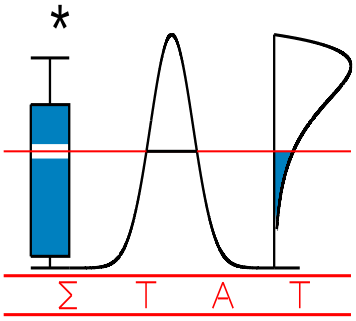
T E C H N I C A L

R E P O R T

0481

PARAMETRIC MODELLING OF THRESHOLDS ACROSS SCALES IN WAVELET REGRESSION

ANTONIADIS, A. and P. FRYZLEWICZ



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Parametric modelling of thresholds across scales in wavelet regression

Anestis Antoniadis Piotr Fryzlewicz*

March 21, 2005

Abstract

Wavelet thresholding is a natural and efficient approach for removing noise from data in nonparametric function estimation. In order to achieve spatial adaptation and benefit from the sparse representation of most signals in the wavelet domain, it is crucial to choose the thresholds correctly. An important property of the universal thresholding advocated by **?** is that the reconstruction is noise-free, i.e. when the true signal is constant, then, with high probability, the estimated function is also constant and equal to the empirical mean of the data. Motivated by this noise-free reconstruction property, we investigate a parametric thresholding procedure which takes advantage of the increasing sparsity of the wavelet coefficients across scales. We show that our estimator possesses the noise-free reconstruction property and achieves near-optimal risk rates for a large variety of signals over the Besov scale. The paper ends with a simulation study which demonstrates the excellent finite-sample performance of our method in comparison to a selection of state-of-the-art, as well as classical, wavelet denoising techniques.

Keywords: Wavelet decomposition, thresholding, noise-free reconstruction, Besov spaces, asymptotic rates.

1 Introduction

We are studying the classical nonparametric regression problem of recovering the values of an unknown function $f : [0, 1] \mapsto \mathbb{R}$ from noisy observations on an equidistant grid:

$$y_i = f(i/n) + \varepsilon_i, \quad i = 1, \ldots, n = 2^J, \tag{1}$$

where ε_i are independent and distributed as $N(0, \sigma^2)$. We measure the performance of an estimate $\hat{f}$ in terms of quadratic loss at the sample points. More specifically, let $\mathbf{f} =$

*Author for correspondence. Until 31 August 2005: Department of Mathematics, South Kensington Campus, Imperial College London, London SW7 2AZ, UK; email: p.fryzlewicz@imperial.ac.uk. From 1 September 2005: Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK.

$(f(i/n))_{i=1,\ldots,n}$ and $\hat{\mathbf{f}} = \left(\hat{f}(i/n)\right)_{i=1,\ldots,n}$ denote the vectors of true and estimated sampled values, respectively. We measure the performance of $\hat{f}$ by the Mean-Squared Error (MSE)

$$\mathrm{MSE}(\hat{\mathbf{f}}, \mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\{\hat{f}(i/n) - f(i/n)\right\}^2, \tag{2}$$

which we wish to make as small as possible. Although the notation f suggests a function of a real variable u, in this paper we only work with the equally spaced sample points $u_i = i/n$.

Throughout the paper, we are concerned with estimators $\hat{f}$ based on *wavelets*: for an overview of wavelet methods in statistics, the reader is referred to the monograph of **?**. Given an orthonormal discrete wavelet transform $W : \mathbb{R}^n \mapsto \mathbb{R}^n$ of regularity r and with the low resolution cut-off $J_0 = 0$, we denote the vector of noisy wavelet coefficients by $y_{j,k} = Wy$, the vector of noise-free wavelet coefficients by $d_{j,k} = W\mathbf{f}$, the vector of estimated wavelet coefficients by $\hat{d}_{j,k} = W\hat{\mathbf{f}}$, and the vector of "wavelet noise" coefficients by $\varepsilon_{j,k} = W\varepsilon_i$, where $j = 0$ ($j = J - 1$) is the coarsest (finest) detail scale. At any given scale j, the detail coefficients are indexed by $k = 1, \ldots, 2^j$. The only smooth coefficient is indexed by $(j, k) = (-1, 1)$.

By the linearity of the wavelet transform W, in the wavelet domain (1) becomes

$$y_{j,k} = d_{j,k} + \varepsilon_{j,k}, \tag{3}$$

where, due to the orthonormality of W, the $\varepsilon_{j,k}$'s are i.i.d. zero mean Gaussian variables with a common variance σ^2. For many signals f, the representation (3) is sparse, i.e. only a few true coefficients $d_{j,k}$ are significantly different from zero, while most $d_{j,k}$'s are close, or equal, to zero. The main idea behind signal denoising via wavelets is to modify the noisy wavelet coefficients $y_{j,k}$ by means of a particular mapping, to obtain $\hat{d}_{j,k}$: a "denoised" version of $y_{j,k}$. The estimate $\hat{f}$ is then obtained upon applying the inverse wavelet transform W^{-1} to $\hat{d}_{j,k}$.

Motivated by the sparsity of the representation (3), **?** proposed *thresholding* as a way of estimating $d_{j,k}$ from $y_{j,k}$. Thresholding annihilates those empirical coefficients $y_{j,k}$ which fall below a certain threshold t, and, provided that t is chosen "correctly", turns out to be an extremely effective denoising technique despite its simplicity. For the *universal* threshold $t = \sigma\sqrt{2\log(n)}$ (**?**), the following *noise-free reconstruction property* is achieved: when the true signal f is constant, then, with high probability, the estimate $\hat{f}$ is also constant and equal to the empirical mean of $\{y_i\}_{i=1}^n$. This is an important and desirable property, for example, in wavelet-based functional ANOVA tests (see **?**). Asymptotically, it ensures that no noise is present in the reconstructed signal, which implies that the reconstruction is "visually appealing". Universal thresholding has also been shown to be asymptotically near-optimal in the minimax MSE sense over a variety of smoothness spaces (see **?**). However, even though the universal threshold is (asymptotically) the lowest threshold which satisfies the noise-free reconstruction property, it is still "too high" in the sense that its application often leads to oversmoothing.

