

# Secondary Publication



Neumann, Michael H.; Leucht, Anne

## Trend estimation for time series with polynomial-tailed noise

Date of secondary publication: 21.05.2026

Submitted Version (Preprint), Article

Persistent identifier:

### Primary publication

Neumann, Michael H.; Leucht, Anne (2025): Trend estimation for time series with polynomial-tailed noise, in: arXiv, pp. 1–25, doi: 10.48550/arxiv.2502.08280

### Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.



This document is made available under a Creative Commons license.



The license information is available online:

<https://creativecommons.org/licenses/by/4.0/legalcode>

# Trend estimation for time series with polynomial-tailed noise

Michael H. Neumann<sup>1</sup>  Anne Leucht<sup>2</sup> 

<sup>1</sup>*Friedrich-Schiller-Universität Jena, Institut für Mathematik, Ernst-Abbe-Platz 2, D – 07743 Jena, Germany, e-mail: [E-mail: michael.neumann@uni-jena.de](mailto:michael.neumann@uni-jena.de)*

<sup>2</sup>*Universität Bamberg, Institut für Statistik, Feldkirchenstraße 21, D – 96052 Bamberg, Germany, e-mail: [E-mail: anne.leucht@uni-bamberg.de](mailto:anne.leucht@uni-bamberg.de)*

**Abstract:** For time series data observed at non-random and possibly non-equidistant time points, we estimate the trend function nonparametrically. Under the assumption of a bounded total variation of the function and low-order moment conditions on the errors we propose a nonlinear wavelet estimator which uses a Haar-type basis adapted to a possibly non-dyadic sample size. An appropriate thresholding scheme for sparse signals with an additive polynomial-tailed noise is first derived in an abstract framework and then applied to the problem of trend estimation.

**Keywords and phrases:** time series, trend estimation, wavelet thresholding.

## 1. Introduction

We consider the problem of estimating the trend function of a process observed at non-random, not necessarily equally spaced time points. We assume that this function has a bounded total variation which includes cases where the function is mostly smooth but has a few jumps. In such scenarios with an inhomogeneous smoothness a locally adapted degree of smoothing is required in order to obtain good rates of convergence. For example, Fourier series methods, kernel estimators with a global bandwidth and linear spline smoothers do not achieve this goal, see Figure 1.1 for illustration. In contrast, a wavelet expansion of such a function provides an efficient representation in terms of the corresponding coefficients: only a small fraction of coefficients are large in magnitude whereas the majority of them is small and therefore negligible. A common strategy consists of separating the large coefficients from the small ones by nonlinear thresholding, which means that the former are estimated and the latter are discarded. Such methods have become popular in the 1990s; see for example [Donoho et al. \(1995\)](#) and references therein. In the case of normally distributed errors and with a sample size  $n$ , a popular choice of the thresholds is  $\sqrt{2\log(n)}$  times the standard deviation of the empirical wavelet coefficients. [Donoho \(1995\)](#) showed that this choice provides a so-called “denoising”, that is, with high probability the estimator of the function is at least as smooth as the true function and its mean squared error is minimax-optimal in many function classes up to a logarithmic factor. Moreover, [Donoho and Johnstone \(1994\)](#) proved that such

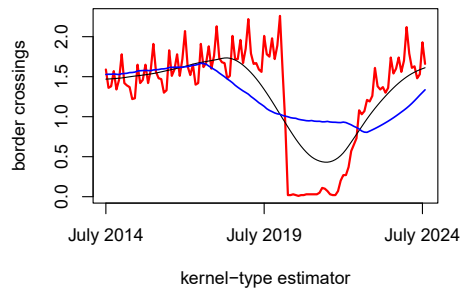


FIG 1.1. **Red**: monthly overseas arrivals in Australia (in millions) with structural breaks due to the COVID pandemic, **black/blue**: Nadaraya Watson estimator with **Epanechnikov** and **rectangular** kernel, bandwidth chosen by Scott's rule of thumb.

estimators have a mean squared error which differs from an ideal estimator by at most a logarithmic factor. Even in cases with non-Gaussian and dependent errors, thresholds as in the Gaussian case can be asymptotically appropriate since empirical wavelet coefficients at the important scales are asymptotically normal in a sufficiently strong sense, see e.g. [Neumann \(1996\)](#), [Neumann and von Sachs \(1997\)](#), and [Dahlhaus and Neumann \(2001\)](#). In the context of independent noise variables which have finite moments of a sufficiently large order, [Averkamp and Houdré \(2005\)](#) showed that soft thresholding achieves the same asymptotic performance as in the Gaussian case. Moreover, for independent variables with heavy tails, these authors proposed a pre-processing of the data by median filtering which compensated for missing finite moments, and finally led to the same rate of convergence as in the Gaussian case. Truly non-Gaussian thresholding rules were proposed by [Gao \(1993\)](#) in the context of spectral density estimation, by [Kolaczyk \(1999\)](#) for estimating the intensity function of a Poisson process. In these cases, the proposed thresholds are larger than those in the Gaussian case. Nonparametric estimation of a function with potentially inhomogeneous smoothness properties observed at non-equidistant data points has already been considered by a number of authors. For example, [Mammen and van de Geer \(1997\)](#) used least squares regression splines regularized by a total variation penalty. [Amato et al. \(2022\)](#) used wavelet thresholding under the assumption of sub-Gaussian noise.

In the present paper, we avoid the standard assumptions of a dyadic sample size, equally spaced sample points and i.i.d. Gaussian or sub-Gaussian noise. In Section 2 we introduce our model and argue that the wavelet coefficients at fine scales are sparse. More importantly, we derive our results under low-order moment conditions which are often imposed in time series analysis. We only require finite moments up to order four and impose a standard condition on the joint cumulants of the errors. This results in a relatively slow polynomial decay of the tails of the distribution of the empirical wavelet coefficients and requires an appropriate adjustment of the thresholds. Section 3 is devoted to a thorough discussion of estimating sparse signals blurred by polynomial-tailed

noise. It is shown in abstract Bayesian and minimax contexts how an optimal rate of convergence depends both on the level of noise as well as the degree of sparsity and how this rate can be attained by an appropriate choice of a threshold. As a further prerequisite for our main results, we discuss the construction of a Haar-type basis for a possibly non-dyadic sample size  $n$  in Section 4.1. We use appropriately adapted basis functions which however share the essential properties of the Haar basis. This deviates from the approach in Amato *et al.* (2022) who intended to apply the standard discrete wavelet transform and therefore proposed to embed the data points  $x_1, \dots, x_n$  into a fine equispaced grid  $\{1/N, 2/N, \dots, 1\}$ , where  $N = 2^J$  for some  $J \in \mathbb{N}$  and  $N \gg n$ . The regularization scheme developed in Section 3 is applied in Section 4.2 to our particular problem of the estimation of a possibly discontinuous trend function of a time series. In Section 4.3 we extend our results to a partially linear model which was also considered in Amato *et al.* (2021). While these authors proposed to estimate the parameters of the linear part and the wavelet coefficients simultaneously, we use a simpler approach where the linear part is first fitted by least squares and wavelet thresholding is applied to the empirical wavelet coefficients afterwards. Section 5 contains some simulations and a real data example. It is shown that the proposed nonlinear wavelet estimator clearly outperforms kernel estimators with an optimally chosen global bandwidth. Proofs of our main results are postponed to Section 6 and a few auxiliary results are collected in a final Section 7.

## 2. Assumptions and a preview of our main results

Suppose that we observe  $Y_1, \dots, Y_n$  which form a not necessarily stationary time series and that  $EY_t = m_0(x_t)$ , where  $x_1 < x_2 < \dots < x_n$  are ordinal variables representing e.g. time. This leads to the model

$$Y_t = m_0(x_t) + \varepsilon_t, \quad t = 1, \dots, n. \quad (2.1)$$

We do not assume any parametric model for the mean function  $m_0$ ; instead we assume that its total variation is small in relation to the sample size  $n$ . For a real-valued function  $f$  defined on a set  $\mathcal{X} \subseteq \mathbb{R}$ , its total variation on a subset  $\tilde{\mathcal{X}}$  is defined by  $\text{TV}(f; \tilde{\mathcal{X}}) = \sup \left\{ \sum_{i=1}^N |f(x_i) - f(x_{i-1})| : \{x_1, \dots, x_N\} \subseteq \tilde{\mathcal{X}}, x_1 < x_2 < \dots < x_N, N \in \mathbb{N} \right\}$ . We assume

$$\text{(A1)} \quad \text{TV}(m_0; \{x_1, \dots, x_n\}) \leq C_0.$$

Regarding the errors  $\varepsilon_1, \dots, \varepsilon_n$  we impose the following weak conditions that are standard in time series analysis.

$$\begin{aligned} \text{(A2)} \quad & \text{(i)} \quad E\varepsilon_t = 0, \\ & \text{(ii)} \quad \sup_s \sum_t \text{cov}(\varepsilon_s, \varepsilon_t) \leq C_1, \\ & \text{(iii)} \quad \sup_s \sum_{t,u,v} \text{cum}(\varepsilon_s, \varepsilon_t, \varepsilon_u, \varepsilon_v) \leq C_2, \end{aligned}$$

Here  $C_0, C_1, C_2$  are fixed finite constants and  $\text{cum}(\varepsilon_s, \varepsilon_t, \varepsilon_u, \varepsilon_v) = E[\varepsilon_s \varepsilon_t \varepsilon_u \varepsilon_v] - E[\varepsilon_s \varepsilon_t] E[\varepsilon_u \varepsilon_v] - E[\varepsilon_s \varepsilon_u] E[\varepsilon_t \varepsilon_v] - E[\varepsilon_s \varepsilon_v] E[\varepsilon_t \varepsilon_u]$  denotes the joint cumulant of  $\varepsilon_s, \varepsilon_t, \varepsilon_u, \varepsilon_v$ .

