HAL
open science

Detection threshold for non-parametric estimation

Abdourrahmane Atto, Dominique Pastor, Grégoire Mercier

To cite this version:

Abdourrahmane Atto, Dominique Pastor, Grégoire Mercier. Detection threshold for non-parametric estimation. Signal, Image and Video Processing, 2008, 2 (3), pp.207-223. 10.1007/s11760-008-0051-x . hal-00732333

HAL Id: hal-00732333

https://hal.science/hal-00732333v1

Submitted on 14 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Detection Threshold for Non-Parametric Estimation

Abdourrahmane M. Atto * Dominique Pastor †
Gregoire Mercier ‡
GET / ENST Bretagne, Technopôle Brest-Iroise,
CNRS UMR 2872 TAMCIC, Team TIME,
CS 83818 - 29238 Brest Cedex 3 - France

Abstract

A new threshold is presented for better estimating a signal by sparse transform and soft thresholding. This threshold derives from a non-parametric statistical approach dedicated to the detection of a signal with unknown distribution and unknown probability of presence in independent and additive white Gaussian noise. This threshold is called the detection threshold and is particularly appropriate for selecting the few observations, provided by the sparse transform, whose amplitudes are sufficiently large to consider that they contain information about the signal. An upper bound for the risk of the soft thresholding estimation is computed when the detection threshold is used. For a wide class of signals, it is shown that, when the number of observations is large, this upper bound is from about twice to four times smaller than the standard upper bounds given for the universal and the minimax thresholds. Many real-world signals belong to this class, as illustrated by several experimental results.

Keywords: Non-parametric estimation, soft thresholding, sparse transform, wavelet transform, non-parametric detection.

1 Introduction

This study concerns the non-parametric estimation of a signal in the sense of [7]. The aim of this estimation is to recover the signal from a noisy observation when noise is independent, additive, white and Gaussian.

The estimation is performed as follows. First, a linear orthonormal transform is applied to the observation. The outcome of this transform is a sequence of coefficients. The transform is chosen so that it represents the signal by a relatively small number of coefficients whose amplitudes are large in comparison to those resulting from noise. The second step is a non-linear filtering of these coefficients. The purpose of this filtering stage is to eliminate the noise components by forcing them to zero and,

* *am.atto@enst-bretagne.fr*
† *dominique.pastor@enst-bretagne.fr*
‡ *gregoire.mercier@enst-bretagne.fr*

possibly, to denoise the signal components. This filtering stage can be performed by a thresholding function $\delta_\lambda(\cdot)$. This function depends on a threshold λ whose main role is to distinguish the noisy signal components from those due to noise alone. A coefficient whose absolute value exceeds the threshold is regarded as a component of the noisy signal; a coefficient with absolute value below the threshold is considered as noise. The last step reconstructs the estimate of the signal on the basis of the filtered coefficients. The performance of this method is evaluated through a cost or risk function $r_\lambda(\cdot,\cdot)$, which will be the Mean Square Error (MSE) of the estimate.

To achieve the estimation described above, we must choose the appropriate transform, the thresholding function $\delta_\lambda(\cdot)$ and the value of the threshold λ used by the thresholding function. For reasons recalled below, the orthonormal Discrete Wavelet Transform (DWT) is appropriate. As far as the thresholding function is concerned, we choose the so-called soft thresholding function because it has the well-known and desirable properties of smoothness and adaptation (see [6]). The last parameter to specify is the value of the threshold.

This paper thus addresses the choice of the threshold to use for the estimation of a signal when the soft thresholding function is applied to the coefficients returned by the wavelet transform of a noisy observation of this signal. The literature on the topic distinguishes between the universal and the minimax thresholds introduced in Donoho and Johnstone's seminal paper [7]. The universal threshold is simply an estimate of the maximum of the amplitude that can be attained by the noise components. The minimax threshold is the largest value attaining the minimax quantity given by Eq. (6) below.

The thresholding function $\delta_\lambda(\cdot)$ basically forces to 0 any coefficient whose amplitude is less than the threshold λ because such a coefficient is considered to contain no or too little information about the signal. On the other hand, any coefficient with amplitude equal to or above λ is expected to relate to the presence of significant information about the signal; such a coefficient is then processed by the thresholding function to reduce the influence of noise. Therefore, in this paper, the choice of the threshold is regarded as a statistical decision problem where it is to be decided whether a given coefficient contains significant information about the signal or not. No assumption about the probability distributions of the signal coefficients is made, nor do we assume that these coefficients are identically distributed. Basically, our solution derives from [16], where a specific threshold is recommended to detect any signal whose amplitude is larger than or equal to a given value, when this signal is additively corrupted by independent White Gaussian Noise (WGN).

