenvalues of its covariance matrix V_T are uniformly bounded away from 0 and ∞ for all $T \geq 1$.*

Then

$$E\ \|\widetilde{\theta} - \theta\|^2 \leq (1 + 2\log T)\{\overline{v_T^2} + \sum_{i=1}^{T}(\widetilde{\theta}_i \wedge v_{ii,T}^2)\}. \tag{28}$$

where $\widetilde{\theta}_i = \delta^{(s)}(Y_i, \sqrt{2\log T}\, v_{ii,T})$.

Moreover:

$$\liminf_{T\to\infty} \frac{1}{2\log T}\ \frac{\overline{v_T^2}}{\overline{\tau_T^2}}\ \inf_{\widetilde{\theta}\in\Theta_T}\ \sup_{\theta\in\Re^T} \frac{E\ \|\widetilde{\theta} - \theta\|^2}{\overline{v_T^2} + \sum_{i=1}^{T}(\theta_i^2 \wedge v_{ii,T}^2)} \geq 1, \tag{29}$$

where Θ_T = set of all estimators of θ based on the $\{Y_i\}$, not just threshold estimators and

$$\overline{v_T^2} + \sum_{i=1}^{T}(\theta_i^2 \wedge v_{ii,T}^2) \tag{30}$$

represents the ideal risk of the best diagonal projection.

The fact that (28) is true, given that (27) is true, was observed in Johnstone and Silverman (1997).

As in Donoho and Johnstone (1994) and Johnstone and Silverman (1997), the main idea of the proof here is to bound the minimax risk in (29) by the Bayes risk relative to a well-chosen point prior on θ. It is clear that a key role will be played by the modified loss function $\widetilde{L}_T$ for θ^*, an arbitrary estimator of θ.

$$\widetilde{L}_T(\theta^*, \theta) = (\overline{v_T^2})^{-1}\ p_T(\theta)^{-1} \sum_{i=1}^{T}(\theta_i^* - \theta_i)^2 \tag{31}$$

and with $p_T(\theta)$ defined for each $\theta \in \Re^T$ by

$$p_T(\theta) = 1 + (\overline{v_T^2})^{-1} \sum_{i=1}^{T}(\theta_i^2 \wedge v_{ii,T}^2)\ . \tag{32}$$

It suffices to show that the relatively mild conditions imposed on the covariance matrix V_T, that is, the existence of positive constants C_0 and C_1, such that

$$\lambda_{min}(V_T) \quad \geq \quad C_0 > 0 \quad \forall\, T \geq 1 \tag{33}$$

$$\lambda_{max}(V_T) \quad \leq \quad C_1 < \infty \quad \forall\, T \geq 1$$

guarantees the (uniform) boundedness of certain ratios of (averaged) moments and covariances needed to prove the theorem. The proof itself is outlined in the Appendix.

We have the following immediate corollary in the wavelet setting for the threshold estimator (12) with appropriate soft thresholding (11):

$$\widetilde{\beta}_{jk} \quad = \quad \delta^{(s)}(\widehat{\beta}_{jk}, \sqrt{2\log T}\ v_{jk})$$

where $\widehat{\beta}_{jk}$ is the empirical wavelet coefficient defined in (9) and v_{jk}^2 denotes its variance.

For convenience, using the single index notation of this section will will denoted by

$$\widetilde{\beta}_i = \delta^{(s)}(\widehat{\beta}_i, \sqrt{2\log T}\ v_{ii,T}) \tag{34}$$

and V_T will denote the covariance matrix of the empirical wavelet coefficients.

Corollary 4.2 *In any Gaussian nonstationary time domain model including the locally stationary error model (2), where the Gaussian errors $\varepsilon_{jk,T}$ satisfy condition (A2) and the eigenvalues of its covariance matrix Γ_T are uniformly bounded away from* 0 *for all* $T \geq 1$,

then for $\widetilde{\beta}$ defined in (34),

$$\boldsymbol{E}\ \|\widetilde{\beta} - \beta\|^2 \leq (1 + 2\log T)\{\overline{v_T^2}\ +\ \sum_{i=1}^{T}(\widetilde{\beta}_i \wedge v_{ii,T}^2)\}. \tag{35}$$

Moreover:

$$\liminf_{T\to\infty} \frac{1}{2\log T} \frac{\overline{v_T^2}}{\tau_T^2} \inf_{\widetilde{\beta}\in\mathbf{B}_T} \sup_{\beta\in\Re^T} \frac{\boldsymbol{E}\ \|\widetilde{\beta} - \beta\|^2}{\overline{v_T^2} + \sum_{i=1}^{T}(\beta_i^2 \wedge v_{ii,T}^2)} \geq 1, \tag{36}$$

where $\mathbf{B}_T$ = set of all estimators of β based on the empirical wavelet coefficients $\widehat{\beta}$.