Motivated by the often observed decreasing sparsity of wavelet coefficients from finer to coarser scales, some authors have proposed using scale-dependent thresholds t_j which often achieve improved MSE performance compared to scale-independent thresholds t, see e.g. **?**. One simple scale-dependent thresholding scheme is to set $t_j = \sigma\sqrt{2\log(n)}2^{(j-J+1)/2}, \quad j = 0, \ldots, J-1$. It is easy to see, however, that the noise-free reconstruction property is then lost.

In this paper, we aim to combine the two important issues mentioned above: the noise-free reconstruction property, and scale-dependent thresholding. The above discussion leads to an interesting question of whether it is possible to devise a scale-dependent thresholding scheme t_j which offers improved MSE performance compared to the universal threshold, but still retains the noise-free reconstruction property. Moreover, we are particularly interested in the case where we can impose some parametric dependence between the threshold values $\{t_j\}_{j=0}^{J-1}$, i.e. assume $t_j = \beta_\theta(j)$, where β_θ is a family of functions parametrized by θ whose dimension is (substantially) less that J. The rationale here is that by choosing t_j "jointly" (and not separately for each scale) we can potentially obtain a stable selection procedure even for coarser scales where only a few wavelet coefficients are available. The function β_θ will often be referred to as a "threshold profile".

The paper is organised as follows: in Section 2, we investigate a general noise-free reconstruction property of scale-dependent thresholding, leading to a specific family of threshold profiles. In Section 3, we investigate the risk properties of this new thresholding procedure, and show that it attains near-optimal MSE rates for signals from a range of Besov spaces. In Section 4, the performance of the new method (with a default choice of parameter θ) is investigated in an extensive simulation study: comparisons are made to a selection of state-of-the-art, as well as classical, wavelet denoising techniques. Finally, in Section 5, we introduce a simple computational technique for selecting the parameter θ in a data-driven way. Section 6 concludes the paper.

2 A generic noise-free reconstruction property

The estimator considered in this paper is the *hard thresholding* estimator

$$\hat{d}_{j,k}^{(h)}(t_j) = y_{j,k}\,\mathbb{I}\{|y_{j,k}| > t_j\}, \tag{4}$$

for $j = 0, \ldots, J-1$ and $k = 1, \ldots, 2^j$. However, results analogous to those obtained in this paper can also be derived for the soft thresholding case. We skip this case for simplicity, and due to the inferior practical performance of soft thresholding estimators, see e.g. **?**. We leave the smooth coefficient unchanged: $\hat{d}_{-1,1} = y_{-1,1}$. Note that due to the orthonormality of W we have

$$\mathrm{MSE}(\hat{\mathbf{f}}, \mathbf{f}) = \frac{1}{n}\sum_{j,k}\mathbb{E}\left\{\hat{d}_{j,k}^{(h)}(t_j) - d_{j,k}\right\}^2. \tag{5}$$

For notational simplicity, we assume $\sigma = 1$ throughout the paper. In practice, the parameter σ is often estimated from the data via the Median Absolute Deviation (MAD) estimator on the finest resolution level $J-1$, see e.g. **?**.

The normality and the i.i.d. nature of the "wavelet noise" coefficients in equation (3) are the key ingredients of the denoising-via-thresholding theory developed by **?**. In particular, the most frequently used universal thresholding procedure is based on the following familiar relation for i.i.d. standard normal variables $\varepsilon_{j,k}$:

$$P\left(\max_{j=0,\ldots,J-1;k=1,\ldots,2^j} |\varepsilon_{j,k}| > \sqrt{2\log n}\right) \to 0 \quad \text{as } n \to \infty.$$

Using the above result, it can easily be shown that applying the scale-independent threshold $t_j = t = \sqrt{2\log n}$ in (4) leads to the noise-free reconstruction property: if $\mathbf{f}$ is a constant signal, then, with high probability, $\hat{\mathbf{f}}$ is also constant and equal to the empirical mean of $\{y_i\}_{i=1}^n$. It can also be demonstrated that the choice $t = \sqrt{2\log n}$ yields near-optimal MSE rates over a range of signal smoothness classes, and produces visually appealing reconstructions for relatively small sample sizes n. However, it is well known that the universal threshold oversmooths: for non-zero signals $\mathbf{f}$, too much signal gets killed in the process of thresholding. There arises a need for lower thresholds; however, replacing $t_j = t = \sqrt{2\log n}$ in (4) with $t_j = t = \sqrt{a\log n}$ for $a < 2$ ruins the noise-free reconstruction property in the sense that

$$P\left(\max_{j=0,\ldots,J-1;k=1,\ldots,2^j} |\varepsilon_{j,k}| > \sqrt{a\log n}\right) \not\to 0 \quad \text{as } n \to \infty$$

if $a < 2$. Thus, the only way of obtaining thresholds which are lower than the universal threshold $t = \sqrt{2\log n}$, but, possibly, still preserve the noise-free reconstruction property, is to resort to scale-dependent thresholds t_j.

As mentioned in Section 1, one other motivation for using scale-dependent thresholds t_j is the varying sparsity of the wavelet representation across scales. For many signals, their wavelet coefficient vectors $\{d_{j,k}\}_{k=1}^{2^j}$ are less sparse at coarser scales j, with more $d_{j,k}$'s significantly different from zero. Thus, in order to prevent the estimation of those $d_{j,k}$'s as zero in (4) (as this would unnecessarily kill significant information), the use of lower thresholds should be considered at coarser scales. Indeed, the fact that the sparsity often decreases from finer to coarser scales suggests using thresholding profiles which decrease from finer to coarser scales.