This paper is organized as follows. After recalling the main principles of the method introduced in [7], the results of [16] and those of [7] are combined in section 3 to introduce a new threshold. The performance of the resulting estimation by soft thresholding is then addressed in section 4. Section 5 concludes this paper.

2 Non-parametric soft thresholding estimation

Let $\boldsymbol{y} = \{y_i\}_{1\leqslant i\leqslant N}$ stand for the sequence of the observed data $y_i = f(t_i)+e_i$, $i = 1, 2, \ldots, N$, where f is an unknown function, the random variables $\{e_i\}_{1\leqslant i\leqslant N}$ are independent and identically distributed (iid), Gaussian with null mean and variance σ^2. For every $i = 1, 2, \ldots, N$, we write, as usual, that

$e_i \sim \mathcal{N}(0, \sigma^2)$.

The problem addressed in this work concerns the non-parametric estimation of the signal $\{f(t_i)\}_{1 \leqslant i \leqslant N}$ according to the approach developed in [7]. In order to recover the function $f(\cdot)$, an orthonormal transform, represented by an orthonormal matrix $\mathcal{W}$, is applied to $\boldsymbol{y}$. The outcome of this transform is the sequence of coefficients

$$c_i = \theta_i + \epsilon_i, \qquad i = 1, 2, \ldots, N, \tag{1}$$

where $\boldsymbol{c} = \{c_i\}_{1 \leqslant i \leqslant N} = \mathcal{W}\boldsymbol{y}$, $\boldsymbol{\theta} = \{\theta_i\}_{1 \leqslant i \leqslant N} = \mathcal{W}\boldsymbol{f}$, $\boldsymbol{f} = \{f(t_i)\}_{1 \leqslant i \leqslant N}$ and $\boldsymbol{\epsilon} = \{\epsilon_i\}_{1 \leqslant i \leqslant N} = \mathcal{W}\boldsymbol{e}$, $\boldsymbol{e} = \{e_i\}_{1 \leqslant i \leqslant N}$. The random variables $\{\epsilon_i\}_{1 \leqslant i \leqslant N}$ are iid and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The transform $\mathcal{W}$ is assumed to achieve a sparse representation [10] of the signal in the sense that, among the coefficients θ_i, $i = 1, 2, \ldots, N$, only a few of these have large amplitudes and, as such, characterize the signal. This heuristic notion of sparsity is sufficient, at this stage, to explain the estimation procedure. The wavelet transform is sparse in the sense given above and, as such, is recommended in [7] and [10].

When the thresholding function is applied to the coefficients $\{c_i\}_{1 \leqslant i \leqslant N}$, the coefficients with small amplitudes are forced to 0 - because they are considered to derive from too small, or even null, components of the signal - whereas, on the other hand, the noise contribution is reduced on those coefficients whose amplitudes exceed the threshold because such coefficients are regarded as large enough to pertain to the signal to estimate.

Denoting by $\widehat{\boldsymbol{\theta}} = \{\delta_\lambda(c_i)\}_{1 \leqslant i \leqslant N}$ the outcome of the non-linear filtering of the coefficients $\{c_i\}_{1 \leqslant i \leqslant N}$ by the thresholding function $\delta_\lambda(\cdot)$, the estimate of $\boldsymbol{f}$ is then $\widehat{\boldsymbol{f}} = \mathcal{W}^\mathsf{T}\widehat{\boldsymbol{\theta}}$, where $\mathcal{W}^\mathsf{T}$ is the transpose, and thus, the inverse, of $\mathcal{W}$.

The thresholding function considered is the soft thresholding function defined by

$$\delta_\lambda(x) = \begin{cases} x - \operatorname{sgn}(x)\lambda & \text{if } |x| \geqslant \lambda, \\ 0 & \text{elsewhere,} \end{cases} \tag{2}$$

where $\operatorname{sgn}(x) = 1$ (resp. -1) if $x \geqslant 0$ (resp. $x < 0$).