Of course, in practice the variance in (34) must be estimated. Now in the wavelet threshold setting with locally stationary errors, as previously explained in Sections 2 and 3, this can be done either by exploiting the leading term of the asymptotic variance in (14) which would call for a plug-in estimator for the unknown evolutionary spectrum $f(u, 0)$. We advertise the use of a segmented variance estimator, e.g. the local MAD $\hat{m}_j^{(i)}$ as in (22), which approximates the square root of the variance, i.e. σ_{jk}, for each k in the segment $S_j^{(i)}$ $(i = 1 \ldots M_j)$ within a fixed level j.

The actual estimator proposed, to which we refer as a semi–location level dependent threshold estimator, can be written as follows: within a fixed level j, for $k \in S_j^{(i)}$,

$$\widetilde{\beta}_{jk} \quad = \quad \delta^{(s)}(\widehat{\beta}_{jk}, \sqrt{2\log T}\ \hat{m}_j^{(i)}).$$

4.2 A proposal for threshold estimators based on Stein's unbiased estimate of risk

Now, we will try to improve the threshold estimators by introducing an adaptive procedure based on minimizing the Stein unbiased estimate of the risk for threshold estimates. This would allow to get rid of the log–term in the resulting L_2–risk by using thresholds which are less conservative than those of the previous section. For the white noise model and the stationary error model, Donoho and Johnstone (1995) and Johnstone and Silverman (1997) respectively successfully used such a method by assigning a threshold level to each (dyadic) resolution level, where the variance is homoskedastic, and then minimizing the Stein unbiased estimate of risk at this level. Here, however, we must cope with the problem of heteroskedastic errors at each level and incorporate minimizing the Stein unbiased estimate of risk into our semi–location level dependent thresholding procedure. A more precise description follows.

The following extension of Stein's lemma Stein (1981) applied to the heteroskedastic case can be shown to be true.

Let Y_i $(i = 1 \dots T)$ be generated by model (23). Now the soft threshold $\delta^s(Y, \tau)$ defined in (11) can be re-expressed as $Y + g(Y)$ where g is defined coordinatewise as follows:

$$g_i(Y_i) = \begin{cases} -\tau & Y_i > \tau \\ -Y_i & |Y_i| \le \tau \\ +\tau & Y_i < -\tau \end{cases} \tag{37}$$

then

$$E\|\delta^{(s)}(Y,\tau) - \theta\|^2 = \sum_i v_{ii,T}^2 + E\sum_i (Y_i^2 \wedge \tau^2) - 2\sum_i v_{ii,T}^2 \; E \; \mathrm{I}\{|Y_i| \le \tau\} = E \; U^*(\tau, Y) \tag{38}$$

Now, the above expression differs from that in the i.i.d. case (Donoho and Johnstone (1995)) or the homoskedastic case (Johnstone and Silverman (1997), Johnstone (1996)), but we could choose our adaptive estimator analogously to minimize this unbiased risk estimate, that is,

$$\widehat{\tau}(Y) = \mathrm{argmin}_{0 \le \tau \le \sqrt{2\log T}} \widehat{U}(\tau, Y)\,, \tag{39}$$

where

$$\widehat{U}(\tau, Y) \;=\; \sum_i \widehat{v}_{ii,T}^2 + \sum_i (Y_i^2 \wedge \tau^2) - 2\sum_i \widehat{v}_{ii,T}^2 \; \mathrm{I}\{|Y_i| \le \tau\}$$

is an unbiased estimator of U^*, with appropriate estimators $\widehat{v}_{ii,T}^2$ of the variances $v_{ii,T}^2$ in (38).

As in Donoho and Johnstone (1995) it might be possible to improve this estimator with the use of a pretest with a given fixed threshold γ_T. Let $s_T^2 = \left(\frac{1}{T}\sum_{i=1}^{T}(Y_i^2 - \widehat{v}_{ii,T}^2)\right)$ denote an unbiased estimator of $\|\theta\|^2$. Then

$$\begin{aligned} \tilde{\tau}(Y) \;&=\; \sqrt{2\log T} \qquad s_T^2 \le \gamma_T^2 \\ &=\; \widehat{\tau}(Y) \qquad s_T^2 > \gamma_T^2 \end{aligned} \tag{40}$$

As in Section 5.1 we propose to use the local MAD estimate for the variance in the SURE estimator developed in Johnstone and Silverman (1997) and Johnstone (1996) over these segments and define for each $k \in S_j^{(i)}$:

$$\lambda_{jk} = \tilde{\tau}\Big(\frac{\widehat{\beta}_{jk}}{\widehat{m}_j^{(i)}}\Big) \tag{41}$$

and our semi–location level dependent SURE threshold estimator is defined with this threshold.

$$\widetilde{\beta}_{jk} = \delta^{(s)}(\widehat{\beta}_{jk}, \hat{m}_j^{(i)} \lambda_{jk}) \ , \quad k \in S_j^{(i)} \ , \tag{42}$$

where the λ_{jk} is set to zero for j below a certain level.