We propose an estimator  $\widehat{m}_n$  of  $m_0$  which will be based on a wavelet expansion of this function. Its performance is measured in terms of the mean squared error at the sample points, i.e.  $E[(1/n) \sum_{t=1}^n (\widehat{m}_n(x_t) - m_0(x_t))^2]$ . Under our assumption **(A1)** a variant of the Haar basis is an appropriate simple choice. We present in Section 4 a version of this basis which is adapted to possibly unevenly spaced design points and a non-dyadic sample size. Then the function  $m_0$  can be represented as

$$m_0(x_t) = \alpha_0^0 + \sum_{j=0}^{J_n} \sum_{k: (j,k) \in \mathcal{I}_n} \beta_{j,k}^0 \psi_{j,k}(x_t),$$

where  $\alpha_0^0 = (1/n) \sum_{t=1}^n m_0(x_t)$ ,  $\beta_{j,k}^0 = (1/n) \sum_{t=1}^n m_0(x_t) \psi_{j,k}(x_t)$ ,  $J_n$  such that  $2^{J_n} < n \leq 2^{J_n+1}$ , and  $\mathcal{I}_n \subseteq \bigcup_{j=0}^{J_n} (\{j\} \times \{1, \dots, 2^j\})$ ; see Section 4 for details. Our wavelet estimator has the form

$$\widehat{m}_n(x_t) = \widehat{\alpha}_0 + \sum_{j=0}^{J_n} \sum_{k: (j,k) \in \mathcal{I}_n} \widehat{\beta}_{j,k} \psi_{j,k}(x_t),$$

where  $\widehat{\alpha}_0$  and the  $\widehat{\beta}_{j,k}$  are estimators of the corresponding coefficients. Since the basis functions form an orthonormal system w.r.t. the inner product  $\langle f, g \rangle = (1/n) \sum_{t=1}^n f(x_t)g(x_t)$ , we have the isometry

$$\frac{1}{n} \sum_{t=1}^n (\widehat{m}_n(x_t) - m_0(x_t))^2 = \sum_{j=0}^{J_n} \sum_{k: (j,k) \in \mathcal{I}_n} (\widehat{\beta}_{j,k} - \beta_{j,k}^0)^2.$$

This separation into the contribution of single coefficients makes an analytic study of the asymptotic performance of  $\widehat{m}_n$  possible.

Under assumption **(A1)** we obtain that

$$2^{-j} \sum_{k: (j,k) \in \mathcal{I}_n} \beta_{j,k}^0 \leq C_0 2^{-3j/2}; \quad (2.2)$$

see (7.1). Empirical versions of the wavelet coefficients can be obtained from the sample as  $\widetilde{\alpha}_0 = (1/n) \sum_{t=1}^n Y_t$  and  $\widetilde{\beta}_{j,k} = (1/n) \sum_{t=1}^n m_0(x_t) Y_t$ . Under assumption **(A2)** we obtain by Lemma 7.2 that

$$P(|\widetilde{\beta}_{j,k} - \beta_{j,k}^0| > t) \leq \frac{E[(\widetilde{\beta}_{j,k} - \beta_{j,k}^0)^4]}{t^4} \leq C (n^{-1/2}/t)^4 \quad \forall t > 0, (j, k) \in \mathcal{I}_n, \quad (2.3)$$

for some  $C < \infty$ . At fine scales  $j$ ,  $2^{-3j/2}$  gets smaller than the noise level  $n^{-1/2}$ , and the degree of sparsity may be described by the ratio  $q_{n,j} = n^{-3j/2}/n^{-1/2}$ . In

the next section we study an abstract model which mimics the situation we are faced with when we estimate the wavelet coefficients at fine scales. This suggests how the empirical wavelet coefficients can be regularized such that the resulting estimator attains an optimal rate of convergence. In Section 4 we introduce a variant of the Haar basis which is adapted to possibly unevenly spaced design points and a non-dyadic sample size. Then we apply the regularization scheme derived in Section 3 and obtain our wavelet estimator of  $m_0$ .

### 3. Optimal reconstruction of sparse signals from data with polynomial-tailed noise

In this section we consider an abstract model which mimics the situation we are faced with when we estimate the wavelet coefficients at fine scales; see in particular (2.2) and (2.3). Suppose first that we observe real-valued random variables  $Y_1, \dots, Y_N$  such that

$$Y_k = \theta_k + \varepsilon_k, \quad k = 1, \dots, N, \quad (3.1a)$$

where

$$\varepsilon_k \sim Q_\varepsilon \in \mathbf{Q}_\varepsilon := \{Q: 1 - Q([-t, t]) \leq (\varepsilon/t)^4 \quad \forall t > 0\} \quad (3.1b)$$

and

$$\theta = (\theta_1, \dots, \theta_N)^T \in \Theta_{N,q} := \left\{ \theta \in \mathbb{R}^N: \frac{1}{N} \sum_{k=1}^N \theta_k \leq \varepsilon q \right\}, \quad (3.1c)$$

for some  $q \in (0, 1)$ . Note that  $\varepsilon_k$  does not have a finite fourth moment in general and it is also not required that  $E\varepsilon_k = 0$ . This includes the case of deterministic noise if  $|\varepsilon_k| \leq \varepsilon$ . In any case, it follows from Lemma 7.1 with  $t = 0$  that

$$E[\varepsilon_k^2] = 2 \int_0^\infty (1 - Q_\varepsilon([-x, x])) x dx \leq 2 \int_0^\varepsilon x dx + 2 \int_\varepsilon^\infty \varepsilon^4 x^{-3} dx = 2\varepsilon^2, \quad (3.2)$$

that is,  $\varepsilon$  may be interpreted as noise level. The constant  $q$  in (3.1c) describes the degree of sparsity of the signals  $\theta_1, \dots, \theta_N$  in relation to  $\varepsilon$ . To obtain a guideline for an appropriate regularization we consider first corresponding Bayes and minimax problems.

To establish a Bayesian framework, we suppose that  $\theta_1, \dots, \theta_N$  are independent and follow a three-point prior, i.e.,

$$\theta_k \sim \pi; \quad \pi(\{x\}) = \begin{cases} p/2, & \text{if } x \in \{-\lambda, \lambda\}, \\ 1-p, & \text{if } x = 0 \end{cases}, \quad (3.3a)$$

where  $p = q^{4/3}$  and  $\lambda = \varepsilon q^{-1/3}$ . Suppose further that the errors  $\varepsilon_1, \dots, \varepsilon_N$  are independent, also independent of the signals  $\theta_1, \dots, \theta_N$ , and

$$\varepsilon_k \sim \pi. \quad (3.3b)$$

Then

$$E\left[\frac{1}{N}\sum_{k=1}^N|\theta_k|\right] = p\lambda = \epsilon q,$$

that is, (3.1c) is satisfied on average. Furthermore,  $P(|\varepsilon_k| > t) \leq (\epsilon/t)^4 \forall t > 0$ , that is,  $\varepsilon_k \sim Q_\epsilon \in \mathbf{Q}_\epsilon$ . The following results shows how the Bayes risk depends on the degree of sparsity.

**Proposition 3.1.** *Suppose that (3.1a), (3.3a), and (3.3b) are fulfilled.*

*Then the unique Bayes estimator  $T^*(Y_k)$  is given by  $T^*(-2\lambda) = -\lambda$ ,  $T^*(-\lambda) = -\lambda/2$ ,  $T^*(0) = 0$ ,  $T^*(\lambda) = \lambda/2$ ,  $T^*(2\lambda) = \lambda$ , and its Bayes risk is equal to*

$$E[(T^*(Y_k) - \theta_k)^2] = \epsilon^2 q^{2/3}/2.$$

This result can be used to obtain a lower bound to a related minimax risk. Suppose in addition that  $q^{-4/3} = O(N^{1-\gamma})$  for some  $\gamma \in (0, 1)$ . Let  $\delta > 0$ , and let  $\pi^{(N)}$  be the  $N$ -fold product of  $\pi$ . Let  $\tilde{T} = \tilde{T}(Y_1, \dots, Y_N) = (\tilde{T}_1, \dots, \tilde{T}_n)^T$  be the Bayes estimator of the vector  $\theta$  w.r.t. the prior given by the truncation of  $\pi^{(N)}$  to  $\Theta_{N,q(1+\delta)}$ . Then a lower bound to the minimax risk over  $\Theta_{N,q(1+\delta)}$  is given by a corresponding Bayes risk,

$$\begin{aligned} & \int_{\Theta_{N,q(1+\delta)}} E_\theta \left[ \frac{1}{N} \sum_{k=1}^N (\tilde{T}_k - \theta_k)^2 \right] d\pi^{(N)}(\theta) \\ &= \int_{\mathbb{R}^N} E_\theta \left[ \frac{1}{N} \sum_{k=1}^N (\tilde{T}_k - \theta_k)^2 \right] d\pi^{(N)}(\theta) - \int_{\mathbb{R}^N \setminus \Theta_{N,q(1+\delta)}} E_\theta \left[ \frac{1}{N} \sum_{k=1}^N (\tilde{T}_k - \theta_k)^2 \right] d\pi^{(N)}(\theta). \end{aligned}$$

The first term on the right-hand side can be estimated from below by  $\int_{\mathbb{R}^N} E_\theta \left[ (1/N) \sum_{k=1}^N (T^*(Y_k) - \theta_k)^2 \right] d\pi^{(N)}(\theta) = \epsilon^2 q^{2/3}/2$ . Regarding the second one, note first that  $\tilde{T}_k \in [-\lambda, \lambda]$  holds with probability 1, which leads to

$$E_\theta \left[ \frac{1}{N} \sum_{k=1}^N (\tilde{T}_k - \theta_k)^2 \right] \leq (2\lambda)^2 = 4\epsilon^2 q^{-2/3}.$$

On the other hand, we obtain from Bernstein's inequality that

$$\begin{aligned} P(\theta \notin \Theta_{N,q(1+\delta)}) &= P\left(\frac{1}{N}\sum_{k=1}^N|\theta_k| - E|\theta_k| > \epsilon q(1+\delta)\right) \\ &\leq \exp\left\{-\frac{(N\epsilon q(1+\delta))^2/2}{\sum_{k=1}^N E[(|\theta_k| - E|\theta_k|)^2] + |\theta_1| - E|\theta_1| \infty (N\epsilon q(1+\delta))/3}\right\} \\ &= \exp\{-R_N\}, \end{aligned}$$

where

$$1/R_N \asymp \frac{N\lambda^2 p + \lambda N\epsilon q}{(N\epsilon q)^2} = O(N^{-1}q^{-4/3}) = O(N^{-\gamma}).$$

This implies that the second term is of order  $O(\epsilon^2 q^{-2/3} \exp\{-R_N\})$  and we obtain that

$$\begin{aligned} & \inf_{\hat{T}} \sup \left\{ E_{\theta} \left[ \frac{1}{N} \sum_{k=1}^N (\hat{T}_k - \theta)^2 \right] : \theta \in \Theta_{N, q(1+\delta)}, \varepsilon_1, \dots, \varepsilon_N \sim Q_{\varepsilon} \in \mathbf{Q}_{\varepsilon} \right\} \\ & \geq (\epsilon^2 q^{2/3}/2)(1 + o(1)), \end{aligned} \quad (3.4)$$

as  $N \rightarrow \infty$ .

Although the form of  $T^*$  might suggest that the linear estimator  $Y_k/2$  of  $\theta_k$  is appropriate, this is no longer true for some other error distributions which satisfy (3.1b). If, for example,  $\varepsilon_k \sim \mathcal{N}(0, \epsilon^2)$ , then it follows from  $E[(Y_k/2 - \theta_k)^2] = \epsilon^2/4 + (\theta_k/2)^2$  that the Bayes risk of  $Y_k/2$  under  $\pi$  is equal to  $\epsilon^2/4 + p\lambda^2/4 = \epsilon^2(1 + q^{2/3})/4$ , that is, it is dominated by the variance of  $Y_k$  and there is not the desired gain due to the sparsity of the signal  $\theta_k$ . This is of course in line with common folklore that nonlinear methods are required in order to efficiently estimate sparse signals.