The risk function or cost used to measure the accuracy of the estimate $\widehat{\boldsymbol{f}}$ of $\boldsymbol{f}$ is the standard MSE. Since the transform $\mathcal{W}$ is orthonormal, this cost is

$$r_\lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \frac{1}{N}\mathbb{E}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big(\theta_i - \delta_\lambda(c_i)\Big)^2.$$

To state the following results, it is convenient to use the standard *oracle risk* introduced in [7]:

$$r_0(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\min\left(\theta_i^2, \sigma^2\right). \tag{3}$$

At this stage, it is time to recall the definition of the universal threshold and that of the minimax threshold (see [7]). These thresholds can be used to achieve the estimation by sparse transform and soft thresholding.

Consider Eq. (1). Since $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, it follows from [2, Eqs. (9.2.1), (9.2.2), Section 9.2, p. 187]; see also [13, p. 454], [19, Section 2.4.4, p. 91], that

$$\lim_{N \to +\infty} \mathrm{P}\left[\lambda_u(N) - \frac{\sigma \ln\ln N}{\ln N} \leqslant \max\Big\{|\epsilon_i|, 1 \leqslant i \leqslant N\Big\} \leqslant \lambda_u(N)\right] = 1, \tag{4}$$

where $\lambda_u(N) = \sigma\sqrt{2\ln N}$. Thus, the maximum amplitude of $\{\epsilon_i\}_{1\leqslant i\leqslant N}$ has a strong probability of being close to $\lambda_u(N)$ when N is large. The threshold $\lambda_u(N)$ is the so-called *universal threshold*. According to [7, Theorem 1], the risk $r_{\lambda_u(N)}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}})$ of the soft thresholding estimation of $\boldsymbol{\theta}$ with universal threshold $\lambda_u(N)$ is such that

$$r_{\lambda_u(N)}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}) \leqslant (1+2\ln N)\left(N^{-1}\sigma^2 + r_0(\boldsymbol{\theta})\right). \tag{5}$$

The *minimax threshold* $\lambda_m(N)$ is defined as the largest value λ among the values attaining the minimax risk bound

$$\Lambda(N) = \inf_{\lambda>0}\sup_{\mu\in\mathbb{R}} \tfrac{r_\lambda(\mu,\widehat{\mu})}{N^{-1}+r_0(\mu)}. \tag{6}$$

It follows from [7, Theorem 2] that the risk r_{λ_m} of the soft thresholding estimation of $\boldsymbol{\theta}$ with minimax threshold $\lambda_m(N)$ satisfies the following inequality

$$r_{\lambda_m(N)}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}) \leqslant \Lambda(N)\left(N^{-1}\sigma^2 + r_0(\boldsymbol{\theta})\right), \tag{7}$$

with $\Lambda(N) \leqslant 1+2\ln N$ and $\Lambda(N) \overset{N\to\infty}{\sim} 2\ln N$.

Remark 1 *According to the inequalities given in Eqs.* (5) *and* (7)*, the upper bound on the risk* $r_\lambda(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}})$ *of the estimation by soft thresholding, whether* λ *is either the universal or the minimax threshold, is of the same order as* $2r_0(\boldsymbol{\theta})\ln N$ *when* N *tends to* ∞*.*

Hereafter, a new threshold, obtained according to [16], will be introduced and its performance will be analysed in comparison with the minimax and universal thresholds. In particular, we will show that, for the soft thresholding estimation based on this threshold, the upper bound on the risk behaves, for a certain class of signals, as $r_0(\boldsymbol{\theta})\ln N$, or even $r_0(\boldsymbol{\theta})\ln N/2$, when N is large. This class is actually large enough to contain many real signals encountered in practice.

3 The detection threshold and its application to non-parametric estimation

In this section, by taking into account the sparsity of the model described by Eq. (1) and following the approach of [16, Theorem VII.1], we derive a threshold that improves the estimation by sparse transform and soft thresholding. The notations introduced in the preceding section are used hereafter with the same meaning as above.

For any real number λ, let $\mathcal{T}_\lambda(\cdot)$ be the *thesholding test with threshold height* λ defined by

$$\mathcal{T}_\lambda(x) = \begin{cases} 1 & \text{if } |x| \geqslant \lambda \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

for every real value x. Then, given any coefficient c_i, $i = 1,2,\ldots,N$, we have $\delta_\lambda(c_i) = \mathcal{T}_\lambda(c_i)(c_i - \operatorname{sgn}(c_i)\lambda)$. This simple equation emphasizes that, in the estimation process by sparse transform and thresholding function, the primary role of the threshold λ is to decide which coefficients must be

processed - because they can reasonably be expected to contain significant information about the signal - and which coefficients must be forced to zero - because they are assumed to contain no or too little information about this same signal.