In the remainder of this section, we derive a sufficient condition for scale-dependent thresholds which (a) are lower than the universal threshold and decrease from finer to coarser scales, and (b) preserve the noise-free reconstruction property. Let $y_{j,k}$ denote wavelet coefficients of a "pure Gaussian noise" signal and assume that (possibly) different thresholds t_j are applied at each scale $j = 0, \ldots, J-1$. It can easily be shown that the noise-free reconstruction property occurs if and only if

$$P\left(|y_{J-1,1}| > t_{J-1} \vee \ldots \vee |y_{J-1,2^{J-1}}| > t_{J-1} \vee \ldots \vee |y_{0,1}| > t_0\right) \to 0 \tag{6}$$

as $n \to \infty$. We now investigate when this is the case. Denoting the pdf (cdf) of a standard normal by ϕ (Φ), we obtain

$$P\left(|y_{J-1,1}| > t_{J-1} \vee \ldots \vee |y_{J-1,2^{J-1}}| > t_{J-1} \vee \ldots \vee |y_{0,1}| > t_0\right) =$$

$$1 - P\left(|y_{J-1,1}| < t_{J-1} \wedge \ldots \wedge |y_{J-1,2^{J-1}}| < t_{J-1} \wedge \ldots \wedge |y_{0,1}| < t_0\right) =$$
$$1 - \prod_{j,k} P\left(|y_{j,k}| < t_j\right) = 1 - \prod_{j=0}^{J-1} \left(2\Phi\left(t_j\right) - 1\right)^{2^j} \leq 1 - \prod_{j=0}^{J-1} \left(1 - \frac{2\phi(t_j)}{t_j}\right)^{2^j} =$$
$$1 - \prod_{j=0}^{J-1} \left\{ \left(1 - \frac{2\phi(t_j)}{t_j}\right)^{\frac{t_j}{2\phi(t_j)}-1} \right\}^{\frac{2\phi(t_j)2^j}{t_j - 2\phi(t_j)}} \leq 1 - \prod_{j=0}^{J-1} \exp\left(-\frac{2\phi(t_j)2^j}{t_j - 2\phi(t_j)}\right) =$$
$$1 - \exp\left(-\sum_{j=0}^{J-1} \frac{2\phi(t_j)2^j}{t_j - 2\phi(t_j)}\right),$$

using

$$\left(1 - \frac{1}{x}\right)^{x-1} \downarrow e^{-1} \tag{7}$$

as $x \to \infty$. Note that for large J,

$$1 - \exp\left(-\sum_{j=0}^{J-1} \frac{2\phi(t_j)2^j}{t_j - 2\phi(t_j)}\right) \leq 1 - \exp\left(-\sum_{j=0}^{J-1} \frac{4\phi(t_j)2^j}{t_j}\right), \tag{8}$$

provided that $\min t_j \to \infty$ as $J \to \infty$, which is a very natural condition. Thus, a sufficient condition for the "noise-free reconstruction" property is

$$\lim_{J\to\infty} \sum_{j=0}^{J-1} \frac{\phi(t_j)2^j}{t_j} = 0. \tag{9}$$

We assume that our thresholds are of the form

$$t_j = \sqrt{2\log(n)}t_\theta\left(\frac{j}{J-1}\right), \tag{10}$$

where $t_\theta(z) : [0,1] \mapsto [\delta, 1]$ is a family of continuous, nondecreasing functions ($\delta > 0$). Note that setting $t_\theta(z) \equiv 1$ yields the classical universal threshold. Continuing from (9), we have

$$\begin{aligned}\sum_{j=0}^{J-1} \frac{\phi(t_j)2^j}{t_j} &= \frac{1}{\sqrt{4\pi J\log(2)}} \sum_{j=0}^{J-1} \frac{2^{-Jt_\theta^2(j/(J-1))+j}}{t_\theta(j/(J-1))} \\ &\leq \frac{1}{\delta\sqrt{4\pi J\log(2)}} \sum_{j=0}^{J-1} 2^{-Jt_\theta^2(j/(J-1))+Jj/(J-1)}.\end{aligned}$$

Since $J \to \infty$, it is enough to investigate when the sum is bounded in J. The sum behaves like

$$J\int_0^1 2^{J(x-t_\theta^2(x))}dx. \tag{11}$$

Clearly, if $t_\theta^2(x) \leq x$ on any set of non-zero measure in $[0,1]$, then (11) is not bounded. Of course, we cannot speak here of the "smallest permitted" $t_\theta^2(x)$, as any such that $t_\theta^2(x) \geq \delta$ and $t_\theta^2(x) > x$ a.e. will do, but for simplicity we single out "almost the smallest permitted" $t_\theta^2(x)$ of the following form:

$$t_\delta^2(x) = \delta + (1-\delta)x, \tag{12}$$

which is a natural lower boundary for the following family of functions parametrized by a one-dimensional parameter $\theta \in [\delta, 1]$:

$$t_\theta(x) = \sqrt{\theta + (1-\theta)x}. \tag{13}$$

Note that with $t_\delta^2(x)$ defined as in (12) the integral (11) is indeed bounded in J as we have

$$J \int_0^1 2^{J(x - t_\delta^2(x))} dx = \frac{1 - 2^{-J\delta}}{\delta \log(2)}. \tag{14}$$

To summarise, $t_\delta(x)$ can be seen as "almost the smallest permitted" threshold profile which still guarantees asymptotically noise-free reconstruction.

Motivated by the above discussion, we propose to estimate $d_{j,k}$ by the hard thresholding estimator (4) with t_j of the form

$$t_j = \sqrt{2\log(n)} \sqrt{\theta + (1-\theta)\frac{j}{J-1}}. \tag{15}$$

In the remaining part of the paper, we study the theoretical risk properties of the proposed estimator, as well as its practical performance. Due to the particular form of the threshold profile (13), we label the new estimator "SQRT".

3 Risk properties of the SQRT estimator

In this section, we consider the Mean-Square Error properties of the SQRT estimator. We assume that the unknown signal f belongs to a Besov ball of radius $C > 0$ on $[0,1]$, $B_{p,q}^\nu(C)$, where $\nu > 0$ and $0 < p, q \leq \infty$. The parameter p can be viewed as the measure of inhomogeneity of f while ν measures its smoothness. Roughly speaking, the (not necessarily integer) parameter ν indicates the number of derivatives of f, where their existence is required in the L^p-sense, while the additional parameter q provides a further finer gradation. Besov classes have an exceptional expressive power. In particular, they include the traditional Hölder and Sobolev classes of smooth functions ($p = q = \infty$ and $p = q = 2$, respectively) but also various classes of spatially inhomogeneous functions like the class of functions of bounded variation, "sandwiched" between $B_{1,\infty}^1$ and $B_{1,1}^1$. In addition, note that if the father and mother wavelets have regularity $r > 0$, then the corresponding wavelet basis is an unconditional basis for the Besov spaces $B_{p,q}^\nu([0,1])$ for $0 < r\nu < r$, $0 < p, q \leq \infty$. This allows one to characterise Besov balls in terms of the wavelet coefficients $\tilde{d}_{j,k} = d_{j,k}/\sqrt{n}$ of