Let $b^m_{p,q}(C)$ denote the Besov family in sequence space form; that is, for $s = m + \frac{1}{2} - \frac{1}{p}$,

$$b^m_{p,q}(C) = \{\beta_{jk} : \sum_{j=0}^{\infty} 2^{jsq} \|\beta_j\|_p^q \leq C^q\}.$$

If the parameters (m, p, q, C) were all known then let the minimax threshold risk be denoted by

$$\inf_{\lambda_{jk}} \sup_{b^m_{p,q}(C)} E\|\beta^* - \beta\|^2 \tag{43}$$

where $\beta^* = (\delta^{(\cdot)}(\beta_{jk}, \lambda_{jk}))_{j,k}$ is a threshold estimator based on thresholds λ_{jk}.

We conjecture that we could then argue as did Johnstone and Silverman (1997) and Johnstone (1996) that the SURE estimator defined in (42) using segmented variance estimators and the proper choice of pretest thresholds get the threshold right asymptotically without needing to know the specific parameters of the Besov smoothness class; that is, asymptotically,

$$\sup_{b^\sigma_{p,q}(C)} E\|\widetilde{\beta} - \beta\|^2 \tag{44}$$

is bounded by (43). A proof could likely follow the arguments of Johnstone (1996) for the case of fractional Brownian motion. It would be too lengthy to discuss these details here.

5 Some further comments

5.1 A different asymptotic model for correlated errors

As was observed in Section 2.5, the asymptotics of Lemma 2.2(ii) emphasizes the localization of the considered wavelet coefficient and its variance in *time* and hence on the coarse scale behaviour. A different possibility for doing asymptotics in regression with correlated errors starts from a model where the obse rvations $X = (X_t)_{t=1,\ldots,T}$ are no longer coupled to functional observations $\mu(t/T) + \varepsilon_t$ with empirical wavelet coefficients as in (9) with a variance of order $O(T^{-1})$. Rather we adopt the point of view of Section 4 (which is widely used, e.g, also in Johnstone and Silverman (1997), Section 2.1), where $\widehat{\beta}_{jk} = (WX)_{jk}$ with a variance of order $O(1)$ (because the orthogonal wavelet weights $W_{jk} = \psi_{jk}$ are now normalized to be in $L_2[0,T]$). This has the advantage of modelling the correlation structure of X so as not to change asymptotically; hence even for stationary X the dependency of the variance of $\widehat{\beta}_{jk}$ on the scale j is no longer lost.

Now let $c(s), s \in \mathbb{Z}$, and $f(\omega)$ denote the autocovariance function and spectrum, respectively, of the stationary time series (X_t). Then

$$\mathrm{var}\{\widehat{\beta}_{jk}\} \ = \mathrm{var}\{\sum_t X_t \ \psi_{jk}(t)\} \ = \sum_s c(s) \ \sum_t \psi_{jk}(t) \ \psi_{jk}(t+s) \ = \int_{-\pi}^{\pi} f(\omega) \ |\widehat{\psi_j}(\omega)|^2 \ d\omega \ ,$$

where we used the fact that

$$\sum_s \sum_t \psi_{jk}(t)\ \psi_{jk}(t+s)\ \exp(-i\omega s) \ = \ |\widehat{\psi_{jk}}(\omega)|^2 \ = \ |\widehat{\psi_j}(\omega)|^2 \ ,$$

as the dependency on k drops out in the squared modulus of the Fourier transform.

We observe that, in contrast to equation (14) the variance remains dependent on scale j, as in the integral over frequency the weight function $|\widehat{\psi_j}(\omega)|^2$ is a certain type of frequency bandpass. We believe that there is a certain advantage in the use of this model which pictures non–parametric regression as estimation of the trend function of a time series. It allows us to study both extreme cases of fine scale and of coarse scale behaviour: If we do asymptotics with the scale j tending to infinity then $|\widehat{\psi_j}(\omega)|^2$ eventually becomes very flat, the integral degenerates to a (weighted) integral over the whole spectrum, and the leading term of the asymptotic variance becomes proportional to $c(0)$, incorporating (almost) all frequencies. In the other extreme, the coarse scale approximation, $|\widehat{\psi_j}(\omega)|^2$ behaves like a delta function concentrated around zero, and only the long–term behavior of the errors is captured, i.e. around zero frequency in the spectrum. Hence we retrieve the expression for the constant in the leading term of equation (14), and the connection back to the classical curve estimation model, as being used in Section 2.