We therefore consider estimators  $\hat{\theta}_k = \hat{T}(Y_k)$  which satisfy the following conditions:

$$\hat{T}(y) = 0, \quad \text{if } |y| < t \quad (3.5a)$$

and

$$\hat{T}(y) - y \leq t, \quad \text{if } |y| \geq t. \quad (3.5b)$$

Note that the popular methods of soft and hard thresholding with threshold  $t$  satisfy these conditions. The following theorem shows that the above estimator attains the optimal rate of convergence.

**Theorem 3.1.** *Suppose that (3.1a), (3.1b), (3.5a), and (3.5b) are fulfilled. Then*

- (i)  $E[(\hat{\theta}_k - \theta_k)^2] \leq \begin{cases} \theta_k^2 + 9E[\varepsilon_k^2 \mathbb{1}(|\varepsilon_k| > t/2)] & \text{if } |\theta_k| \leq t/2, \\ 2t^2 + 4\epsilon^2 & \text{if } |\theta_k| > t/2 \end{cases}.$
- (ii) *If in addition  $t = K\lambda = K\epsilon q^{-1/3}$  for some  $K > 0$ , then*

$$\sup_{\theta \in \Theta_{N, q(1+\delta)}, \varepsilon_1, \dots, \varepsilon_N \sim Q_{\varepsilon} \in \mathbf{Q}_{\varepsilon}} E_{\theta} \left[ \frac{1}{N} \sum_{k=1}^N (\hat{\theta}_k - \theta_k)^2 \right] = O(\epsilon^2 q^{2/3}).$$

## 4. Main results

### 4.1. A Haar-type basis for unevenly spaced data

Let us first consider the simple case where we have to deal with a real-valued function  $f$  on the unit interval  $(0,1]$ . This interval can be decomposed into intervals

$$I_{j,k} = ((k-1)2^{-j}, k2^{-j}], \quad j = 0, 1, 2, \dots; \quad k = 1, \dots, 2^j.$$

To define the Haar basis we start with a scaling function  $\phi_0 = \mathbb{1}_{I_{0,1}} = \mathbb{1}_{(0,1]}$  and a mother wavelet  $\psi = \mathbb{1}_{I_{1,1}} - \mathbb{1}_{I_{1,2}} = \mathbb{1}_{(0,1/2]} - \mathbb{1}_{(1/2,1]}$ . Using dilations and translations we obtain wavelets  $\psi_{j,k} = 2^{j/2} (\mathbb{1}_{I_{j+1,2k-1}} - \mathbb{1}_{I_{j+1,2k}}) = 2^{j/2} \psi(2^j \cdot - (k-1))$ , where  $j \geq 0$ ;  $k = 1, \dots, 2^j$ . The collection of these functions  $\{\phi_0, \psi_{j,k} (j \geq 0; k = 1, \dots, 2^j)\}$  forms an orthonormal basis of  $L_2((0, 1])$ . An arbitrary function  $f \in L_2((0, 1])$  can be expanded as

$$f(x) = \alpha_0 \phi_0(x) + \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k} \psi_{j,k}(x),$$

where  $\alpha_0 = \int \phi_0(x) f(x) dx$  and  $\beta_{j,k} = \int \psi_{j,k}(x) f(x) dx$ . Since the wavelets  $\psi_{j,1}, \dots, \psi_{j,2^j}$  are supported on respective disjoint intervals  $I_{j,1}, \dots, I_{j,2^j}$ , we obtain the following estimate for the size of the wavelet coefficients:

$$\sum_{k=1}^{2^j} \beta_{j,k} = \sum_{k=1}^{2^j} \|\psi_{j,k}\|_1 \cdot (\text{TV}(f; I_{j,k})/2) \leq 2^{-j/2-1} \text{TV}(f; (0, 1]). \quad (4.1)$$

Suppose for the time being that we have a dyadic sample size  $n = 2^{J_n+1}$  for some  $J_n \in \mathbb{N}$  and that  $x_t = t/n$  for all  $t = 1, \dots, n$ . Then we can directly use the Haar basis for our purposes. Let  $\bar{m}_0$  be a piecewise constant continuation of  $m_0(x_1), \dots, m_0(x_n)$  on the intervals  $(x_{t-1}, x_t]$ , i.e.,  $\bar{m}_0(x) = m_0(x_t) \forall x \in (x_{t-1}, x_t]$ . Then  $\bar{m}_0$  can be perfectly represented by the first  $n$  basis functions,

$$\bar{m}_0(x) = \alpha_0^0 \phi_0(x) + \sum_{j=0}^{J_n} \sum_{k=1}^{2^j} \beta_{j,k}^0 \psi_{j,k}(x),$$

where  $\alpha_0^0 = \int \phi_0(x) \bar{m}_0(x) dx = (1/n) \sum_{t=1}^n m_0(x_t)$  and  $\beta_{j,k}^0 = \int \psi_{j,k}(x) \bar{m}_0(x) dx = (1/n) \sum_{t=1}^n \psi_{j,k}(x_t) m_0(x_t)$ . We have in particular that

$$m_0(x_t) = \alpha_0^0 + \sum_{j=0}^{J_n} \sum_{k=1}^{2^j} \beta_{j,k}^0 \psi_{j,k}(x_t) \quad \forall t = 1, \dots, n.$$

The observations  $Y_1, \dots, Y_n$  can also be described this way,

$$Y_t = \tilde{\alpha}_0 \phi_0(x_t) + \sum_{j=0}^{J_n} \sum_{k=1}^{2^j} \tilde{\beta}_{j,k} \psi_{j,k}(x_t),$$

where  $\tilde{\alpha}_0 = (1/n) \sum_{t=1}^n \phi_0(x_t) Y_t = \bar{Y}_n$  and  $\tilde{\beta}_{j,k} = (1/n) \sum_{t=1}^n \psi_{j,k}(x_t) Y_t$  are empirical versions of the respective wavelet coefficients. Since the vectors  $(\phi_0(x_1), \dots, \phi_0(x_n))^T$  and  $(\psi_{j,k}(x_1), \dots, \psi_{j,k}(x_n))^T$  ( $j = 0, \dots, J_n$ ;  $k = 1, \dots, 2^j$ ) form an orthonormal basis of  $\mathbb{R}^n$  w.r.t. the inner product  $\langle \cdot, \cdot \rangle$  given by  $\langle a, b \rangle = (1/n) \sum_{t=1}^n a_t b_t$ , we also obtain, for  $\hat{m}_n(x) = \hat{\alpha}_0 + \sum_{j=0}^{J_n} \sum_{k=1}^{2^j} \hat{\beta}_{j,k} \psi_{j,k}(x)$ , that

$$\frac{1}{n} \sum_{t=1}^n (\hat{m}_n(x) - m_0(x_t))^2 = (\hat{\alpha}_0 - \alpha_0^0)^2 + \sum_{j=0}^{J_n} \sum_{k=1}^{2^j} (\hat{\beta}_{j,k} - \beta_{j,k}^0)^2.$$

If the sample size  $n$  is not a dyadic one, then we have to modify the above approach slightly. Since the  $x_t$  can be considered as ordinal variables, we simplify our notation by sticking to our assumption that  $x_t = t/n$ . We choose the finest scale  $J_n$  such that

$$2^{J_n} < n \leq 2^{J_n+1}.$$

As above, we take as a starting point dyadic intervals on  $(0, 1]$ :

$$I_{j,k} = ((k-1)2^{-j}, k2^{-j}], \quad j = 0, \dots, J_n + 1; k = 1, \dots, 2^j.$$

It suffices to specify the functions  $\phi_0$  and  $\psi_{j,k}$  at the points  $x_1, \dots, x_n$ . Let

$$\phi_0(x_t) = 1 \quad \forall t = 1, \dots, n.$$

We define wavelet functions  $\psi_{j,k}$  such that the essential properties of the Haar basis are retained: These functions shall form an orthonormal system w.r.t. the inner product  $\langle \cdot, \cdot \rangle$  and they should efficiently describe functions with jumps. Let

$$n_{j,k} := \#\{1 \leq t \leq n: x_t \in I_{j,k} \}.$$

It is not difficult to see that

$$[n2^{-j}] \leq n_{j,k} < n2^{-j} + 1.$$

(Since the length of the interval  $I_{j,k}$  is  $2^{-j}$  and the distance between adjacent points  $x_{t-1}$  and  $x_t$  is  $1/n$ , we obtain that  $n_{j,k} = n2^{-j}$  if  $n2^{-j}$  is an integer. Otherwise, if  $n2^{-j}$  is not an integer, then  $n_{j,k} \geq [n2^{-j}]$ . On the other hand,  $n_{j,k} \geq n2^{-j} + 1$  is impossible since this implies  $n_{j,k} \geq [n2^{-j}] + 2$  and so the length of  $I_{j,k}$  would exceed  $([n2^{-j}] + 1)/n$ . This, however, leads to a contradiction since  $([n2^{-j}] + 1)/n > n2^{-j}/n = 2^{-j}$ .) First we obtain a system of orthogonal functions by

$$\tilde{\psi}_{j,k} = \frac{1}{n_{j+1,2k-1}} \mathbb{1}_{I_{j+1,2k-1}} - \frac{1}{n_{j+1,2k}} \mathbb{1}_{I_{j+1,2k}} \quad \forall (j, k) \in \mathcal{I}_n,$$

where  $\mathcal{I}_n := \{(j, k): n_{j+1,2k-1} \geq 1 \text{ and } n_{j+1,2k} \geq 1\}$ . Since  $n_{J_n+1,k} < n2^{-(J_n+1)} + 1 \leq 2$ , we obtain  $n_{J_n+1,k} \leq 1$  for all  $k = 1, \dots, 2^{J_n+1}$ . This implies that  $(j, k) \notin \mathcal{I}_n$ , if  $j > J_n$ , i.e.,  $J_n$  is the finest scale where functions  $\tilde{\psi}_{j,k}$  are defined. Moreover, the ‘decomposition pyramid’ does not stop before the interval  $(0, 1]$  is decomposed into the smallest possible intervals  $((t-1)/n, t/n]$ . For example, if  $n_{j,k} = 2$ , then  $n_{j+1,2k-1} = n_{j+1,2k} = 1$ , and so  $(j, k) \in \mathcal{I}_n$ . This implies in particular that  $\#\mathcal{I}_n = n - 1$ .