the function f in the following way. Define the *Besov sequence ball* of radius C as

$$b^{\nu}_{p,q}(C) = \left\{ \tilde{d}_{j,k} : \sum_{j=0}^{\infty} 2^{jsq} \|\tilde{d}_j\|_p^q \leq C^q \right\}, \tag{16}$$

where $s = \nu + 1/2 - 1/p$ and $\|\tilde{d}_j\|_p^p = \sum_{k=1}^{2^j} |\tilde{d}_{j,k}|^p$. The membership of f in $B^{\nu}_{p,q}(C)$ can be thought of as being equivalent to the membership of $\{\tilde{d}_{j,k}\}_{j,k}$ in $b^{\nu}_{p,q}(C)$. The reader is referred to **?** for rigorous definitions and a detailed study of Besov spaces.

The following theorem establishes the MSE near-optimality of our SQRT estimator over a wide range of Besov sequence spaces.

Theorem 3.1 *Given the regression problem (1), let* $\hat{\mathbf{f}}$ *be the SQRT estimator of* $\mathbf{f}$*, constructed by applying the inverse Discrete Wavelet Transform to the sequence of estimated wavelet coefficients* $\hat{d}^{(h)}_{j,k}(t_j)$ *with thresholds* t_j *defined by (15), for any fixed* $\theta \in [\delta, 1]$*. Denote* $\tilde{\hat{d}}^{(h)}_{j,k}(t) = \hat{d}^{(h)}_{j,k}(t)/\sqrt{n}$*. If* $0 < p, q \leq \infty$ *and* $\nu > 1/p$*, then*

$$\begin{aligned} \sup_{\tilde{d}_{j,k} \in b^{\nu}_{p,q}(C)} \mathrm{MSE}(\hat{\mathbf{f}}, \mathbf{f}) &= \frac{\sigma^2}{n} + \sup_{\tilde{d}_{j,k} \in b^{\nu}_{p,q}(C)} \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} \mathbb{E} \left\{ \tilde{\hat{d}}^{(h)}_{j,k}(t_j) - \tilde{d}_{j,k} \right\}^2 \\ &\leq C_0 \log(n) n^{-\frac{2\nu}{2\nu+1}}, \end{aligned}$$

where C_0 *is independent of* n*.*

The rate $O(n^{-\frac{2\nu}{2\nu+1}})$ is the best possible MSE rate for Besov spaces, and SQRT achieves it up to the logarithmic term: hence the name "near-optimality". The above rate is identical to that achieved by the classical universal thresholding estimator. The proof of Theorem 3.1 can be found in the Appendix.

4 Empirical performance of the SQRT estimator

In this section, we compare the finite-sample performance of the SQRT estimator to a selection of state-of-the-art, as well as classical, wavelet denoising methods. The material in this section is divided into 3 subsections. In section 4.1, we attempt to demonstrate that the new method is superior to classical universal thresholding in terms of MSE, and that the reconstructions have a similar visual quality. In section 4.2, we compare the translation-invariant (TI) version of our technique to classical TI universal thresholding, as well as to the empirical Bayes (eBayes) procedure of **?**. Finally, in section 4.3, we compare our procedure to denoising algorithms based on complex-valued wavelets, proposed by **?**.

	UNI0	UNI3	SQRT
bumps	391	382	314
doppler	148	145	123
heavisine	99	91	70
blocks	204	202	165

Table 1: ISE averaged over 100 sample paths ($\times 1000$ and rounded) for the 3 competing methods. Box indicates best result. See the discussion in Section 4.1.

4.1 Comparison with classical universal thresholding

The aim of this section is to argue that our method offers a good replacement for classical universal thresholding, in that it produces estimates which are superior in terms of MSE but have an equally high visual quality, due to the noise-free reconstruction property. Our test functions are Donoho and Johnstone's bumps, doppler, heavisine and blocks, sampled at 1024 equispaced points, and rescaled in such a way that their (minima, maxima) are $(0, 10.11)$, $(-2.49, 2.47)$, $(-6, 4)$ and $(-2, 5.2)$, respectively. The respective root signal-to-noise ratios are: 1.33, 1.45, 2.97, 1.91 (note that these signal-to-noise ratios are relatively low, i.e. the observed signals have a considerably noisy appearance). We use a randomly selected smooth wavelet (Daubechies' Extremal Phase with 5 vanishing moments) except for blocks where we use Haar. Periodic boundary conditions are assumed. The standard deviation of the noise is always 1 but it is unknown to the estimation procedures and always estimated using Median Absolute Deviation on the finest detail level. All the procedures are based on the Decimated Discrete Wavelet Transform and use hard thresholding. The competitors are:

- UNI0 — universal thresholding with all levels thresholded;
- UNI3 — universal thresholding with all but the 3 coarsest levels thresholded;
- SQRT — our method with the thresholding profile defined by $t_{0.01}(x) = \sqrt{0.01 + 0.99x}$.

Table 1 shows the ISE for each method averaged over 100 simulated sample paths (and multiplied by 1000 and rounded for clarity of presentation). The best results for each signal are indicated by a box. The new method outperforms the classical universal thresholding by 15–23%.

Figure 1 goes some way to demonstrating that, as with classical universal thresholding, reconstructions obtained using our method also enjoy a "noise-free" character. In all of the plots, solid lines are estimates of the signals represented by the corresponding dotted lines, contaminated by noise simulated in S-Plus with the random seed set to 11 (chosen at random). The left column shows estimates obtained using the SQRT method, and the right column — using the UNI3 method.