Finally, to do the same in the corresponding model of a *locally stationary* process, quite naturally, both a time and a frequency localization can be observed in the variance of the empirical wavelet coefficients. This has been studied in von Sachs et al. (1996), and hence we cite the results from Lemma 3.4, equation (3.17) of von Sachs et al. (1996)), i.e.

$$\mathrm{var}\{\widehat{\beta}_{jk}\} \ = \ \sum_t \int_{-\pi}^{\pi} f_{t,T}(\omega)\ W_{jk}^{\psi}(t,\omega)\ d\omega$$

with $\quad f_{t,T}(\omega) := \sum_s \mathrm{cov}\{X_{t,T}; X_{t+s,T}\}\ \exp(-i\omega s)$
and $\quad W_{jk}^{\psi}(t,\omega) := \sum_s \psi_{jk}(t)\ \psi_{jk}(t+s)\ \exp(-i\omega s)$. Here we observe that, for the fine scales (coupled to the asymptotically growing finest scale $J = \log_2(T)$),

$$\mathrm{var}\{\widehat{\beta}_{jk}\} \ \asymp \ \int_{-\pi}^{\pi} f(u_0,\omega)\ |\widehat{\psi_j}(\omega)|^2\ d\omega \ , \qquad\qquad j = j(T) \to \infty \ ,$$

for all k with $k/2^j \to u_0$, as $T \to \infty$.
Hence, we have the corresponding *time–dependent* analog of the stationary case.

5.2 Conclusion

We have studied non–linear wavelet thresholding for nonparametric regression with correlated and non–stationary errors. In our work, the model of local stationarity was useful as one possibility to control deviations from the classical situation. Quite generally we have observed how some more subtle (asymptotic) investigations of the variance of the empirical wavelet coefficients could give insights into the question of appropriate threshold choice. In particular, we discussed a new approach of how to possibly circumvent plug–in rules with their need for good pilot estimators in determining the threshold directly from the data. We suggested the use of a local MAD to estimate the variability of the empirical wavelet coefficients in segments of appropriate length of quasi–stationarity within one level. This generalizes the approach of Johnstone and Silverman (1997) where the use of only level–dependent thresholds was sufficient because of the true stationarity within each level of coefficients. We demonstrated the performance of our new approach both in a simulated Doppler example, to which we added a locally stationary AR(2)–noise process, and a stationary counterpart, for comparison, and also with a real data set of sheep hormone level measurements.

From the theoretical point of view we showed that under fairly moderate assumptions on the errors, i.e. uniformly bounded cumulants, the near–optimal L_2–rate between wavelet threshold estimator and the true unknown function, as a member of a Besov class, is attained. Moreover, in the situation of *Gaussian* errors with quite a general form of a (non–stationary) covariance matrix with uniformly bounded eigenvalues, we also investigated the lower minimax bound on the asymptotic minimax rate in the wavelet coefficient domain, by comparison to the ideal benchmark risk. In this respect we generalized again results from Johnstone and Silverman (1997) and Johnstone (1996). Finally we briefly indicated a possible approach to adaptive SURE threshold choices for the locally stationary situation.

We conclude with the observation that though there certainly exist well–suited alternatives to treat time–varying correlation structure in the problem of trend estimation of a possibly nonstationary time series, soft or hard wavelet thresholding (or a mixture of these) is one possibility which, in spite of its limitation to the case of sufficiently high local signal-to-noise ratio, seems to be as promising as it appeared for curve estimation with stationary errors.

6 Appendix

In this section we give the remaining proofs.

Proof of Lemma 2.2:

(i) With $E\ X_{t,T} = \mu(t/T)$, there is no stochastic bias. Hence,

$$E\ \widehat{\beta}_{jk} = T^{-1} \sum_t \mu(t/T)\ \psi_{jk}(t/T) \ = \ \int_0^1 \mu(u)\ \psi_{jk}(u)\ du \ + \ O\ (2^{j/2} \cdot T^{-1}) \ ,$$

as both μ and ψ are at least of bounded total variation.
(ii) Let

$$c_T(t/T, s) := \mathrm{cov}\{\varepsilon_{[t-s/2],T}; \varepsilon_{[t+s/2],T}\} \ ,$$

and let

$$c(u,s) := (2\pi)^{-1} \int_{-\pi}^{\pi} f(u,\omega)\ \exp(i\omega s)\ d\omega \ = \ (2\pi)^{-1} \int_{-\pi}^{\pi} |A(u,\omega)|^2\ \exp(i\omega s)\ d\omega \ .$$