Now, consider that the transform $\mathcal{W}$ satisfies the next two assumptions. These assumptions formalize, more specifically than above, that the coefficients pertaining to the signal are few and large.

(**F**) [Few:] Only a few coefficients of the sequence $\{c_i\}_{1 \leqslant i \leqslant N}$ contain significant information about the signal in the following sense: first, each coefficient c_i, $i = 1, 2, \ldots, N$, follows a binary hypothesis model where the null hypothesis is that c_i is noise only, so that $c_i = \epsilon_i$, and the alternative hypothesis is that c_i is the sum of signal and noise, so that $c_i = \theta_i + \epsilon_i$ with $\theta_i \neq 0$; second, the probability of occurrence of the alternative hypothesis is unknown but less than or equal to one half.

(**L**) [Large:] When the alternative hypothesis described above is true for a given coefficient c_i, the amplitude of the corresponding coefficient θ_i is larger than or equal to the universal threshold $\lambda_u(N) = \sigma\sqrt{2 \ln N}$. We recall that, according to Eq. (4), the universal threshold can be regarded as the maximum amplitude of the coefficients returned for noise when N is large enough.

Assumptions (**F**) and (**L**) are acceptable to model the statistical behaviour of the wavelet coefficients for smooth or piecewise regular signals ([7, 10]). Summarizing these assumptions, we can write that for every coefficient c_i, $i = 1, 2, \ldots, N$, the decision about the presence or the absence of significant information about the signal amounts to testing the null hypothesis $c_i \sim \mathcal{N}(0, \sigma^2)$ against the alternative hypothesis $c_i \sim \mathcal{N}(\theta_i, \sigma^2)$ where $|\theta_i| \geqslant \lambda_u(N)$. We hereafter assume that the noise standard deviation σ is known.

If assumptions (**F**) and (**L**) did not bound our lack of prior knowledge about the coefficients of the signal and the probabilities of occurrence of the alternative hypotheses, the use of Wald's test ([20]) would be recommended since the coefficients θ_i, $i = 1, 2, \ldots, N$, are unknown and the noise standard deviation is known. Given some test level, any positive real value r, and any coefficient θ_i, $i = 1, 2, \ldots, N$, such that $|\theta_i| = r$, Wald's test has best constant power for accepting the alternative hypothesis. We recall the following: the test level, or probability of error of the first type, is the probability of accepting the alternative hypothesis when the null hypothesis is true; the power of the test is the probability of rejecting the alternative hypothesis when this alternative hypothesis is true. The power of the test is also the complementary probability of the so-called probability of error of the second type, that is, the probability of accepting the alternative hypothesis when the latter is true.

Since we assume an upper-bound equal to one half for the probabilities of occurrence of the alternative hypotheses and a lower-bound equal to $\lambda_u(N)$ for the amplitudes of the coefficients θ_i, $i = 1, 2, \ldots, N$, when the alternative hypotheses occur, we can use proposition 1 below. This proposition derives from [16, Theorem VII.1]. In contrast with Wald's test, the criterion for the quality of the tests propounded in [16, Theorem VII.1] and the following statement is not the power of the test given some level, but the probability of error, that is, the probability of accepting the wrong hypothesis and, thus, the weighted average of the probabilities of the first and the second type. Therefore, the test proposed below does not require any *a priori* test level to be chosen.

In the following statement, $V(\rho, p)$ stands for the function defined for every non-negative real number ρ and every $0 \leqslant p \leqslant 1$ by

$$V(\rho, p) = p\left[\mathrm{F}(\rho + \xi(\rho, p)) - \mathrm{F}(\rho - \xi(\rho, p))\right] + 2(1-p)\left(1 - \mathrm{F}(\xi(\rho, p))\right), \tag{9}$$

where F is the cumulative distribution function of the standard normal distribution $\mathcal{N}(0,1)$ and

$$\xi(\rho, p) = \frac{\rho}{2} + \frac{1}{\rho}\left[\ln\frac{1-p}{p} + \ln\left(1 + \sqrt{1 - \frac{p^2}{(1-p)^2}e^{-\rho^2}}\right)\right]. \tag{10}$$

As usual, if a property $\mathcal{P}$ holds true almost surely, we write $\mathcal{P}$ (a-s).