We have

$$\langle \phi_0, \tilde{\psi}_{j,k} \rangle := \frac{1}{n} \sum_{t=1}^n \phi_0(x_t) \tilde{\psi}_{j,k}(x_t) = 0 \quad \forall (j, k) \in \mathcal{I}_n$$

and

$$\langle \tilde{\psi}_{j,k}, \tilde{\psi}_{j',k'} \rangle := \frac{1}{n} \sum_{t=1}^n \tilde{\psi}_{j,k}(x_t) \tilde{\psi}_{j',k'}(x_t) = 0 \quad \forall (j, k), (j', k') \in \mathcal{I}_n, (j, k) \neq (j', k'),$$

i.e., these functions form an orthogonal system w.r.t. the inner product  $\langle \cdot, \cdot \rangle$ . It remains to normalize these functions. Since

$$\langle \tilde{\psi}_{j,k}, \tilde{\psi}_{j,k} \rangle = \frac{1}{n} \left( \frac{1}{n_{j+1,2k-1}} + \frac{1}{n_{j+1,2k}} \right),$$

we obtain by

$$\psi_{j,k} := \tilde{\psi}_{j,k} / \sqrt{\langle \tilde{\psi}_{j,k}, \tilde{\psi}_{j,k} \rangle} = \frac{\sqrt{n}}{\sqrt{\frac{1}{n_{j+1,2k-1}} + \frac{1}{n_{j+1,2k}}}} \left( \frac{1}{n_{j+1,2k-1}} \mathbb{1}_{I_{j+1,2k-1}} - \frac{1}{n_{j+1,2k}} \mathbb{1}_{I_{j+1,2k}} \right)$$

that  $\langle \psi_{j,k}, \psi_{j,k} \rangle = 1$ . The vectors

$$\phi_0 = (1/\sqrt{n})(\phi_0(x_1), \dots, \phi_0(x_n))^T$$

and

$$\psi_{j,k} = (1/\sqrt{n})(\psi_{j,k}(x_1), \dots, \psi_{j,k}(x_n))^T \quad \forall (j,k) \in \mathcal{I}_n$$

form an orthonormal system in  $\mathbb{R}^n$ . Since  $\#\mathcal{I}_n = n - 1$ , this is even an orthonormal basis. As in the case of  $n = 2^{J_n+1}$  we have, for  $\hat{m}_n(x) = \hat{\alpha}_0 + \sum_{j=0}^{J_n} \sum_{k:(j,k) \in \mathcal{I}_n} \hat{\beta}_{j,k} \psi_{j,k}(x)$ , the isometry

$$\frac{1}{n} \sum_{t=1}^n (\hat{m}_n(x_t) - m_0(x_t))^2 = (\hat{\alpha}_0 - \alpha_0^0)^2 + \sum_{j=0}^{J_n} \sum_{k:(j,k) \in \mathcal{I}_n} (\hat{\beta}_{j,k} - \beta_{j,k}^0)^2. \quad (4.2)$$

#### 4.2. A nonlinear wavelet estimator of the trend function

Now we consider an estimator  $\hat{m}_n$  of  $m_0$ , where

$$\hat{m}_n(x_t) = \hat{\alpha}_0 + \sum_{j=0}^{J_n} \sum_{k:(j,k) \in \mathcal{I}_n} \hat{\beta}_{j,k} \psi_{j,k}(x_t).$$

The coefficients  $\hat{\alpha}_0$  and  $\hat{\beta}_{j,k}$  of this wavelet expansion are derived from corresponding empirical versions  $\tilde{\alpha}_0 = (1/n) \sum_{t=1}^n \phi_0(x_t) Y_t$  and  $\tilde{\beta}_{j,k} = (1/n) \sum_{t=1}^n \psi_{j,k}(x_t) Y_t$  of the true coefficients  $\alpha_0^0 = (1/n) \sum_{t=1}^n \phi_0(x_t) m_0(x_t)$  and  $\beta_{j,k}^0 = (1/n) \sum_{t=1}^n \psi_{j,k}(x_t) m_0(x_t)$  of the function  $m_0$ . In view of the isometry (4.2) above, we direct our attention to the estimation of the coefficients. It follows from Lemma 7.2 that

$$E[(\tilde{\beta}_{j,k} - \beta_{j,k}^0)^4] \leq C_3 n^{-2} \quad \forall (j,k) \in \mathcal{I}_n, \quad (4.3)$$

for some  $C_3 < \infty$ , which implies that

$$P(|\tilde{\beta}_{j,k} - \beta_{j,k}^0| > t) \leq C_3 (n^{-1/2}/t)^4 \quad \forall t > 0. \quad (4.4)$$

On the other hand, we obtain similarly to (4.1) that

$$2^{-j} \sum_{k:(j,k) \in \mathcal{I}_n} |\beta_{j,k}^0| \leq C_0 2^{-3j/2}. \quad (4.5)$$

This means that the signal becomes sparse in relation to the noise level  $n^{-1/2}$  at scales  $j$  where  $2^{-3j/2}$  is of smaller order than  $n^{-1/2}$ . The degree of sparsity is expressed by  $q_{n,j} = n^{1/2}2^{-3j/2}$ . In view of the message provided by Theorem 3.1 we will modify the empirical coefficients  $\tilde{\beta}_{j,k}$  nonlinearly at scales  $j \geq J_n^*$ , where  $J_n^*$  is the critical level. Let, for definiteness,  $J_n^*$  be such that  $2^{J_n^*-1} < n^{1/3} \leq 2^{J_n^*}$  and let  $t_{n,j} = K n^{-1/2} q_{n,j}^{-1/3} = K n^{-2/3} 2^{j/2}$ , where  $K$  is an arbitrary positive constant. We focus on the estimator

$$\hat{m}_n(x_t) = \tilde{\alpha}_0 + \sum_{j=0}^{J_n^*-1} \sum_{k: (j,k) \in \mathcal{I}_n} \tilde{\beta}_{j,k} \psi_{j,k}(x_t) + \sum_{j=J_n^*}^{J_n} \sum_{k: (j,k) \in \mathcal{I}_n} \hat{\beta}_{j,k} \psi_{j,k}(x_t),$$

where, for  $(j, k) \in \mathcal{I}_n$ ,  $j \geq J_n^*$ ,

$$\hat{\beta}_{j,k} = 0, \quad \text{if } |\tilde{\beta}_{j,k}| < t_{n,j} \quad (4.6a)$$

and

$$\hat{\beta}_{j,k} - \tilde{\beta}_{j,k} \leq t_{n,j}, \quad \text{if } |\tilde{\beta}_{j,k}| \geq t_{n,j}. \quad (4.6b)$$

Popular examples of such a strategy are hard thresholding,

$$\hat{\beta}_{j,k}^{(h)} = \begin{cases} 0, & \text{if } |\tilde{\beta}_{j,k}| < t_{n,j}, \\ \tilde{\beta}_{j,k}, & \text{if } |\tilde{\beta}_{j,k}| \geq t_{n,j} \end{cases},$$

and soft thresholding,

$$\hat{\beta}_{j,k}^{(s)} = \text{sgn}(\tilde{\beta}_{j,k}) (|\tilde{\beta}_{j,k}| - t_{n,j})^+.$$

Note, in passing, that there is a well-known connection between soft thresholding and  $l_1$ -penalization, i.e.,  $\hat{\beta}_{j,k}^{(s)}$  is the unique minimizer of the function  $\beta \mapsto (\beta - \tilde{\beta}_{j,k})^2 + 2t_{n,j}|\beta|$ . It follows from (4.3) that

$$E[(\tilde{\alpha}_0 - \alpha_0^0)^2] + \sum_{j=0}^{J_n^*-1} \sum_{k: (j,k) \in \mathcal{I}_n} E[(\tilde{\beta}_{j,k} - \beta_{j,k}^0)^2] = O(2^{J_n^*} n^{-1}) = O(n^{-2/3}) \quad (4.7)$$

In order to estimate the contribution to the risk by the other coefficients we use the following result. It improves the simple upper estimate  $\sum_{k: (j,k) \in \mathcal{I}_n} (\beta_{j,k}^0)^2 \wedge t_{n,j}^2 \leq t_{n,j} \sum_{k: (j,k) \in \mathcal{I}_n} |\beta_{j,k}^0| = O(n^{-2/3})$  and it is essential for the proof of Theorem 4.2 below.

**Lemma 4.1.** *Let  $m_0: (0, 1] \rightarrow \mathbb{R}$  be such that  $TV(m_0; \{x_1, \dots, x_n\}) \leq C_0$  and let  $t_{n,j} = K n^{-2/3} 2^{j/2}$ . Then, for  $\beta_{j,k}^0 = (1/n) \sum_{t=1}^n \psi_{j,k}(x_t) m(x_t)$ ,*

$$S_n(m_0) := \sum_{j: 2^j \geq n^{1/3}} \sum_{k: (j,k) \in \mathcal{I}_n} (\beta_{j,k}^0)^2 \wedge t_{n,j}^2 = O(n^{-2/3}).$$

Since  $\text{var}(\tilde{\beta}_{j,k}) = O(t_{n,j})$  for all  $(j,k) \in \mathcal{I}_n$ ,  $j \geq J_n^*$ , we obtain from (i) of Theorem 3.1, that for  $j \geq J_n^*$ ,

$$\begin{aligned} & \sum_{k:(j,k) \in \mathcal{I}_n} E[(\hat{\beta}_{j,k} - \beta_{j,k}^0)^2] \\ &= \sum_{k:(j,k) \in \mathcal{I}_n} O\left(\min\{(\beta_{j,k}^0)^2, t_{n,j}^2\} + E[(\tilde{\beta}_{j,k} - \beta_{j,k}^0)^2 \mathbb{1}(|\tilde{\beta}_{j,k} - \beta_{j,k}^0| > t_{n,j}/2)]\right). \end{aligned}$$

Since

$$\sum_{k:(j,k) \in \mathcal{I}_n} \min\{(\beta_{j,k}^0)^2, t_{n,j}^2\} \leq t_{n,j} \sum_{k:(j,k) \in \mathcal{I}_n} |\beta_{j,k}^0| = O(n^{-2/3})$$

and

$$E[(\tilde{\beta}_{j,k} - \beta_{j,k}^0)^2 \mathbb{1}(|\tilde{\beta}_{j,k} - \beta_{j,k}^0| > t_{n,j}/2)] \leq E[(\tilde{\beta}_{j,k} - \beta_{j,k}^0)^4] / (t_{n,j}/2)^2 = O(n^{-2/3} 2^{-j}),$$

it follows that

$$\sum_{k:(j,k) \in \mathcal{I}_n} E[(\hat{\beta}_{j,k} - \beta_{j,k}^0)^2] = O(n^{-2/3}). \quad (4.8)$$

Since  $J_n - J_n^* = O(\log(n))$ , we obtain from (4.7) and (4.8) the following result.