Observe that, by (4),

$$c_T(u,s) = (2\pi)^{-1} \int_{-\pi}^{\pi} A([uT - s/2]/T, \lambda)\ \overline{A([uT + s/2]/T, \lambda)}\ \exp(i\lambda s)\ d\lambda \ + \ O\ (T^{-1}) \ .$$

The proof now runs similarly to the proof of Lemma 3.2 in Neumann and von Sachs (1997). Let $A_t(\omega) := A(t/T, \omega)$, and let,

$$R_T = \sum_{s,t} (c_T(t/T, s) \ - c(t/T, s)) \ .$$

Then, R_T is composed of two similar terms for each of which the estimate

$$|R_T| \;\leq\; \sum_s \sum_t \,|\int \overline{A_t(\omega)} \sum_{n=0}^{s/2-1} \{A_{t-n}(\omega) - A_{t-n-1}(\omega)\} \exp(i\omega s)\, d\omega|$$

holds, such that, by (A4) (a), (b), for some positive constant C,

$$|R_T| \;\leq\; \sum_s |s|/2 \,\sum_\ell \sup_u |\widehat{A}(u, s-\ell)|\; TV_{[0,1]}(\widehat{A}(., \ell)) \;\leq C,$$

That is,

$$T^{-2} \sum_{t,s} (c_T(t/T, s) \; - c(t/T, s)) \;\; \psi_{jk}(t/T) \;\psi_{jk}((t+s)/T) \;\; = \;\; O\;(2^j \; T^{-2}) \; .$$

Further, we want to use that (A4)(a) implies $\sum_s \sup_u |s| \; |c(u,s)| \; < \infty$, by

$$|\sum_t c(\frac{t}{T}, s) \psi_{jk}(t/T) \psi_{jk}((t{+}s)/T) - \sum_t c(\frac{t}{T}, s) \psi^2_{jk}(t/T)| \;\leq 2^j \sup_u |c(u,s)\; \psi(u)| \; |s| \; TV(\psi) \; ,$$

with $TV(\psi)$ denoting the total variation of ψ.
Hence we get the following rates:

$$\begin{aligned} \mathrm{var}\{\widehat{\beta}_{jk}\} \;\; &= \;\; T^{-2} \sum_{t,s} c_T(t/T, s) \;\; \psi_{jk}(t/T) \psi_{jk}((t+s)/T) \\ &= \;\; T^{-2} \sum_{t,s} c(t/T, s) \;\; \psi_{jk}(t/T) \psi_{jk}((t+s)/T) \;\; + \;\; O\;(2^j T^{-2}) \\ &= \;\; T^{-2} \sum_{t,s} c(t/T, s) \;\; \psi^2_{jk}(t/T) \;\; + \;\; O\;(2^j T^{-2}) \;\; + \;\; O\;(2^j T^{-2}) \end{aligned}$$

Further, using that for a function g of bounded variation,

$$T^{-1} \sum_t g(t/T) \;\; \psi^2_{jk}(t/T) \;\; - \;\; \int_0^1 g(u) \;\; \psi^2_{jk}(u) \; du \;\; = \;\; O\;(2^j T^{-1}) \; ,$$

we get

$$\begin{aligned} T^{-2} \sum_{t,s} c(t/T, s) \;\; \psi^2_{jk}(t/T) \;\; &= \;\; T^{-2} \sum_t f(t/T, 0) \;\; \psi^2_{jk}(t/T) \;\; + \;\; O\;(2^j T^{-2}) \\ &= \;\; T^{-1} \int_0^1 f(u, 0) \;\; \psi^2_{jk}(u) \; du \;\; + \;\; O\;(2^j T^{-2}) \;\; + \;\; O\;(2^j T^{-2}) \; . \end{aligned}$$

Here we have used that

$$\sum_{s \leq T} c(., s) = \sum_{s=-\infty}^{\infty} c(., s) \;\; + \;\; \sum_{s>T} c(., s) \;\; = \;\; f(., 0) + \sum_{s>T} c(., s) \; ,$$

where

$$\sum_{s>T} \sup_u |c(u,s)| \;\; \leq \sum_{s>T} \frac{|s|}{T} \;\; \sup_u |c(u,s)| = O\;(T^{-1}) \; ,$$

as, again, (A4)(a) implies $\sum_s \sup_u |s| \;\; |c(u,s)| \;\; < \;\; \infty$. Note also that, with (A3)(b), $\inf_u f(u,0) > 0$.

(iii) Using assumption (A2),

$$\begin{aligned}\mathrm{cum}_p(\widehat{\beta}_{jk}) &= \sum_{t_1}\sum_{t_2,\ldots,t_p}(T^{-1}\psi_{jk}(t_1/T))\ \ldots(T^{-1}\psi_{jk}(t_p/T))\ \ \mathrm{cum}(\varepsilon_{t_1,T},\ldots,\varepsilon_{t_p,T})\\ &= O\ \left(T^{-1}\,(T^{-1}2^{j/2})^{p-2}\ C^p(p!)^{1+\gamma}\right)\ ,\end{aligned}$$

uniformly in $p \geq 2$.