Proposition 1 *Consider the following binary hypothesis testing problem*

$$\begin{cases} \mathcal{H}_0: & U \sim \mathcal{N}(0, \sigma^2) \\ \mathcal{H}_1: & U = S + X,\ S \neq 0 \text{ (a-s)}, |S| \geqslant \mathsf{a} \geqslant 0 \text{ (a-s)}, X \sim \mathcal{N}(0, \sigma^2), \end{cases}$$

where U, S, X are real random variables such that S and X are independent.

If the a priori *probability of occurrence of hypothesis $\mathcal{H}_1$ is less than or equal to some value $\mathrm{p}^* \leqslant 1/2$, then $V(\mathsf{a}/\sigma, \mathrm{p}^*)$ is a sharp upper bound for the probabilities of error of the Bayes test $\mathcal{L}$ with the least probability of error among all possible tests and the thresholding test $\mathcal{T}_{\sigma\xi(\mathsf{a}/\sigma, \mathrm{p}^*)}$ with threshold height $\sigma\xi(\mathsf{a}/\sigma, \mathrm{p}^*)$. The bound $V(\mathsf{a}/\sigma, \mathrm{p}^*)$ is sharp because attained by both $\mathcal{L}$ and $\mathcal{T}_{\sigma\xi(\mathsf{a}/\sigma, \mathrm{p}^*)}$ if $|S| = \mathsf{a}$ (a-s), with $\mathrm{P}[S = \mathsf{a}] = \mathrm{P}[S = -\mathsf{a}] = 1/2$ and the probability of occurrence of hypothesis $\mathcal{H}_1$ is p^*.*

Proof: [See appendix]. ∎

In the foregoing result, if random variables are replaced by n-dimensional real random vectors and the absolute values by the standard Euclidean norm in $\mathbb{R}^n$, the statement thus obtained still holds true, turns out to be an extension of [16, Theorem VII.1] and can be established by mimicking the proof of [16, Theorem VII.1]. Proposition 1 could therefore be considered as a straightforward corollary of this extension. However, for self-completeness of the present paper, we prefer to prove proposition 1 without resorting to [16, Theorem VII.1] and the somewhat sophisticated material of its proof. In fact, dealing with random variables instead of n-dimensional random vectors significantly eases the task.

If the distribution of the signal to recover were known, which is rarely the case, $\mathcal{L}$ could be used to decide whether observed data contain significant information or not, and this decision would be optimal in the sense that $\mathcal{L}$ yields the least possible probability of error. In the non-parametric and frequent practical case where the probability distribution of S is unknown or cannot be estimated accurately enough, then $\mathcal{L}$ is not workable; but, according to the foregoing proposition, we can apply $\mathcal{T}_{\sigma\xi(\rho/\sigma, \mathrm{p}^*)}$, which guarantees the same sharp upper bound for the probability of error as $\mathcal{L}$. Therefore, when the transform $\mathcal{W}$ is sparse in the sense specified by assumptions (**F**) and (**L**), it follows from proposition 1 and Eq. (10) with $p = 1/2$ that the thresholding test $\mathcal{T}_{\lambda_d(N)}$ with

$$\begin{aligned} \lambda_d(N) &= \sigma\xi(\lambda_u(N)/\sigma) \\ &= \sigma\sqrt{\ln N/2} + \sigma\ln\left(1 + \sqrt{1 - 1/N^2}\right)/\sqrt{2\ln N}, \end{aligned} \tag{11}$$

accepts or rejects the null hypothesis with a probability of error less than or equal to $V(\sqrt{2\ln N})$, which is a decreasing function of N. Table 1 gives the values of $V(\sqrt{2\ln N})$ for some values of N.

Table 1: Upper bound $V(\sqrt{2\ln N})$ of the probability of error of the thresholding test $\mathcal{T}_{\lambda_d(N)}$.

N	2	4	8	16	32
$V(\sqrt{2\ln N})$	0.3645	0.2743	0.2110	0.1648	0.1302
N	64	128	512	1024	2048
$V(\sqrt{2\ln N})$	0.1036	0.0830	0.0540	0.0437	0.0356
N	4096	8192	16384	32768	65536
$V(\sqrt{2\ln N})$	0.0290	0.0236	0.0193	0.0158	0.0130

The threshold $\lambda_d(N)$ is henceforth called the *detection threshold*. It is easy to see that the detection threshold $\lambda_d(N)$ is close to $\lambda_u(N)/2$ when N is large enough. Table 2 gives the values of $\lambda_d(N)$, $\lambda_m(N)$, and $\lambda_u(N)$ for some values of N. It shows that for small values of N, the threshold $\lambda_d(N)$ is close to the minimax threshold, and for large values of N (above or equal to 2048), the value of the threshold $\lambda_d(N)$ is, as mentioned above, about $\lambda_u(N)/2$. The threshold $\lambda_d(N)$ is smaller than the minimax threshold and almost two times smaller than the universal threshold when N is large.