**Theorem 4.1.** *Suppose that **(A2)** is fulfilled. Then*

$$\sup \left\{ E \left[ \frac{1}{n} \sum_{t=1}^n (\hat{m}_n(x_t) - m_0(x_t))^2 \right] : TV(m_0; [0, 1]) \leq C_0 \right\} = O(n^{-2/3} \log(n)).$$

This result can be improved if we replace condition **(A2)** by the following majorization condition for the distribution of the empirical wavelet coefficients.

- (A2')** (i)  $E[(\tilde{\alpha}_0 - \alpha_0^0)^2] + \sum_{j=0}^{J_n^*-1} \sum_{k:(j,k) \in \mathcal{I}_n} E[(\tilde{\beta}_{j,k} - \beta_{j,k}^0)^2] = O(n^{-2/3}).$   
(ii) There exists a distribution function  $G$  on  $[0, \infty)$  (not necessarily a probability distribution function) such that  $\int_0^\infty x^4 dG(x) < \infty$  and

$$P(n^{1/2} \tilde{\beta}_{j,k} - \beta_{j,k}^0 > x) \leq 1 - G(x) \quad \forall x \geq 0 \quad (4.9)$$

holds for all  $(j,k) \in \mathcal{I}_n$ ,  $j \geq J_n^*$ .

*Remark 1.* Condition (4.9), as it stands, is a high-level condition and such conditions should be avoided wherever possible. There are however scenarios where this relation follows from simple conditions on the errors  $\varepsilon_1, \dots, \varepsilon_n$ .

- 1) If  $\varepsilon_1, \dots, \varepsilon_n$  are jointly normal with zero mean, then

$$n^{1/2}(\tilde{\beta}_{j,k} - \beta_{j,k}^0) \sim N(0, \sigma_{j,k}^2),$$

where

$$\begin{aligned}\sigma_{j,k}^2 &= n^{-1} \sum_{s,t=1}^n \psi_{j,k}(x_s) \psi_{j,k}(x_t) \text{cov}(\varepsilon_s, \varepsilon_t) \leq n^{-1} \sum_{s=1}^n \psi_{j,k}^2(x_s) \sum_{t=1}^n \text{cov}(\varepsilon_s, \varepsilon_t) \\ &\leq \bar{\sigma}^2 := \max_{1 \leq s \leq n} \sum_{t=1}^n \text{cov}(\varepsilon_s, \varepsilon_t).\end{aligned}$$

In this case,  $G(x) = 2\Phi(x/\bar{\sigma}) - 1$  satisfies (4.9).

- 2) If  $\varepsilon_1, \dots, \varepsilon_n$  are independent with zero mean

$$\max_{1 \leq t \leq n} E[|\varepsilon_t|^\gamma] < \infty$$

for some  $\gamma > 4$ , then we obtain by Rosenthal's inequality (see Theorem 3 in Rosenthal (1970)) that

$$\begin{aligned}& E\left[ \sqrt{n}(\tilde{\beta}_{j,k} - \beta_{j,k}^0)^\gamma \right] \\ &= E\left[ \frac{1}{\sqrt{n}} \sum_{t=1}^n \psi_{j,k}(x_t) \varepsilon_t^\gamma \right] \\ &\leq C_\gamma \max \left\{ \sum_{t=1}^n E\left[ \psi_{j,k}(x_t) \varepsilon_t / \sqrt{n} \right]^\gamma, \left( \sum_{t=1}^n E\left[ \psi_{j,k}(x_t) \varepsilon_t / \sqrt{n} \right]^2 \right)^{\gamma/2} \right\} \\ &\leq C_\gamma \max \{ E[|\varepsilon_t|^\gamma] : 1 \leq t \leq n \} =: C'_\gamma.\end{aligned}$$

(The latter inequality follows from  $(\psi_{j,k}(x_t)/\sqrt{n})^2 \leq (1/n) \sum_{t=1}^n \psi_{j,k}^2(x_t) = 1$ .) Then

$$P(n^{1/2} \tilde{\beta}_{j,k} - \beta_{j,k}^0 > x) \leq \min \{1, C'_\gamma / x^\gamma\} =: 1 - G(x) \quad \forall x \geq 0.$$

and it holds that  $\int_0^\infty x^4 dG(x) < \infty$ , as required.

- 3) If, for  $\gamma > 4$ ,  $\epsilon > 0$ , and some even number  $c > \gamma$ ,  $\varepsilon_1, \dots, \varepsilon_n$  are strong  $(\alpha-)$  mixing with coefficients satisfying

$$\sum_{r=1}^{\infty} (r+1)^{c-2} (\alpha(r))^{\epsilon/(c+\epsilon)} < \infty,$$

and if the  $\varepsilon_t$  have zero mean and  $\max_{1 \leq t \leq n} E[|\varepsilon_t|^{\gamma+\epsilon}] < \infty$ , then we obtain from a Rosenthal-type inequality (see e.g. Doukhan (1994, page 26, Theorem 2)) that

$$E\left[ \sqrt{n}(\tilde{\beta}_{j,k} - \beta_{j,k}^0)^\gamma \right] = E\left[ \frac{1}{\sqrt{n}} \sum_{t=1}^n \psi_{j,k}(x_t) \varepsilon_t^\gamma \right] \leq C \left( E[|\varepsilon_t|^{\gamma+\epsilon}] \right)^{\gamma/(\gamma+\epsilon)} =: C''_\gamma.$$

Then  $G$  given by  $G(x) := 1 - \min \{1, C''_\gamma / x^\gamma\}$  satisfies (4.9).

Under (A2') we obtain a rate of convergence without a logarithmic factor which is known to be optimal in many similar estimation problems.

**Theorem 4.2.** *If (A2') is fulfilled, then*

$$\sup \left\{ E \left[ \frac{1}{n} \sum_{t=1}^n (\hat{m}_n(x_t) - m_0(x_t))^2 \right] : TV(m_0; \{x_1, \dots, x_n\}) \leq C_0 \right\} = O(n^{-2/3}).$$

### 4.3. A partially linear model

In this section we add a linear trend and a seasonal component to our original model (2.1). To simplify notation, we suppose again that the time points  $x_1, \dots, x_n$  are equidistant and that the seasonal component has period  $p$ . This leads to the partially linear model

$$Y_t = ((x_1 - \bar{x}_n)/(x_n - x_1))\gamma_0^0 + \gamma_{(t \bmod p)+1}^0 + m_0(x_t) + \varepsilon_t, \quad t = 1, \dots, n. \quad (4.10)$$

Throughout this section we assume that (A1) and (A2) are fulfilled.

In a first step, the parameter  $\gamma^0 = (\gamma_0^0, \gamma_1^0, \dots, \gamma_p^0)^T$  is estimated by least squares. We rewrite (4.10) in vector/matrix form,

$$Y = X\gamma^0 + \bar{m}_0 + \varepsilon,$$

where  $Y = (Y_1, \dots, Y_n)^T$ ,  $\bar{m}_0 = (m_0(x_1), \dots, m_0(x_n))^T$ ,  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ , and

$$X = \begin{pmatrix} (x_1 - \bar{x}_n)/(x_n - x_1) & 1 & 0 & \dots & \dots & 0 \\ (x_2 - \bar{x}_n)/(x_n - x_1) & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ (x_{p-1} - \bar{x}_n)/(x_n - x_1) & 0 & \dots & 0 & 1 & 0 \\ (x_p - \bar{x}_n)/(x_n - x_1) & 0 & \dots & \dots & 0 & 1 \\ (x_{p+1} - \bar{x}_n)/(x_n - x_1) & 1 & 0 & \dots & \dots & 0 \\ (x_{p+2} - \bar{x}_n)/(x_n - x_1) & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix},$$

with  $\bar{x}_n = (x_1 + \dots + x_n)/n$ . To ensure identifiability of the parameters we choose  $m_0$  such that  $\sum_{t=1}^n t m_0(x_t) = 0$  and  $\sum_{1 \leq t \leq n: t \bmod p=k} m_0(x_t) = 0$  for  $k = 1, \dots, p$ , which implies that

$$X^T \bar{m}_0 = 0_{p+1}. \quad (4.11)$$

It is easy to see that

$$n^{-1} X^T X \xrightarrow[n \rightarrow \infty]{} \text{Diag}[1/12, 1/p, \dots, 1/p],$$

which means that  $X^T X$  is regular for sufficiently large  $n$  and

$$\begin{aligned}\widehat{\gamma}_n &= \arg \min_{\gamma} \|Y - X\gamma\|^2 \\ &= \arg \min_{\gamma} \|X\gamma^0 + \bar{m}_0 + \varepsilon - X\gamma\|^2 \\ &= \arg \min_{\gamma} \|X\gamma^0 + \varepsilon - X\gamma\|^2 \\ &= (X^T X)^{-1} X^T (X\gamma^0 + \varepsilon).\end{aligned}$$

Then

$$\begin{aligned}E\widehat{\gamma}_n &= \gamma_0, \\ E[(\widehat{\gamma}_n - \gamma_0)(\widehat{\gamma}_n - \gamma_0)^T] &= (X^T X)^{-1} X^T \text{Cov}(\varepsilon) X (X^T X)^{-1} = O(n^{-1}).\end{aligned}$$

In a second step we estimate  $m_0$  nonparametrically by wavelet thresholding. Let

$$\widetilde{Y}_t := Y_t - (X\widehat{\gamma}_n)_t = m_0(x_t) + (X(\gamma^0 - \widehat{\gamma}_n))_t + \varepsilon_t.$$

It follows from our identifiability condition (4.11) that  $\sum_{t=1}^n m_0(x_t) = 0$ . Hence,  $m_0$  can be represented as a linear combination of the wavelets and  $\phi_0$  is not needed. Let, for  $(j, k) \in \mathcal{I}_n$ ,

$$\begin{aligned}\widetilde{\beta}_{j,k} &= \frac{1}{n} \sum_{t=1}^n \psi_{j,k}(x_t) \widetilde{Y}_t \\ &= \beta_{j,k}^0 + \frac{1}{n} \sum_{t=1}^n \psi_{j,k}(x_t) (X(\gamma^0 - \widehat{\gamma}_n))_t + \frac{1}{n} \sum_{t=1}^n \psi_{j,k}(x_t) \varepsilon_t.\end{aligned}$$

It follows from Lemma 7.2 that  $E\left[\|\widehat{\gamma}_n - \gamma^0\|^4\right] = O(n^{-2})$ , which implies

$$\sup_t E\left[(X(\widehat{\gamma}_n - \gamma^0))_t^4\right] = O(n^{-2}),$$

and therefore, in conjunction with (4.3),

$$E[(\widetilde{\beta}_{j,k} - \beta_{j,k}^0)^4] = O(n^{-2}) \quad \forall (j, k) \in \mathcal{I}_n.$$

This is an analog to equation (4.3) which was the starting point for our calculations in the previous section and we obtain the following result.