If $\sigma_{jk}^2 \geq C' \cdot T^{-1}$ for some positive C', we have, for $p \geq 3$ and for some appropriate $\widetilde{C}, \nu > 0$,

$$\mathrm{cum}_p(\widehat{\beta}_{jk}/\sigma_{jk})\ =\ O\ \left((T^{-1/2}2^{j/2})^{p-2}\ C^p(p!)^{1+\gamma}\right)\ \ =\ O\ \left((p!)^{1+\gamma}\ (\widetilde{C}T^{\nu})^{-(p-2)}\right)\ .$$

Proof of Theorem 4.1

Here we give the outline of the proof in our case following the general method used in Donoho and Johnstone (1994) and Johnstone and Silverman (1997). When necessary, we refer to several lemmas in the appendix of Johnstone and Silverman (1997).

A three point prior can be defined on $\mathbb{R}$ as follows:

As in Johnstone and Silverman (1997) let us choose $a \gg 0$ and define $\gamma_{T,a}$ (denoted μ_n in Donoho and Johnstone (1994) and Johnstone and Silverman (1997)) by

$$\phi(a+\gamma_{T,a})\ =\ \frac{\log T}{T}\ \phi(a).$$

Note that $\gamma_{T,a} \sim \sqrt{2\log T}$ as $T \to \infty$.

Let $F(\eta,\gamma)$ denote the three point prior that places mass $\frac{1}{2}\eta$ on each $\pm\gamma$ and mass $1-\eta$ on 0. The prior $\pi_{T,a}$ is then defined by setting the components θ_i to be independent $F[T^{-1}\ \log T, \gamma_{T,a}\tau_{ii,T}]$, where $\tau_{ii,T}^2 = 1/v^{ii,T}$ was defined in Section 4.1. Now let $\varrho_T(\pi_{T,a})$ and $\widetilde{\varrho}_T(\pi_{T,a})$ denote its Bayes risk and let θ^b and $\widetilde{\theta}^b$ denote the corresponding Bayes estimators with respect to the quadratic loss and the modified quadratic loss function defined by (31) respectively.

By the minimax theorem of decision theory, it suffices to show for the prior $\pi_{T,a}$ that as $T \to \infty$ then for all a and all η,

$$\frac{\overline{v_T^2}}{\overline{\tau_T^2}}\,\widetilde{\varrho}_T(\pi_{T,a})\ \geq\ (1+\eta)^{-1}\ \gamma_{T,a}^2\ \Phi(a)\ \{1+o(1)\}\ =\ 2(1+\eta)^{-1}\ \Phi(a)\ \log T\ \{1+o(1)\}. \qquad (*)$$

This is accomplished by considering A_T, the event $\{p_T(\theta) \leq 1 + (1+\eta)\log T\}$ where $p_T(\theta)$ is defined in (32) and proceeding to show that as $T \to \infty$, the probability $P(A_T) \to 1$.

We also need the technical result in Lemma 4 of the Appendix in [JS]:

$$E_{\pi_{T,a}}\,E_\theta\ (\|\widetilde{\theta}^b - \theta\|^2\, I[A_T^c])\ =\ o\ (\gamma_{T,a}^2\ \overline{v_T^2}\ \log T),$$

where $E_{\pi_{T,a}}$ denote expectation with respect to the prior $\pi_{T,a}$.

Now clearly,

$$E(p_T(\theta)) \le 1 + \log T$$

and

$$E|p_T(\theta) - E(p_T(\theta))|^2 \; = \; (\overline{v_T^2})^{-2}\, E\Big|\sum_{i=1}^{T}\{(\gamma_{T,a}^2\ \tau_i^2)\ \ \wedge v_{ii,T}^2\}\ (I_i - T^{-1}\log T)\ \Big|^2$$

where the I_i are independent Bernoulli $(T^{-1}\log T)$ random variables $I[\theta_i \neq 0]$. Thus clearly as in Lemma 3, [JS] there is a constant such that

$$\begin{aligned} E|p_T(\theta) - E(p_T(\theta))|^2 \quad &\le \quad C^*\ (\overline{v_T^2})^{-2} \sum_{i=1}^{T} v_{ii,T}^4\ E|I_i - T^{-1}\log T|^2 \\ &\le \quad C^*\ TC_1 2T^{-1}\log T \ = \ C^{**}\ \log T. \qquad (**) \end{aligned}$$

The second inequality follows from the lower bound on the eigenvalues of V_T.