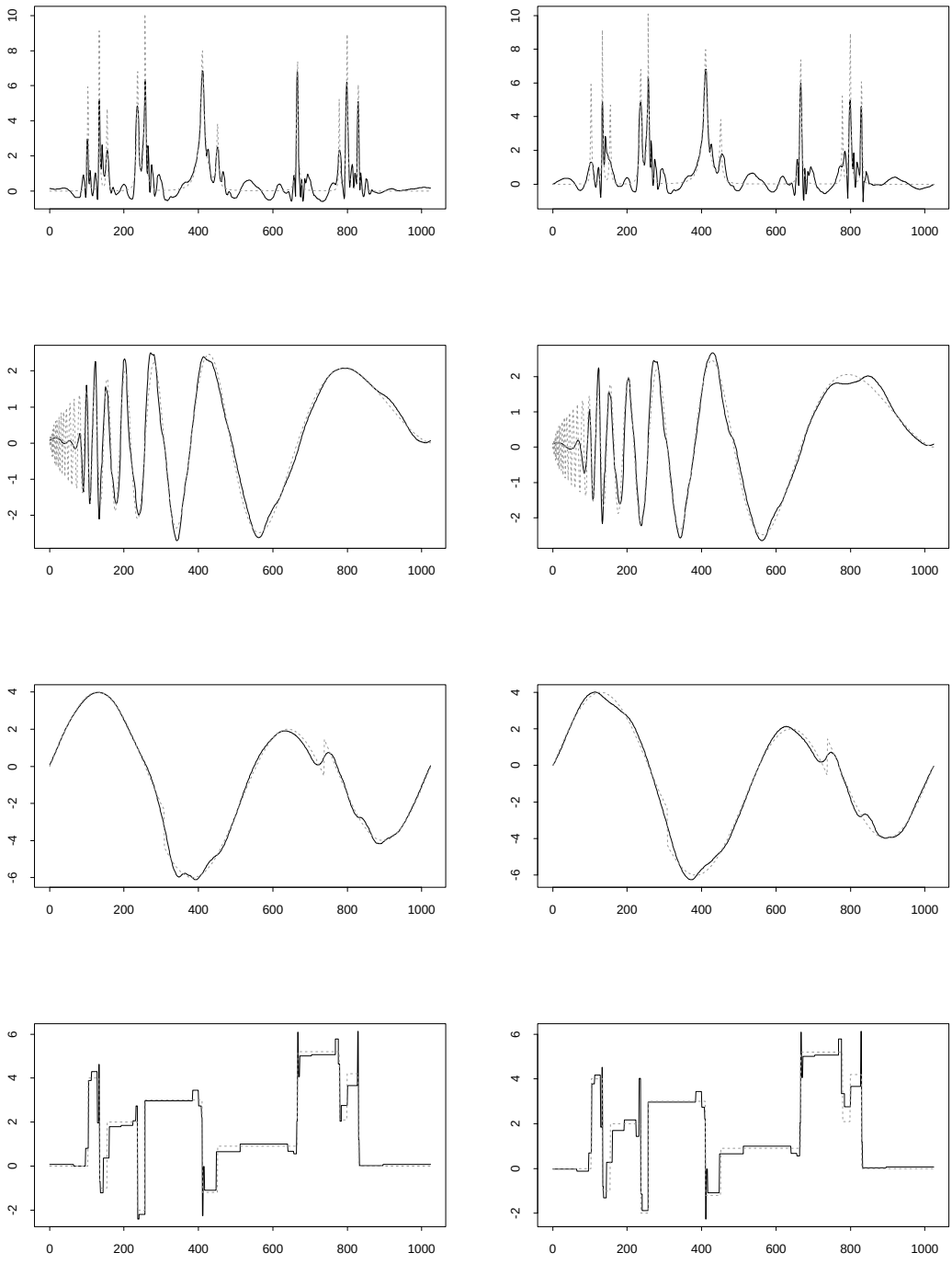


Figure 1: Sample reconstructions using SQRT (left column) and UNI3 (right column). See the discussion in Section 4.1.

4.2 Comparison with other techniques in a translation-invariant setting

Despite its simplicity, the translation-invariant (TI) version of the universal hard thresholding procedure (**?**) is a very powerful denoising tool. In a recent study assessing the empirical performance of various wavelet-based denosing methods (**?**), TI universal hard thresholding consistently performed the best, or nearly the best, among various modern wavelet smoothing techniques. Its performance was particularly good for longer signals. At the same time, it was shown to be a lot faster than several more sophisticated Bayesian methods. The variant of the TI universal hard thresholding considered in **?** uses thresholds of the form

$$t_j = \hat{\sigma}\sqrt{2J\log(2) + 2\log(J)}, \tag{17}$$

where $\hat{\sigma}$ is the MAD estimate of the noise level. In this section, we compare the performance of the TI version of our method to the classical TI universal hard thresholding, as well as to the TI version of the empirical Bayes (eBayes) procedure of **?** (in the simulation study reported therein, eBayes is shown to outperform several other denoising techniques, including the classical univeral thresholding (**?**), the SureShrink technique (**?**), techniques based on the False Discovery Rate (FDR; **?**), the block thresholding techniques of **?** as well as the QL method of **?**).

The competitors considered in this section are:

- TI-UNI3-TITH — translation-invariant hard thresholding with all but the 3 coarsest levels thresholded and thresholds of the form $t_j = \hat{\sigma}(2J\log(2) + 2\log(J))^{1/2}$;
- TI-UNI3 — translation-invariant hard thresholding with all but the 3 coarsest levels thresholded and thresholds of the form $t_j = \hat{\sigma}(2J\log(2))^{1/2}$;
- TI-SQRT-TITH — translation-invariant version of our method with the thresholding profile defined by $t_j = \hat{\sigma}(2J\log(2) + 2\log(J))^{1/2}(0.01 + 0.99j/(J-1))^{1/2}$;
- TI-SQRT — translation-invariant version of our method with the thresholding profile defined by $t_j = \hat{\sigma}(2J\log(2))^{1/2}(0.01 + 0.99j/(J-1))^{1/2}$;
- TI-EB — translation-invariant version of eBayes (note: the number of levels thresholded is equal to the maximum number of levels computable by the S-Plus routine `nd.dwt` from the `wavelets` module).

The experimental setup is exactly the same as in Section 4.1. Table 2 shows the ISE for each method averaged over 100 simulated sample paths (and multiplied by 1000 and rounded for clarity of presentation). The best results for each signal are indicated by a box. TI-SQRT outperforms TI-UNI3-TITH by 25–37%, TI-UNI3 by 10–26%, and TI-EB by 0–17%.