**Proposition 4.1.** *Suppose that  $(Y_t)_t$  satisfies (4.10) and that (A1) and (A2) are fulfilled. Then*

$$\begin{aligned}E[\|\widehat{\gamma} - \gamma_0\|^2] &= O(n^{-1}), \\ \frac{1}{n} \sum_{t=1}^n (\widehat{m}_n(x_t) - m_0(x_t))^2 &= O_P(n^{-2/3} \log(n)).\end{aligned}$$

## 5. Simulations and data examples

### 5.1. Simulations

We illustrate the finite sample performance of the wavelet estimator with soft thresholding proposed in Section 4.2 using the following two trend functions:

$$f(t) = \begin{cases} 1.5 + t, & t \in [0, 1/2), \\ 0.1, & t \in [1/2, 2/3), \\ 3\sqrt{t - 2/3} + 0.1, & t \in [2/3, 1) \end{cases}$$

$$g(t) = \begin{cases} 10t - \lfloor 10t \rfloor, & t \in [0, 0.7), \\ 0.5, & t \in [0.7, 1). \end{cases}$$

The shape of the first function is motivated by our data example displayed in Figure 1.1 and analyzed in Section 5.2. The second function is used to illustrate that our approach can successfully estimate rough functions with several jumps. In both cases, we simulate an AR(1) noise process  $(\varepsilon_t)_t$  with autoregressive param . 5.1.

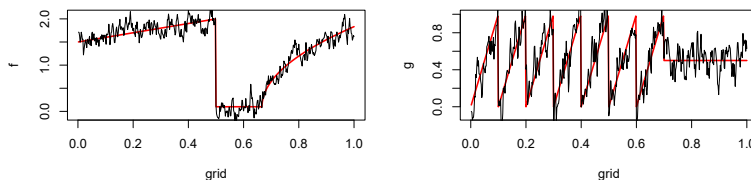


FIG 5.1. **Red:** function  $f$  (left) and  $g$  (right), **black:** data generated according to  $Y_t = f(t/n) + \varepsilon_t$  (left) and  $Y_t = g(t/n) + \varepsilon_t$  (right).

We calculate the wavelet estimator proposed in Section 4.2 and compare it to the Nadaraya-Watson estimator generated with the function `kreg` from the R package `gplm`. For the Nadaraya-Watson estimator we consider both, the rectangular and the Epanechnikov kernel. For our estimator as well as the kernel estimator we choose the tuning parameters (threshold parameter  $K$  and bandwidth  $b$ , respectively) in an MSE-optimal manner using a grid search. Figures 5.2 and 5.3 display the resulting box plots based on 1000 Monte Carlo iterations. In both settings, the wavelet estimator outperforms the competing kernel estimators. Moreover, note that the MSE-optimal bandwidths for the kernel estimators are much smaller than the corresponding default values chosen by Scott's rule of thumb (Scott (1992, p. 152, eq. 6.42)) which are  $b = 0.186$  for the rectangular kernel and  $b = 0.145$  for the Epanechnikov kernel, respectively, in case of the function  $g$ . The latter bandwidths lead to oversmooth estimators of the trend function, while the choices in Figures 5.2 and 5.3 result in an overfitting.

### 5.2. A real data example

The data set contains monthly overseas arrival data in Australia. More precisely, it consists of monthly recordings of international border crossings (in millions)

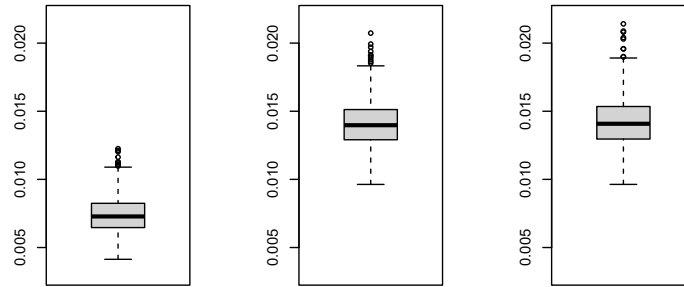


FIG 5.2. left: MSE of the wavelet estimator of  $f$  with  $K = 0.1$ , middle: MSE of the NW estimator of  $f$  (rectangular kernel with  $b = 0.009$ ) right: MSE of the NW estimator of  $f$  (Epanechnikov)

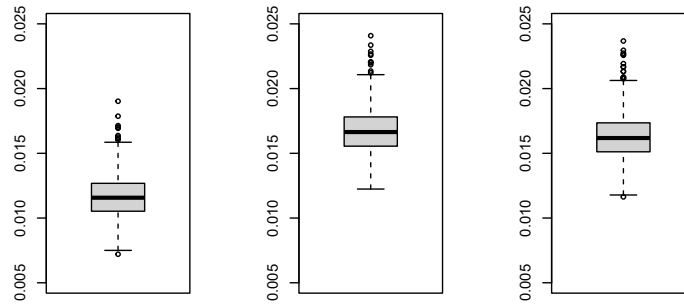


FIG 5.3. left: MSE of the wavelet estimator of  $g$  with  $K = 0.045$ , middle: MSE of the Nadaraya-Watson estimator of  $g$  (rectangular kernel with  $b = 0.006$ ) right: MSE of the Nadaraya-Watson estimator of  $g$  (Epanechnikov kernel with  $b = 0.007$ ).

from July 2014 to August 2024, retrieved from the Australian Bureau of Statistics<sup>1</sup>. As it can be seen from Figure 5.4, there was an abrupt decay of the arrivals in April 2020 resulting from travel restrictions due to the COVID pandemic.

We applied the partially linear model with soft thresholding, introduced in Section 4.3 with three different choices for the thresholding parameter  $K = 0.05$ ,  $K = 0.1$ ,  $K = 0.2$ . For comparison, we additionally fitted a partially linear model, where the nonlinear part is estimated via classical kernel regression (Nadaraya-Watson).

<sup>1</sup><https://www.abs.gov.au/statistics/industry/tourism-and-transport/overseas-arrivals-and-departures-australia/latest-release>

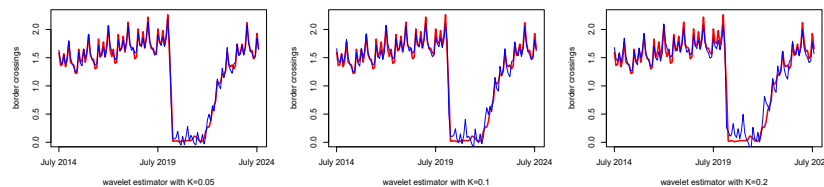


FIG 5.4. *red*: original data, *blue*: (partially linear) wavelet approximations with different choices of  $K$

From Figure 5.4 one can see that the wavelet-based estimator is perfectly able to adapt to the sharp decrease of the arrivals in 2020 irrespective of the choice of  $K$ .

In contrast, the decay of the kernel-type estimator is clearly slower if the bandwidth is chosen via Scott's rule of thumb, see Figure 5.5 (left). Reducing the bandwidth, we observe that the estimator can adapt better to the rough path of the data. The question of choosing an appropriate bandwidth is much more important here than choosing the truncation parameter  $K$  for our wavelet estimator. The latter perfectly captures the COVID jump for all three choices of  $K$

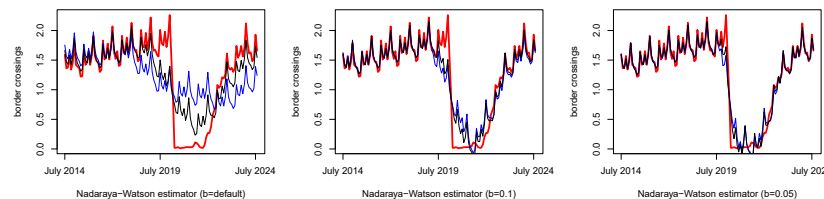


FIG 5.5. *red*: original data, *blue*: (partially linear) Nadaraya-Watson approximation with the rectangular kernel, *black*: (partially linear) Nadaraya-Watson approximation with the Epanechnikov kernel.

## 6. Proofs of the main results

*Proof of Proposition 3.1.* Since the pairs  $(\theta_1, \varepsilon_1), \dots, (\theta_N, \varepsilon_N)$  are by assumption independent and identically distributed, it suffices to consider the Bayes risk for a single coefficient. Let  $P_\theta$  be the distribution of  $Y_k$  given  $\theta_k = \theta$  and let  $T(Y_k)$  be an arbitrary estimator of  $\theta_k$ . Then its Bayes risk w.r.t. prior  $\pi$  and

squared error loss is given by

$$\begin{aligned}
& E[(T(Y_k) - \theta_k)^2] \\
&= \pi(\{0\}) \left\{ P_0(Y_k = -\lambda)T(-\lambda)^2 + P_0(Y_k = 0)T(0)^2 + P_0(Y_k = \lambda)T(\lambda)^2 \right\} \\
&\quad + \pi(\{-\lambda\}) \left\{ P_{-\lambda}(Y_k = -2\lambda)(T(-2\lambda) + \lambda)^2 + P_{-\lambda}(Y_k = -\lambda)(T(-\lambda) + \lambda)^2 \right. \\
&\quad\quad \left. + P_{-\lambda}(Y_k = 0)(T(0) + \lambda)^2 \right\} \\
&\quad + \pi(\{\lambda\}) \left\{ P_\lambda(Y_k = 0)(T(0) - \lambda)^2 + P_\lambda(Y_k = \lambda)(T(\lambda) - \lambda)^2 \right. \\
&\quad\quad \left. + P_\lambda(Y_k = 2\lambda)(T(2\lambda) - \lambda)^2 \right\} \\
&= (1-p) \left\{ (p/2)T(-\lambda)^2 + (1-p)T(0)^2 + (p/2)T(\lambda)^2 \right. \\
&\quad \left. + (p/2) \left\{ (p/2)(T(-2\lambda) + \lambda)^2 + (1-p)(T(-\lambda) + \lambda)^2 + (p/2)(T(0) + \lambda)^2 \right\} \right. \\
&\quad \left. + (p/2) \left\{ (p/2)(T(0) - \lambda)^2 + (1-p)(T(\lambda) - \lambda)^2 + (p/2)(T(2\lambda) - \lambda)^2 \right\} \right. \\
&= (p/2)^2 (T(-2\lambda) + \lambda)^2 \\
&\quad + (1-p)(p/2) \left\{ T(-\lambda)^2 + (T(-\lambda) + \lambda)^2 \right\} \\
&\quad + (1-p)^2 T(0)^2 + (p/2)^2 \left\{ (T(0) - \lambda)^2 + (T(0) + \lambda)^2 \right\} \\
&\quad + (1-p)(p/2) \left\{ T(\lambda)^2 + (T(\lambda) - \lambda)^2 \right\} \\
&\quad + (p/2)^2 (T(2\lambda) - \lambda)^2.
\end{aligned}$$