Let $\{l_m, m = 1, \dots, T\}$ denote the eigenvalues of V_T.

$$\overline{v_T^2} \quad = \quad T^{-1}\ tr(V_T) = T^{-1}\sum_{m=1}^{T} l_m \ge C_0\ .$$

Furthermore, as the average of a convex function of the diagonal elements of V_T is bounded by the average of the function of the eigenvalues of V_T,

$$\overline{v_T^4} \ \ \le \ \ T^{-1}\ tr(V_T^2)\ \le\ C_1^2\ ,$$

as by assumption $\lambda_{max}(V_T) \le C_1$.

Now the fact that $P(A_T) \to 1$ as $T \to \infty$ can now be deduced from $(**)$ using Chebyshev's inequality.

Now the Bayes risk for $\pi_{T,a}$ with respect to the modified loss $\widetilde{L}_T$ (where $\widetilde{\theta}^b$ denotes the corresponding Bayes estimator) satisfies

$$\begin{aligned} \overline{v_T^2}\tilde{\varrho}_T(\pi_{T,a}) \quad &= \quad E_{\pi_{T,a}}\, E_\theta\ \frac{\|\widetilde{\theta}^b - \theta\|^2}{p_T(\theta)} \\ &\ge \quad \frac{E_{\pi_{T,a}}\, E_\theta\ (\|\widetilde{\theta}^b - \theta\|^2\ I[A_T])}{1 + (1+\eta)\log T} \\ &= \quad \frac{E_{\pi_{T,a}}\, E_\theta\ (\|\widetilde{\theta}^b - \theta\|^2)}{1 + (1+\eta)\log T}\ \ -\ o\ (\gamma_{T,a}^2\ \overline{v_T^2}), \qquad (***) \end{aligned}$$

applying Lemma 4 of Johnstone and Silverman (1997).

We observe that, as the arithmetic mean of the quantities $1/v^{ii,T}$ is greater than their harmonic mean, then

$$\overline{\tau_T^2} = T^{-1}\sum_{i=1}^{T}(v^{ii,T})^{-1} \ge (T^{-1}\sum_{i=1}^{T} v^{ii,T})^{-1} = (T^{-1}tr(V_T^{-1}))^{-1}\ .$$

The minimum eigenvalue condition implies that the $\lambda_{max}(V_T^{-1}) \leq C_0^{-1}$ for all T. Hence,

$$\overline{\tau_T^2} \geq (\lambda_{max}(V_T^{-1}))^{-1} \geq C_0 \ .$$

As,

$$\overline{v_T^2} = T^{-1} \ tr(V_T) \leq C_1 \ ,$$

we get, $\overline{\tau_T^2} \geq C_0/C_1 \ \overline{v_T^2}$.

This fact allows us to deduce the following:

$$\overline{v_T^2}\widetilde{\varrho}_T(\pi_{T,a}) \ = \ \frac{E_{\pi_{T,a}} E_\theta \ (\|\widetilde{\theta}^b - \theta\|^2)}{1 + (1+\eta)\log T} \ - \ o \ (\gamma_{T,a}^2 \ \overline{\tau_T^2}), \qquad (****)$$

Using equation (****) and Lemma 1 of Johnstone and Silverman (1997) we have that

$$\begin{aligned} \liminf_{T\to\infty} \ \frac{E_{\pi_{T,a}} E_\theta \ (\|\widetilde{\theta}^b - \theta\|^2)}{\overline{\tau_T^2}\gamma_{T,a}^2 \log T} \ &\geq \ \liminf_{T\to\infty} \ \frac{\varrho_T(\pi_{T,a})}{\overline{\overline{\tau_T^2}}\gamma_{T,a}^2 \log T} \\ &\geq \ \Phi(a) \ . \end{aligned}$$

Clearly this implies (*) and the theorem is proved.

Proof of Corollary 4.2

We need again to estimate $\overline{v_T^2}$ and $\overline{v_T^4}$ from above and $\overline{\tau_T^2}$ from below. For this it is sufficient to note that Γ_T, the covariance matrix of the error terms, fulfills $\Gamma_T = W'V_TW$ for the orthogonal matrix W of wavelet coefficient filter weights, and that

$$\overline{v_T^2} \ = \ T^{-1} \ tr(V_T) \ = \ T^{-1} \ tr(\Gamma_T) = T^{-1} \ \sum_{m=1}^{T} k_m \ \geq \ C_0$$

where $\{k_m \ : \ m = 1, \ldots, T\}$ denote the eigenvalues of Γ_T.