Figure 2, again produced with the random seed set to 11, shows sample reconstructions obtained using TI-SQRT (left column) and TI-EB (right column). The visual quality of the TI-SQRT estimates is clearly better than TI-EB for bumps and blocks, and very similar to TI-EB for doppler. For the heavisine function, TI-SQRT produces a spurious spike around

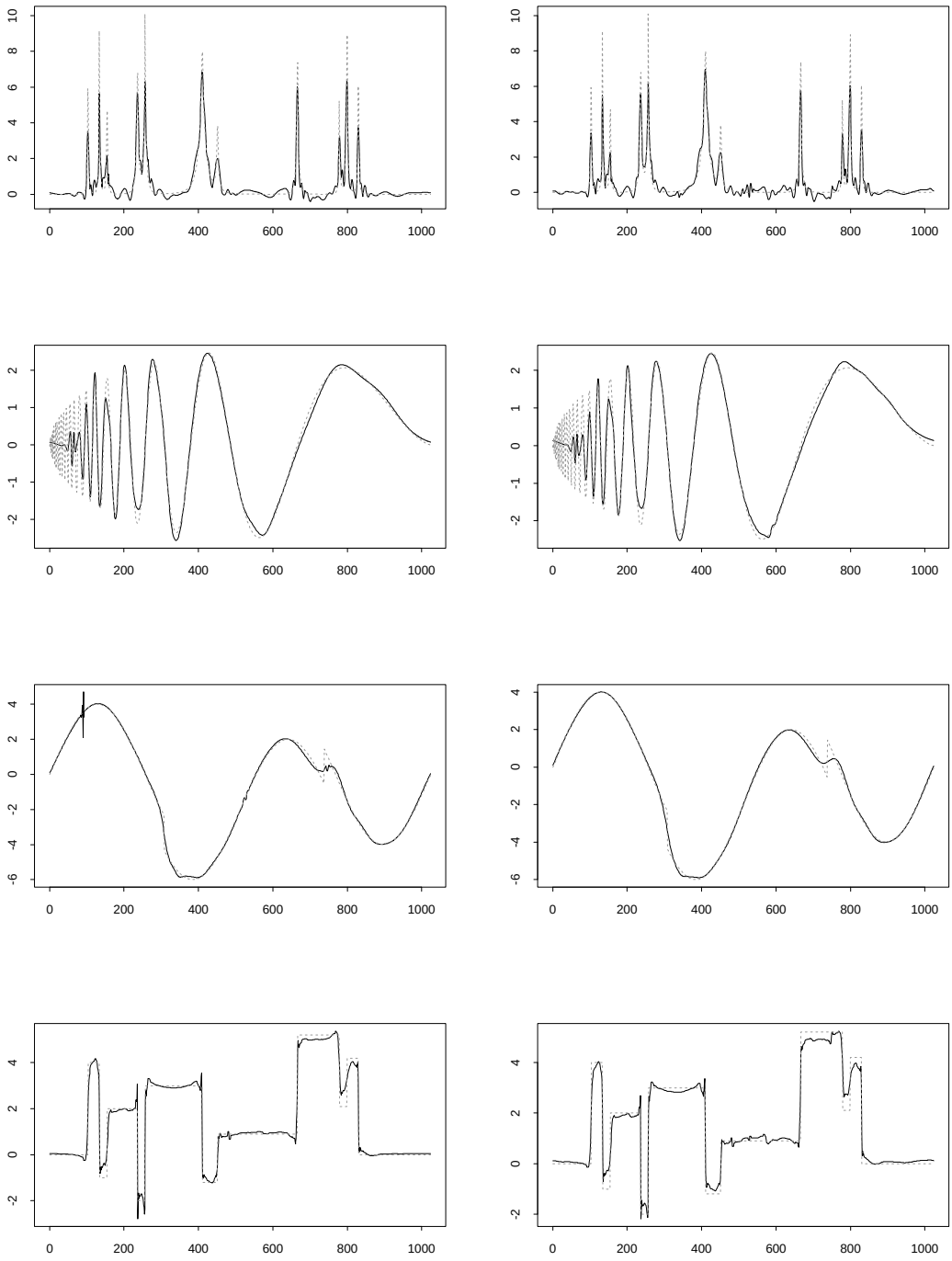


Figure 2: Sample reconstructions using TI-SQRT (left column) and TI-EB (right column). See the discussion in Section 4.2.

	TI-UNI3-TITH	TI-UNI3	TI-SQRT-TITH	TI-SQRT	TI-EB
bumps	214	171	167	145	151
doppler	85	73	69	64	69
heavisine	59	50	39	37	37
blocks	98	80	80	72	87

Table 2: ISE averaged over 100 sample paths ($\times 1000$ and rounded) for the 5 competing methods. Box indicates best result. See the discussion in Section 4.2.

the 100th observation. On the other hand, it indicates the second discontinuity in a clearer fashion than TI-EB.

4.3 Comparison with techniques based on complex wavelets

? report excellent performance of their denoising algorithms based on complex-valued wavelet decompositions. In a comprehensive simulation study, the proposed methods are shown to outperform several techniques based on real-valued wavelets. The basic idea underlying the proposed algorithms is as follows: decompose the n-element noisy vector into n "real" and n "imaginary" wavelet coefficients, couple them together according to their (scale, location) indices, and shrink the bivariate coefficients towards zero in a prescribed fashion to ensure noise removal. The inverse discrete complex wavelet transform then yields an estimate of the original signal.

When comparing the performance of a given denoising method based on complex-valued wavelets to a method which uses real-valued wavelets, care must be taken not to give "unfair advantage" to the former: it must be remembered that in the complex-valued wavelet case, the arising estimate is built of $2n$ shrunk real-valued detail coefficients as opposed to n in the real-valued wavelet case. Therefore, to ensure a fair comparison, it seems appropriate to us to consider estimators which arise as averages of two estimators, each computed using a different real-valued wavelet basis. Indeed, in this way, the final estimate is also built of $2n$ shrunk real-valued detail coefficients (like in the complex-valued wavelet case), and its performance can now be meaningfully compared to the complex-valued wavelet technique under consideration.