Table 2: Detection, minimax, and universal thresholds for different values of the sample size N.

N	128	256	512	1024	2048	4096	8192
$\lambda_d(N)$	1.78	1.87	1.96	2.05	2.13	2.21	2.29
$\lambda_m(N)$	1.67	1.86	2.05	2.23	2.40	2.58	2.74
$\lambda_u(N)$	3.12	3.33	3.53	3.72	3.91	4.08	4.25

Under assumptions (**F**) and (**L**), the detection threshold is appropriate for deciding whether a coefficient returned by the sparse transform $\mathcal{W}$ pertains to the signal or not. However, assumptions (**F**) and (**L**) may not be satisfied in practice, especially if the signal is not smooth enough or not sufficiently regular. Hence, we now address the performance of the estimation by sparse transform and soft thresholding when the detection threshold is used, without assuming that the transform is sparse in the sense of assumptions (**F**) and (**L**). In this respect, proposition 2 below gives a bound on the risk for the estimation of $\boldsymbol{\theta}$ when the estimate is performed by using $\delta_{\lambda_d(N)}(\cdot)$. In fact, proposition 2 below relies on the following result, which is an easy extension of [7, Theorem 1] (see also [13, Theorem 10.4]) about the risk of the soft thresholding estimation. The extension is that the subsequent result holds true for any positive real value λ and not only for the universal threshold.

Lemma 1 *Given the model described by Eq.* (1)*, consider the estimation of* $\boldsymbol{\theta}$ *by soft thresholding where the threshold is any positive real value* λ*.*

The risk r_λ *of this estimation is such that*

$$r_\lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \leqslant (1 + \lambda^2/\sigma^2) \times \left(\sigma^2 e^{-\lambda^2/2\sigma^2} + r_0(\boldsymbol{\theta})\right). \tag{12}$$

Proof: [See appendix]. ∎

The foregoing result actually extends [7, Theorem 1] since, by putting $\lambda = \lambda_u(N)$ in Eq. (12), we obtain Eq. (5) again.

Proposition 2 *With respect to the model described by Eq.* (1)*, assume that* $N \geqslant 2$ *and consider the estimation of* $\boldsymbol{\theta}$ *by soft thresholding with threshold value* $\lambda_d(N)$*. The risk* $r_{\lambda_d(N)}$ *of this estimation satisfies the inequality*

$$r_{\lambda_d(N)}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \leqslant (\ln N/2 + \eta(N)) \left(\sigma^2 \zeta(N) + r_0(\boldsymbol{\theta})\right), \tag{13}$$

with

$$\eta(N) = 1 + \ln\left(1 + \sqrt{1 - 1/N^2}\right) + \ln^2\left(1 + \sqrt{1 - 1/N^2}\right)/2\ln N, \tag{14}$$

and

$$\zeta(N) = N^{-1/4}\left(1 + \sqrt{1 - 1/N^2}\right)^{-1/2} \times e^{-\ln^2\left(1+\sqrt{1-1/N^2}\right)/4\ln N} \tag{15}$$

Proof: It suffices to inject the value of the detection threshold given by Eq. (11) into Eq. (12) to obtain Eqs. (13), (14) and (15). ∎

Although the detection threshold derives from the binary hypothesis testing problem associated with the sparsity model described by hypotheses (**F**) and (**L**), proposition 2 is established without resorting whatsoever to these hypotheses or, more generally, to any sparsity model. Proposition 2 is, thus, very general.

For a specific class of signals, the upper bound provided by proposition 2 is asymptotically smaller than $2r_0 \ln N(\boldsymbol{\theta})$. Consider the subset

$$\boldsymbol{\Theta}_N = \left\{\boldsymbol{\theta} = \{\theta_i\}_{1 \leqslant i \leqslant N} \in \mathbb{R}^N : r_0(\boldsymbol{\theta}) \geqslant \sigma^2 \zeta(N)\right\}$$

of $\mathbb{R}^N$. The elements of this subset are sequences of coefficients returned by transform $\mathcal{W}$ for a certain class of signals. This is why, with some slight abuse of language, the subset $\boldsymbol{\Theta}_N$ will hereafter be regarded as a class of signals.