These equalities imply that $\lambda_{max}(\Gamma_T) = \lambda_{max}(V_T)$ and $\lambda_{min}(\Gamma_T) = \lambda_{min}(V_T)$. The last inequality follows from the assumption that

$$\lambda_{min}(\Gamma_T) \ \geq \ C_0 > 0 \qquad \forall \, T \geq 1 \ .$$

Furthermore, as the average of a convex function of the diagonal elements of V_T is bounded by the average of the function of the eigenvalues of V_T,

$$\overline{v_T^4} \ \leq \ T^{-1} \ tr(V_T^2) \ = \ T^{-1} \ tr(\Gamma_T^2) \ \leq \ \tilde{C}^2 \ ,$$

as

$$\lambda_{max}(\Gamma_T) \ \leq \ \sup_{1\leq t\leq T} \sum_s |\operatorname{cov}\{\varepsilon_{t,T}, \varepsilon_{s,T}\}| \ \leq \ \tilde{C},$$

by assumption (A2).

7 Acknowledgements

We would like to thank M.B. Brown and J. Raz, Department of Biostatistics, University of Michigan, Ann Arbor, for drawing our attention to the sheep luteinizing hormone data and for the use of these data, which have been collected by F. Karsch, Reproductive Sciences Program of the University of Michigan.

Many thanks also to B.W. Silverman, I.M. Johnstone, M.H. Neumann and J. Raz for helpful discussions.

Parts of the paper were written in a very hospitable atmosphere while the first author visited the second at GERAD in Montreal, in October 1996. The second author would like to thank the NSERC of Canada and FCAR of Quebec for partial support of this research.

References

Brillinger, D. R. (1994). Some asymptotics of wavelet fits in the stationary error case. Technical report 415, Dept. of Statist., U. C. Berkeley.

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

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

Dahlhaus, R., Neumann, M., and von Sachs, R. (1995). Non-linear wavelet estimation of time-varying autoregressive processes. Technical Report "Berichte der AG Technomathematik" 145, Universität Kaiserslautern (1995), tentatively accepted by *Bernoulli*.

Donoho, D. L. and Johnstone, I. M. (1992). Minimax estimation via wavelet shrinkage. Technical Report No. 402, Department of Statistics, Stanford University, to appear in *Ann. Statist.* 1997.

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

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200–1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion) *J. R. Statist. Soc., Ser. B* **57**, 301–369.

Donoho, D. L., Mallat, S., and von Sachs, R. (1996). Estimating covariances of locally stationary processes: Consistency of Best Basis methods. *Proc. IEEE TFTS-96, June 18-21, Paris, France.*

Diggle, P. and Zeger, S. (1989). A non-Gaussian model for time series with pulses. *J. Amer. Statist. Assoc.* **84**, 354–359.

Frazier, M., Jawerth, B., and Weiss, G. (1991). *Littlewood–Paley theory and the study of function spaces.* CBMS–Conference Lecture Notes **79**, American Mathematical Society, Providence, RI.

Johnstone, I. M., and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc., Ser. B* **59**, 319–351.

Johnstone, I. M. (1996). Wavelet shrinkage for correlated data and inverse problems: adaptivity results. Preprint, Stanford University.

Karsch, F. J., Robinson, J. E., Woodfill, C., and Brown, M. B. (1989). Circannual cycles of luteinizing hormone and prolactin secretion in ewes during prolonged exposure to a fixed photoperiod: Evidence for an endogenous reproductive rhythm. *Biology of Reproduction* **41**, 1034–1046.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *J. Amer. Statist. Assoc.* **82**, 1032–1041.

Mallat, S., Papanicolaou, G. and Zhang, Z. (1995). Adaptive covariance estimation of locally stationary processes. Manuscript.

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

Neumann, M. H., and von Sachs, R. (1995). Wavelet thresholding: Beyond the Gaussian i.i.d. situation. in: A. Antoniadis, G. Oppenheim (eds), ''Wavelets and Statistics", LN Statistics 103, Springer Verlag (1995), 301–329.

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

Normolle, D. P. and Brown, M. B. (1994). Identification of aperiodic seasonality in non–Gaussian time series. *Biometrics* **50**, 798–812.

O'Sullivan, F. and O'Sullivan, J. (1988). Deconvolution of episodic hormone data: An analysis of the role of season on the onset of puberty in cows. *Biometrics*, **44**, 339–353.

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

von Sachs, R., Nason, G. P., and Kroisandt, G. (1996) Spectral representation and estimation for locally stationary wavelet processes. *Proc. Workshop on Spline Functions and Theory of Wavelets. Montreal, Canada, 1996.*

von Sachs, R., and Schneider, K. (1996). Wavelet smoothing of evolutionary spectra by non-linear thresholding. *Applied Computational Harmonic Analysis.* **3**, 268–282.

Stein, C. (1981) Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–1151.

Wang, Y. (1996) Function estimation via wavelet shrinkage for long–memory data. *Ann. Statist.* **24**, 466–484.