To compare our techniques with the methods proposed by **?**, we reproduce the simulation setup of **?**: we still use Donoho and Johnstone's bumps, doppler, heavisine and blocks sampled at 1024 equispaced points, but this time we rescale them in such a way that their sample variance equals 1. The standard deviation of the noise is set to $\sigma = 1/3$ to ensure that the root signal-to-noise ratio equals 3 (as before, σ is unknown to the estimation procedures and estimated via MAD). Table 3 shows the ISE for:

- BEST TI-CPLX — the best-performing translation-invariant method based on complex wavelets; the ISE results quoted are taken from **?**;
- BEST TI-SQRT — the best result obtained for TI-SQRT (see Section 4.2 for a de-

	BEST TI-CPLX	BEST TI-SQRT	2nd BEST SQRT	AVG 2 BEST TI-SQRT
bumps	1603	1695 (DEP2)	1803 (DEP1)	1496
doppler	710	830 (DLA9)	857 (DLA10)	832
heavisine	470	412 (DLA8)	416 (DLA6)	406
blocks	1727	744 (DEP1)	1691 (DEP2)	1064

Table 3: ISE averaged over 100 sample paths ($\times 1000$ and rounded) for the 2 competing methods. Box indicates best result. See the discussion in Section 4.3.

scription of the method) across a range of real-valued wavelet filters: Daubechies Extremal Phase (DEP) 1 to 10, and Daubechies Least Asymmetric (DLA) 4 to 10. The abbreviation in brackets indicates the wavelet basis which attains the minimum ISE;

- 2nd BEST TI-SQRT — as above, but 2nd best result;
- AVG 2 BEST TI-SQRT — estimator arising as the average of TI-SQRT for the two filters which attain the best and 2nd best ISE.

The results in Table 3 can be briefly summarised as follows: the best method based on complex-valued wavelets was outperformed by the technique which combined the best 2 TI-SQRT estimates, in 3 cases out of 4. Also, in 2 cases out of 4, the best complex-valued wavelet technique was outperformed by the best TI-SQRT method.

A few remarks are in order:

1. The SQRT estimator is of computational order $O(n)$, and the translation-invariant TI-SQRT estimator — of computational order $O(n \log(n))$. In practice, the software is extremely fast, which is partly due to the fact that the threshold choice is straightforward and requires no computationally intensive procedures.
2. While in practice the "optimal" analysing wavelet for the signal at hand is obviously unknown, a "good" wavelet can be chosen, for example, via the fast cross-validation algorithm of **?**.
3. The (TI-)SQRT algorithm is very easy to code in any package which implements the Discrete Wavelet Transform, e.g. the *WaveThresh* package for the *R* environment (both freeware).

5 Data-driven choice of θ

In the simulations reported in Section 4, we used the default value of $\theta = 0.01$, having found that it performed the best, or nearly the best, for a variety of spatially inhomogeneous signals. However, small values of θ cannot be expected to perform well for all signals. As

	UNI0	SQRT	SQRT-CV
zero	6	19	7
bumps	391	314	319
doppler	148	123	125
heavisine	99	70	75
blocks	204	165	170

Table 4: ISE averaged over 100 sample paths ($\times 1000$ and rounded) for the 3 competing methods. See the discussion in Section 5.

a counterexample, consider the zero signal, where the MSE of the SQRT estimator is a decreasing function of θ. Thus, there arises a need for a data-driven choice of θ.

In this section, we briefly describe a computational procedure, based on the "leave-half-out" cross-validation procedure of **?**, for choosing a "good" value of θ from the data. Given the value θ_l from a pre-selected grid $\{\theta_l\}_{l=1}^L$, we split the data $\{y_i\}_{i=1}^n$ into the "odd" subsample $\{y_{2i-1}\}_{i=1}^{n/2}$ and the "even" subsample $\{y_{2i}\}_{i=1}^{n/2}$. We then run the SQRT algorithm with parameter θ_l on the two subsamples to obtain the "odd" and "even" estimates, respectively. Finally, we measure the distance between the odd estimate and the even subsample, and add it to the distance between the even estimate and the odd subsample. The selected value of θ_l is the one which minimises the sum of these two distances. In practice, we have found that the grid $\theta_l = l/10$ for $l = 2, \ldots, 10$ and $\theta_1 = 0.01$ performs well.

The version of our SQRT algorithm which includes the above "cross-validatory" procedure for choosing θ is labelled SQRT-CV. To investigate the practical performance of SQRT-CV, we revisit the simulation setup of Section 4.1. The ISE values for bumps, doppler, heavisine, blocks, and the zero signal are given in Table 4. While SQRT-CV is never the best performing estimator in terms of ISE, it is always extremely close to the best one and is clearly the preferred option here, especially that the code is still fast, being of computational order $O(Ln) = O(10n)$ (or $O(n(L + \log(n))) = O(n(10 + \log(n)))$ for the translation-invariant version). The SQRT-CV algorithm is fully automatic, i.e. does not require the choice of any parameters by the user.

6 Conclusion

In this paper, we have proposed a new method for selecting threshold values in wavelet function estimation. Our proposed threshold values increase from coarser to finer scales, to reach the level of the classical universal threshold at the finest scale. They are parametrized by one scalar parameter, jointly for all scales.

The arising estimator, labelled SQRT due to the particular "square-root" shape of the threshold profile, preserves the important property of the classical universal threshold: the noise-free reconstruction property, which guarantees good visual quality of the SQRT estimates. At the same time, it achieves the usual near-optimal Mean-Square Error convergence

rates over a range of Besov smoothness classes.

In a detailed simulation study, we have investigated the finite-sample performance of our SQRT estimator and demonstrated its high visual quality and improved Mean-Square Error performance for a variety of spatially inhomogeneous signals, compared to the classical universal thresholding. Also, we have shown that it outperforms a number of state-of-the-art techniques in the translation-invariant setting. Those results have been obtained for the "default" value of the parameter of our procedure. Furthermore, we have proposed a simple and robust cross-validation-type technique for selecting the value of the parameter from the data, and have shown it to perform well on simulated examples (the arising estimator was labelled SQRT-CV).

Our SQRT algorithm and all of its variants investigated in this paper are fast and easy to code. Moreover, the SQRT-CV algorithm is fully automatic, i.e. does not require the choice of any parameters by the user.