Clearly, if $\boldsymbol{\theta}$ belongs to this class, the upper bound given by equation (13) for the risk of the soft thresholding estimation with detection threshold behaves as $r_0(\boldsymbol{\theta}) \ln N$ when N tends to ∞. This follows straightforwardly from the fact that $\lim_{N\to\infty} \eta(N) = 1 + \ln 2$ and that $\sigma^2\zeta(N) + r_0(\boldsymbol{\theta}) \leqslant 2r_0(\boldsymbol{\theta})$ when $\boldsymbol{\theta} \in \boldsymbol{\Theta}_N$.

Moreover, for any element $\boldsymbol{\theta}$ of $\boldsymbol{\Theta}_N$ such that $r_0(\boldsymbol{\theta}) \gg \sigma^2\zeta(N)$, the order for the upper bound on the risk $r_{\lambda_d(N)}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$ is now $r_0(\boldsymbol{\theta}) \ln N/2$ when N tends to ∞. Indeed, from $\lim_{N\to\infty} \zeta(N) = 0$, it follows that if $r_0(\boldsymbol{\theta})$ is very large in comparison with $\sigma^2\zeta(N)$, then $\sigma^2\zeta(N) + r_0(\boldsymbol{\theta}) \sim r_0(\boldsymbol{\theta})$ for sufficiently large values of N.

On the other hand, for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}_N$, the upper bound in Eq. (5) (resp. Eq. (7)) for the risk of the soft thresholding estimation with universal threshold (resp. minimax threshold) behaves as $2r_0(\boldsymbol{\theta}) \ln N$ when N increases to ∞.

Therefore, to estimate an element of $\boldsymbol{\Theta}_N$ by soft thresholding when N is large, the detection threshold leads to an order for the upper bound of the estimation risk two to four times smaller than the order obtained when either the universal or the minimax threshold is used.

These results do not contradict [7, Theorem 3], which states that $2r_0(\boldsymbol{\theta}) \ln N$ is the optimal order for the upper bound of the estimation risk when diagonal estimators such as soft thresholding are used. There is no contradiction because our discussion concerns the subset $\boldsymbol{\Theta}_N$ of $\mathbb{R}^N$, whereas [7, Theorem 3] holds true over $\mathbb{R}^N$.

At this stage, it is worth wondering whether $\boldsymbol{\Theta}_N$ is not too small a class and what kind of signals this class can be expected to contain. On the one hand, if α stands for the proportion of coefficients whose amplitude is larger than or equal to σ, it follows from

$$\sum_{i=1}^{N} \min\left(\theta_i^2, \sigma^2\right) = \alpha N \sigma^2 + \sum_{|\theta_i|<\sigma} \theta_i^2, \tag{16}$$

that $r_0(\boldsymbol{\theta}) \geqslant \alpha\sigma^2$. Therefore, any signal such that $\alpha \geqslant \zeta(N)$ belongs to $\boldsymbol{\Theta}_N$. In this respect, $\boldsymbol{\Theta}_N$ must contain piecewise regular signals. In fact, a singularity creates approximately the same number of large coefficients at each resolution level, whereas the number of wavelet coefficients at resolution level $j \geqslant 1$ decreases when j increases [13, p. 460]. On the other hand, let us assume that the function to recover belongs to some Besov class $B^s_{p,q}$, $p, q \geq 1$, $sp > 1$. According to [5, 8, 11], the oracle risk $r_0(\boldsymbol{\theta}) = O(N^{-2s/(2s+1)})$ whereas $\zeta(N) \sim N^{-1/4}$ when N tends to infinity. This suggests that most elements of $\boldsymbol{\Theta}_N$ belong to Besov classes $B^s_{p,q}$ such that $s < 1/6$. The elements of these Besov classes tend to be non-smooth functions since, roughly speaking, s indicates the 'number' of derivatives of the elements of $B^s_{p,q}$. The condition $r_0(\boldsymbol{\theta}) \geqslant \sigma^2\zeta(N)$ is however not as restrictive as the foregoing could suggest. Since real-world signals and images are often non-smooth and rather piecewise regular, $\boldsymbol{\Theta}_N$ can in fact be expected to contain the wavelet representations of many of these signals. This is confirmed by the experimental results of the next section: briefly, $r_0(\boldsymbol{\theta})$ is actually much larger than $\sigma^2\zeta(N)$ for every natural image considered below; half of the synthetic signals given in the `WaveLab` toolbox are elements of $\boldsymbol{\Theta}_N$. Note also that the 'Blocks' signal is an example of a piecewise regular signal which is an element of $\boldsymbol{\Theta}_N$ for reasonable values of the noise standard deviation σ.