Therefore, the unique Bayes estimator  $T^*(Y_k)$  is given by  $T^*(-2\lambda) = -\lambda$ ,  $T^*(-\lambda) = -\lambda/2$ ,  $T^*(0) = 0$ ,  $T^*(\lambda) = \lambda/2$ ,  $T^*(2\lambda) = \lambda$ , and its Bayes risk is equal to

$$\begin{aligned}
& E[(T^*(Y_k) - \theta_k)^2] \\
&= (p/2)^2 0 \\
&\quad + (1-p) \left\{ (p/2)(\lambda/2)^2 + (p/2)(\lambda/2)^2 \right\} \\
&\quad + (1-p)^2 0 + (p/2) \left\{ (p/2)\lambda^2 + (p/2)\lambda^2 \right\} \\
&\quad + (1-p) \left\{ (p/2)(\lambda/2)^2 + (p/2)(\lambda/2)^2 \right\} \\
&\quad + (p/2)^2 0 \\
&= 4(1-p)(p/2)(\lambda/2)^2 + 2(p/2)^2 \lambda^2 \\
&= p\lambda^2/2 = \epsilon^2 q^{2/3}/2.
\end{aligned}$$

□

*Proof of Theorem 3.1.* (i) To prove (i) we distinguish between two cases.

$$1) \quad |\theta_k| < t/2$$

If  $|\varepsilon_k| = |Y_k - \theta_k| \leq t/2$ , then  $|Y_k| < t$  and so  $\hat{\theta}_k = 0$ . Otherwise, we

use the estimate  $|\widehat{\theta}_k - \theta_k| \leq |\widehat{\theta}_k - Y_k| + |Y_k - \theta_k| \leq t + |\varepsilon_k| \leq 3|\varepsilon_k|$ .  
Therefore,

$$E[(\widehat{\theta}_k - \theta_k)^2] \leq \theta_k^2 + 9E[\varepsilon_k^2 \mathbb{1}(|\varepsilon_k| > t/2)].$$

2)  $|\theta_k| \geq t/2$

In this case we use that  $|\widehat{\theta}_k - \theta_k| \leq t + |\varepsilon_k|$ , which implies by (3.2)

$$E[(\widehat{\theta}_k - \theta_k)^2] \leq 2t^2 + 2E[\varepsilon_k^2] \leq 2t^2 + 4\epsilon^2.$$

(ii) We have

$$\begin{aligned} (1/N) \sum_{k: |\theta_k| \leq t/2} \min\{\theta_k^2, t^2\} &\leq (t/2) (1/N) \sum_{k=1}^N |\theta_k| \\ &= (K\epsilon q^{-1/3}/2) \epsilon q = O(\epsilon^2 q^{2/3}). \end{aligned}$$

Moreover, it follows from Lemma 7.1 that

$$\begin{aligned} E[\varepsilon_k^2 \mathbb{1}(|\varepsilon_k| > t/2)] &= 2 \int_{t/2}^{\infty} (1 - Q_\varepsilon([-x, x])) x dx + (1 - Q_\varepsilon([-t/2, t/2])) (t/2)^2 \\ &\leq 2 \int_{t/2}^{\infty} (\epsilon/x)^4 x dx + (\epsilon/(t/2))^4 (t/2)^2 = 8\epsilon^4/t^2. \end{aligned}$$

This implies

$$\frac{1}{N} \sum_{k: |\theta_k| \leq t/2} E_\theta[(\widehat{\theta}_k - \theta_k)^2] = O(\epsilon^2 q^{2/3}). \quad (6.1a)$$

Since  $\#\{k \in \{1, \dots, N\}: |\theta_k| \geq t/2\} \leq (1 + \delta)\epsilon q/(t/2)$ , we obtain

$$\begin{aligned} \frac{1}{N} \sum_{k: |\theta_k| > t/2} E_\theta[(\widehat{\theta}_k - \theta_k)^2] &\leq (2t^2 + 4\epsilon^2) \#\{k \in \{1, \dots, N\}: |\theta_k| \geq t/2\} \\ &\leq (2t^2 + 4\epsilon^2) \epsilon q(1 + \delta)/(t/2) = O(\epsilon^2 q^2 \beta). \end{aligned} \quad (6.1b)$$

The second result follows from (6.1a) and (6.1b).  $\square$

*Proof of Theorem 4.2.* We have by assumption that

$$E[(\widetilde{\alpha}_0 - \alpha_0^0)^2] + \sum_{j=0}^{J_n^* - 1} \sum_{k: (j,k) \in \mathcal{I}_n} E[(\widetilde{\beta}_{j,k} - \beta_{j,k}^0)^2] = O(n^{-2/3}).$$

Since  $\text{var}(\widetilde{\beta}_{j,k}) = O(t_{n,j})$  for all  $(j, k) \in \mathcal{I}_n$ ,  $j \geq J_n^*$ , we obtain from (i) of Theorem 3.1 that

$$\begin{aligned} &\sum_{j=J_n^*}^{J_n} \sum_{k: (j,k) \in \mathcal{I}_n} E[(\widehat{\beta}_{j,k} - \beta_{j,k}^0)^2] \\ &= \sum_{j=J_n^*}^{J_n} \sum_{k: (j,k) \in \mathcal{I}_n} O\left(\min\{(\beta_{j,k}^0)^2, t_{n,j}^2\} + E[(\widetilde{\beta}_{j,k} - \beta_{j,k}^0)^2 \mathbb{1}(|\widetilde{\beta}_{j,k} - \beta_{j,k}^0| > t_{n,j}/2)]\right). \end{aligned}$$

Lemma 4.1 reveals that

$$\sum_{j=J_n^*}^{J_n} \sum_{k:(j,k) \in \mathcal{I}_n} \min \{ (\beta_{j,k}^0)^2, t_{n,j}^2 \} = O(n^{-2/3}).$$

Let now  $(j, k) \in \mathcal{I}_n$  with  $j \geq J_n^*$ . Then, since  $P(\tilde{\beta}_{j,k} - \beta_{j,k}^0 > x) \leq 1 - G(x/\sqrt{n}) \forall x \geq 0$ ,

$$\begin{aligned} & E \left[ (\tilde{\beta}_{j,k} - \beta_{j,k}^0)^2 \mathbb{1}(|\tilde{\beta}_{j,k} - \beta_{j,k}^0| > t_{n,j}/2) \right] \\ & \leq \int (n^{-1/2}x)^2 \mathbb{1}(|n^{-1/2}x| > Kn^{-2/3}2^{j/2-1}) dG(x) \\ & = n^{-1} \sum_{l=0}^{\infty} \int_{(Kn^{-1/6}2^{(j+l)/2-1}, Kn^{-1/6}2^{(j+l+1)/2-1}]} x^2 dG(x) \\ & \leq n^{-1} \sum_{l=0}^{\infty} \int_{(Kn^{-1/6}2^{(j+l)/2-1}, Kn^{-1/6}2^{(j+l+1)/2-1}]} \frac{x^4}{K^2 n^{-1/3} 2^{j+l-2}} dG(x) \\ & = 2^{-j} n^{-2/3} / K^2 \sum_{l=0}^{\infty} \int_{(Kn^{-1/6}2^{(j+l)/2-1}, Kn^{-1/6}2^{(j+l+1)/2-1}]} x^4 / 2^{l-2} dG(x). \end{aligned}$$

This implies

$$\begin{aligned} & \sum_{j=J_n^*}^{J_n} \sum_{k:(j,k) \in \mathcal{I}_n} E \left[ (\tilde{\beta}_{j,k} - \beta_{j,k}^0)^2 \mathbb{1}(|\tilde{\beta}_{j,k} - \beta_{j,k}^0| > t_{n,j}/2) \right] \\ & \leq \frac{4n^{-2/3}}{K^2} \sum_{m=0}^{\infty} \left( \frac{1}{2^0} + \dots + \frac{1}{2^m} \right) \int_{(Kn^{-1/6}2^{m/2}, Kn^{-1/6}2^{(m+1)/2}]} x^4 dG(x) \\ & \leq \frac{8n^{-2/3}}{K^2} \int_{(0, \infty)} x^4 dG(x), \end{aligned}$$

which completes the proof.  $\square$

## 7. A few auxiliary results

**Lemma 7.1.** *Let  $F$  be the cumulative distribution function of a nonnegative random variable. Then, for any  $t \geq 0$ ,*

$$\int_t^{\infty} x^2 dF(x) = 2 \int_t^{\infty} (1 - F(x))x dx + (1 - F(t))t^2.$$

*Proof.* Let  $b > t$  and

$$G(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ x^2, & \text{if } 0 \leq x \leq b, \\ b^2, & \text{if } x \geq b \end{cases}.$$

Since  $G$  is a continuous distribution function which has derivative  $2x$  on  $(0, b)$ , we obtain by integration by parts

$$\begin{aligned} \int_t^b x^2 dF(x) &= F(b)b^2 - F(t)t^2 - \int_t^b F(x) dG(x) \\ &= 2F(b) \int_0^b x dx - F(t)t^2 - 2 \int_t^b F(x) x dx \\ &= 2 \int_t^b (F(b) - F(x))x dx + (F(b) - F(t))t^2. \end{aligned}$$

The result follows with  $b \rightarrow \infty$ .  $\square$

**Lemma 7.2.** *Suppose that (A2) is fulfilled. Then*

$$E\left[\left(\sum_{s=1}^n a_s \varepsilon_s\right)^4\right] \leq 3C_1^2 \left(\sum_{s=1}^n a_s^2\right)^2 + C_2 \sum_{s=1}^n a_s^4.$$

*Proof.* We have that

$$\begin{aligned} E\left[\left(\sum_{s=1}^n a_s \varepsilon_s\right)^4\right] &= \sum_{s,t,u,v=1}^n a_s a_t a_u a_v E[\varepsilon_s \varepsilon_t \varepsilon_u \varepsilon_v] \\ &= \sum_{s,t,u,v=1}^n a_s a_t a_u a_v \{E[\varepsilon_s \varepsilon_t] E[\varepsilon_u \varepsilon_v] + E[\varepsilon_s \varepsilon_u] E[\varepsilon_t \varepsilon_v] + E[\varepsilon_s \varepsilon_v] E[\varepsilon_t \varepsilon_u]\} \\ &\quad + \sum_{s,t,u,v=1}^n a_s a_t a_u a_v \text{cum}(\varepsilon_s, \varepsilon_t, \varepsilon_u, \varepsilon_v) \\ &=: T_{n,1} + T_{n,2}. \end{aligned}$$