The software implementing the SQRT, TI-SQRT and (TI)-SQRT-CV estimators is available upon request from the second author.

A Proof of Theorem 3.1

The first equality is due to the orthonormality of the Discrete Wavelet Transform. For n large enough (such that $n^\theta \geq 4$), we apply the "oracle inequality" from Theorem 7 of **?** to obtain

$$\begin{aligned}
\sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\mathbb{E}\left\{\tilde{\hat{d}}^{(h)}_{j,k}(t_j)-\tilde{d}_{j,k}\right\}^2 &\leq \sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\left(2.4+\left(\theta+(1-\theta)\frac{j}{J-1}\right)2J\log(2)\right)\\
&\times \left(2^{-J(\theta+(1-\theta)\frac{j}{J-1}+1)}+\min\left\{\tilde{d}^2_{j,k},n^{-1}\right\}\right)\\
&= \frac{1}{2^J}\sum_{j=0}^{J-1}2^j\left(2.4+\left(\theta+(1-\theta)\frac{j}{J-1}\right)2J\log(2)\right)2^{-J(\theta+(1-\theta)\frac{j}{J-1})}\\
&+ \sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\left(2.4+\left(\theta+(1-\theta)\frac{j}{J-1}\right)2J\log(2)\right)\min\left\{\tilde{d}^2_{j,k},n^{-1}\right\}\\
&= I+II.
\end{aligned}$$

Note that I is at most of order

$$\frac{2.4+2J\log(2)}{2^J}\int_0^J 2^{x-J(\theta+(1-\theta)\frac{x}{J-1})}dx \leq \frac{2.4+2J\log(2)}{2^J}2^{-1}2^{\frac{\theta J-1}{J-1}} = O\left(\frac{\log(n)}{n}\right), \quad (18)$$

which, incidentally, is of the same order as the corresponding quantity for the universal threshold:

$$\frac{2.4+2J\log(2)}{2^J}\sum_{j=0}^{J-1}2^{j-J} = O\left(\frac{J}{2^J}\right) = O\left(\frac{\log(n)}{n}\right). \quad (19)$$

We now focus on II. Since $\theta \in [\delta, 1]$ and $j \leq J-1$, note that II is less than the corresponding quantity for the classical universal threshold, which is

$$(2.4 + \log(n)) \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} \min \left\{ n^{-1}, \tilde{d}_{j,k}^2 \right\}. \tag{20}$$

Thus, instead of II, we shall consider (20). As $b_{p,q}^{\nu}(C) \subset b_{p,\infty}^{\nu}(C)$ for all q, we only have to consider the case $\tilde{d}_{j,k} \in b_{p,\infty}^{\nu}(C)$, so we can assume

$$\|\tilde{d}_j\|_p :=\leq C 2^{-js} \quad \text{for all} \quad j, \tag{21}$$

where C is a generic constant. The following argument was considered e.g. in **?**. We need to consider the cases $p \leq 2$ and $p > 2$ separately. For $p \leq 2$, we first note the simple inequality

$$\begin{aligned} \min\{|a|^2, |b|^2\} &= \min\{|a|^p, |b|^p\} \min\{|a|^{2-p}, |b|^{2-p}\} \leq |a|^{2-p} \min\{|a|^p, |b|^p\} \\ &= \min\{|a|^2, |a|^{2-p}|b|^p\}. \end{aligned}$$

Applying it with $a = n^{-1/2}$, $b = \tilde{d}_{j,k}$, we bound the double sum in (20) as follows:

$$\begin{aligned} \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} \min\{n^{-1}, \tilde{d}_{j,k}^2\} &\leq \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} \min\{n^{-1}, \left|\tilde{d}_{j,k}\right|^p n^{p/2-1}\} \\ &\leq \sum_{j=0}^{J-1} \min\{2^j n^{-1}, C^p 2^{-jsp} n^{p/2-1}\}. \end{aligned} \tag{22}$$

Note that

$$2^j n^{-1} \leq C^p 2^{-jsp} n^{p/2-1} \tag{23}$$

if and only if

$$j \leq J^* := \frac{2\log_2(C) + J}{2\nu + 1}. \tag{24}$$

Observe that asymptotically, we always have $J^* < J$. Assuming that J^* is an integer (it has no impact on the rates), we split (22) into two parts

$$\sum_{j=0}^{J*-1} 2^j n^{-1} + \sum_{j=J^*}^{J-1} C^p 2^{-jsp} n^{p/2-1}. \tag{25}$$

The first part is a partial sum of an increasing geometric series so, without going into details, it is bounded from above by a multiple of $n^{-1}2^{J^*} = O(n^{-2\nu/(2\nu+1)})$. The second part is a tail of decreasing geometric series so it is bounded from above by a multiple of $n^{p/2-1}2^{-J^*sp} = O(n^{-2\nu/(2\nu+1)})$. This proves the rate for $p \leq 2$.

For $p > 2$, first note that the Hölder inequality gives

$$\|\tilde{d}_j\|_2^2 \leq 2^{j(1-2/p)} \|\tilde{d}_j\|_p^2. \tag{26}$$

With this in mind, we bound the double sum in (20) as follows:

$$\begin{aligned}\sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\min\{n^{-1},\tilde{d}_{j,k}^2\} &\leq \sum_{j=0}^{J-1}\min\{2^j n^{-1},\|\tilde{d}_j\|_2^2\} \\ &\leq \sum_{j=0}^{J-1}\min\{2^j n^{-1},C^2 2^{-2js}2^{j(1-2/p)}\}. \qquad (27)\end{aligned}$$

As before, note that

$$2^j n^{-1} \leq C^2 2^{-2js}2^{j(1-2/p)} \qquad (28)$$

if and only if $j < J^*$. Again splitting the sum in (27) into two, we obtain

$$\sum_{j=0}^{J*-1} 2^j n^{-1} + \sum_{j=J^*}^{J-1} C^2 2^{-2js}2^{j(1-2/p)}. \qquad (29)$$

As we have already noted, the first part behaves like $O(n^{-2\nu/(2\nu+1)})$. The second part is a decreasing geometric series, so it is bounded from above by a multiple of $2^{J^*(-2s+1-2/p)} = O(n^{-2\nu/(2\nu+1)})$. This proves the desired rate for $p > 2$. □