Then

$$\begin{aligned} T_{n,1} &= 3 \left( \sum_{s,t=1}^n a_s a_t E[\varepsilon_s \varepsilon_t] \right)^2 \\ &\leq 3 \left( \sum_{s,t=1}^n (a_s^2 + a_t^2)/2 |\text{cov}(\varepsilon_s, \varepsilon_t)| \right)^2 \\ &\leq 3 C_1^2 \left( \sum_{s=1}^n s_s^2 \right)^2 \end{aligned}$$

and since  $|a_s a_t a_u a_v| \leq (a_s^4 + a_t^4 + a_u^4 + a_v^4)/4$ ,

$$T_{n,2} = \sum_{s=1}^n a_s^4 \sum_{t,u,v=1}^n |\text{cum}(\varepsilon_s, \varepsilon_t, \varepsilon_u, \varepsilon_v)| \leq C_2 \sum_{s=1}^n a_s^4.$$

$\square$

*Proof of Lemma 4.1.* To simplify notation, let  $K = 1$ , which means that  $t_{n,j} = n^{-2/3}2^{j/2}$ . Before we delve into details of the proof we consider an extreme case in order to provide some intuition. If  $m_0 = \mathbb{1}_{[c,x_n]}$  for any  $c$  between  $x_1$  and  $x_n$ , then we obtain that at each scale  $j$  only one of the coefficients  $\beta_{j,k}^0$  can be non-zero with  $\beta_{j,k}^0 = O(2^{-j/2})$ . Since  $2^{-j/2} \geq t_{n,j}$  if and only if  $2^j \leq n^{2/3}$ , we obtain that

$$\begin{aligned} S_n(m_0) &\leq \sum_{j: n^{1/3} \leq 2^j \leq n^{2/3}} t_{n,j}^2 + \sum_{j: 2^j > n^{2/3}} \sum_{k: (j,k) \in \mathcal{I}_n} (\beta_{j,k}^0)^2 \\ &= O\left(n^{-4/3} \sum_{j: n^{1/3} \leq 2^j \leq n^{2/3}} 2^j\right) + O\left(\sum_{j: 2^j > n^{2/3}} 2^{-j}\right) = O(n^{-2/3}). \end{aligned}$$

In such a case, and in other cases as well, the desired result for  $S_n(m_0)$  will be obtained by cutting the terms  $|\beta_{j,k}^0|$  at the corresponding thresholds  $t_{n,j}$  up to a certain scale.

Now we turn to the general case. We have that

$$\begin{aligned} \beta_{j,k}^0 &\leq (1/n) \sum_{t=1}^n \psi_{j,k}^{(n)}(x_t) \left( \max \{m_0(x_t) : x_t \in I_{j,k}\} - \min \{m_0(x_t) : x_t \in I_{j,k}\} \right) / 2 \\ &\leq 2^{-j/2} \text{TV}(m_0; I_{j,k}) =: c_{j,k}. \end{aligned} \quad (7.1)$$

In order to estimate  $S_n(m_0)$  we replace the terms  $|\beta_{j,k}^0|$  by their upper estimates  $c_{j,k}$ . In what follows we make use of the relations between the  $c_{j,k}$  at adjacent scales and of the nested structure of the intervals  $I_{j,k}$ . If  $c_{j,k} > t_{n,j}$ , which is equivalent to  $\text{TV}(m_0; I_{j,k}) > n^{-2/3}2^j$ , then we obtain for  $(j', k') \subseteq \mathcal{I}_n$  with  $I_{j,k} \subseteq I_{j',k'}$  that

$$c_{j',k'} = 2^{-j'/2} \text{TV}(m_0; I_{j',k'}) \geq 2^{-j'/2} \text{TV}(m_0; I_{j,k}) > n^{-2/3}2^{j'/2} = t_{n,j'}. \quad (7.2a)$$

On the other hand, if  $c_{j,k} \leq t_{n,j}$ , which is equivalent to  $\text{TV}(m_0; I_{j,k}) \leq n^{-2/3}2^j$ , then we obtain for  $(j', k') \in \mathcal{I}_n$  with  $I_{j',k'} \subseteq I_{j,k}$  that

$$c_{j',k'} = 2^{-j'/2} \text{TV}(m_0; I_{j',k'}) \leq 2^{-j'/2} \text{TV}(m_0; I_{j,k}) \leq n^{-2/3}2^{j'/2} = t_{n,j'}. \quad (7.2b)$$

Let  $\mathcal{I}_{n,t} := \{(j, k) \in \mathcal{I}_n : 2^j \geq n^{1/3}, c_{j,k} > t_{n,j}\}$  and  $\mathcal{I}_{n,c} := \{(j, k) \in \mathcal{I}_n : 2^j \geq n^{1/3}, c_{j,k} \leq t_{n,j}\}$ . Then

$$S_n(m_0) \leq \sum_{(j,k) \in \mathcal{I}_{n,t}} t_{n,j}^2 + \sum_{(j,k) \in \mathcal{I}_{n,c}} c_{j,k}^2 =: S_{n,t}(m_0) + S_{n,c}(m_0). \quad (7.3)$$

In order to estimate the two terms on the right-hand side of (7.3) we make use of the nested structure of the subsets  $\mathcal{I}_{n,t}$  and  $\mathcal{I}_{n,c}$ . Let

$$\mathcal{I}_{n,t}^{\text{top}} := \{(j, k) \in \mathcal{I}_{n,t} : (j', k') \notin \mathcal{I}_{n,t} \text{ for all } I_{j',k'} \subseteq I_{j,k}\}$$

be the collection of those intervals which are on top of the ‘‘pyramid’’ of intervals  $I_{j,k}$  with  $(j, k) \in \mathcal{I}_{n,t}$ . Furthermore, let

$$\mathcal{I}_{n,c}^{\text{bottom}} := \{(j, k) \in \mathcal{I}_{n,c} : I_{j-1, [k/2]} \in \mathcal{I}_{n,t}\}$$

be the collection of those intervals from  $\mathcal{I}_{n,c}$  which reside on top of an interval  $I_{j,k}$  with  $(j, k) \in \mathcal{I}_{n,t}$ .

We have, for  $(j, k) \in \mathcal{I}_{n,t}^{\text{top}}$ , that

$$t_{n,j}^2 < t_{n,j} 2^{-j/2} \text{TV}(m_0; I_{j,k}) = n^{-2/3} \text{TV}(m_0; I_{j,k}).$$

Therefore, and since the intervals  $I_{j,k}$  with  $(j, k) \in \mathcal{I}_{n,t}^{\text{top}}$  are disjoint, we obtain that

$$\begin{aligned} S_{n,t}(m_0) &= \sum_{(j,k) \in \mathcal{I}_{n,t}^{\text{top}}} \sum_{(j',k') \in \mathcal{I}_n : I_{j,k} \subseteq I_{j',k'}} t_{n,j'}^2 \\ &\leq \sum_{(j,k) \in \mathcal{I}_{n,t}^{\text{top}}} \sum_{j' : n^{1/3} \leq 2^{j'} \leq 2^j} t_{n,j'}^2 \\ &\leq \sum_{(j,k) \in \mathcal{I}_{n,t}^{\text{top}}} 2 n^{-2/3} \text{TV}(m_0; I_{j,k}) \\ &\leq 2 n^{-2/3} \text{TV}(m_0; (0, 1]). \end{aligned} \quad (7.4a)$$

Note that we have for  $(j, k) \in \mathcal{I}_{n,c}^{\text{bottom}}$  that  $2^{j/2} c_{n,j} \leq n^{-2/3}$  and for  $(j', k') \in \mathcal{I}_n$  with  $I_{j',k'} \subseteq I_{j,k}$  that  $c_{j',k'} \leq 2^{(j-j')/2} c_{j,k}$ . Therefore, and since the intervals  $I_{j,k}$  with  $(j, k) \in \mathcal{I}_{n,c}^{\text{bottom}}$  are disjoint, we obtain that

$$\begin{aligned} S_{n,c}(m_0) &= \sum_{(j,k) \in \mathcal{I}_{n,c}^{\text{bottom}}} \sum_{(j',k') \in \mathcal{I}_n : I_{j',k'} \subseteq I_{j,k}} c_{j',k'}^2 \\ &\leq \sum_{(j,k) \in \mathcal{I}_{n,c}^{\text{bottom}}} \sum_{(j',k') \in \mathcal{I}_n : I_{j',k'} \subseteq I_{j,k}} 2^{(j-j')/2} c_{j,k} 2^{-j'/2} \text{TV}(m_0; I_{j',k'}) \\ &\leq \sum_{(j,k) \in \mathcal{I}_{n,c}^{\text{bottom}}} 2^{j/2} c_{j,k} \sum_{j' \geq j} 2^{(j-j')} \text{TV}(f; I_{j,k}) \\ &\leq 2 n^{-2/3} \text{TV}(m_0; (0, 1]). \end{aligned} \quad (7.4b)$$

The result follows from (7.3), (7.4a), and (7.4b).  $\square$

## Funding

The second author was supported in part by the Oberfrankenstiftung (project: FP01054).

## References

- AMATO, U., ANTONIADIS, A., DE FEIS, I., AND GIJBELS, I. (2021). Penalized robust estimators for sparse and high-dimensional linear models. *Statist. Meth. Appl.* **30**(1), 1–48.
- AMATO, U., ANTONIADIS, A., DE FEIS, I., AND GIJBELS, I. (2022). Penalized wavelet estimation and robust denoising for irregular spaced data. *Comp. Statist.* **37**(4), 1621–1651.
- AVERKAMP, R. AND HOUDRÉ, C. (2005). Wavelet thresholding for non-necessarily Gaussian noise: Functionality. *Ann. Statist.* **33**(5), 2164–2193.
- DAHLHAUS, R. AND NEUMANN, M.H. (2001). Locally adaptive fitting of semi-parametric models to nonstationary time series. *Stoch. Proc. Appl.* **91**(2), 277–308.
- DONOHO, D.L. (1995). De-noising by soft-thresholding. *IEEE Transactions on information theory* **41**(3), 613–627.
- DONOHO, D.L. AND JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**(3), 425–455.
- DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G., AND PICARD, D. (1995). Wavelet shrinkage: Asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**(2), 301–370.
- DOUKHAN, P. (1994). *Mixing: Properties and Examples*. Lecture Notes in Statistics **85**, Springer, New York.
- GAO, H.Y. (1993). Wavelet estimation of spectral densities in time series analysis. Ph.D. dissertation, Univ. California, Berkeley.
- KOLACZYK, E.D. (1999). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statist. Sinica* **9**, 119–136
- MAMMEN, E. AND VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25**(1), 387–413.
- NEUMANN, M.H. (1996). Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. *J. Time Ser. Anal.* **17**(6), 601–633.
- 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**(1), 38–76. (1997).
- ROSENTHAL, H.P. (1970). On the subspaces of  $L^p$  ( $p > 2$ ) spanned by sequences of independent random variables. *Israel Journal of Mathematics* **8**(3), 273–303.
- SCOTT, D.W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization*. New York: Wiley.