We can summarize the discussion above as follows. The minimax and universal thresholds are suitable for recovering smooth signals, whereas the detection threshold is suitable for estimating less smooth signals, including piecewise regular signals, which are known to be over-smoothed when using the minimax or the universal threshold. For instance, smooth signals yield very sparse wavelet representations in the sense given by [7]: for such signals, large coefficients are very few in number. In contrast, wavelet representations of natural images, which are piecewise regular rather than smooth, fail to be sparse enough since large coefficients are not very few. This justifies the introduction of assumption **(F)**, which makes it possible to derive thresholds adapted to less smooth signals.

However, note that the null signal does not belong to $\boldsymbol{\Theta}_N$. To estimate $\boldsymbol{\theta} = 0$, the larger the threshold, the smaller the risk. Therefore, the detection and minimax thresholds are less suitable for estimating the null function than the universal threshold because the two former are smaller than the latter. In fact, when $\boldsymbol{\theta} = 0$, the risk is zero when the threshold is infinitely large. This is coherent

with proposition 1: if only the null hypothesis actually occurs, the probability of occurrence of the alternative hypothesis is 0; it then suffices to set $p = \mathrm{p}^* = 0$ in equation Eq. (10) to derive that the threshold to use in this case is actually $\xi(\mathsf{a}, 0) = \infty$.

The foregoing discussion about class $\mathbf{\Theta}_N$ suggests that the universal and the minimax thresholds are actually too large for many practical applications, as already reported by several authors (see [3, 13] among others). It also suggests that, when the sample size N is large enough, the detection threshold should perform better than the universal and minimax thresholds for the estimation by soft thresholding of many signals and images of practical interest. This is what we experimentally verify in the next section. In fact, the experimental results of the following section confirm that the wavelet representations of many standard signals and images actually belong to class $\mathbf{\Theta}_N$, that the detection threshold performs better than the universal and the minimax thresholds for most of the signals and images tested and that the detection threshold achieves better results than the universal and minimax thresholds, even when the signal representation $\boldsymbol{\theta}$ does not belong to $\mathbf{\Theta}_N$.

4 Experimental results

The previous section suggests using the detection threshold instead of the universal and minimax thresholds for the estimation by soft thresholding of many signals. We now verify experimentally that for a large class of synthetic signals and standard images, the detection threshold makes it possible to achieve smaller risks for the estimation by soft thresholding than the universal and the minimax thresholds.

4.1 Risk evaluation on synthetic signals

In the experiments whose results are presented below, the transform represented by the orthonormal matrix $\mathcal{W}$ is the DWT based on the Symlet wavelet of order 8 ('sym8' in the Matlab Wavelet toolbox).

The synthetic signals considered in this section are generated from the WaveLab toolbox [1]. As in [7], the sample size is $N = 2048$ for every signal tested. We choose $\sigma = 1$ and the signals are rescaled for every Signal-to-Noise Ratio (SNR) tested. The SNRs tested are $1, 3, 5$ and 7. Soft thresholding is applied to the detail coefficients of the decomposition levels $j = 1, 2, \ldots, J$ where J is either 6 or 10. The signals under consideration have different sparsity degree according to their wavelet representations. This can be seen, for instance, in figure 1, which gives the DWT representations of the 'Blocks', 'Doppler', 'Cusp', and 'HypChirps' signals.

For every signal tested, table 3 gives the average risk computed over 25 noise realizations, for SNR $= 1, 7$ and when $J = 6, 10$. Experiments of the same type were carried out for SNR $= 3, 5$, with the same signals and the same decomposition levels. The results thus obtained are very similar: at a given SNR and for most of the signals tested, the smallest risk is achieved with the detection threshold. As far as the four signals 'Blocks', 'Doppler', 'Cusp', and 'HypChirps' are concerned, this holds true except for the 'Cusp' signal (see figure 2). In fact, depending on the SNR and the maximum

[1] available at `http://www-stat.stanford.edu/~wavelab/`

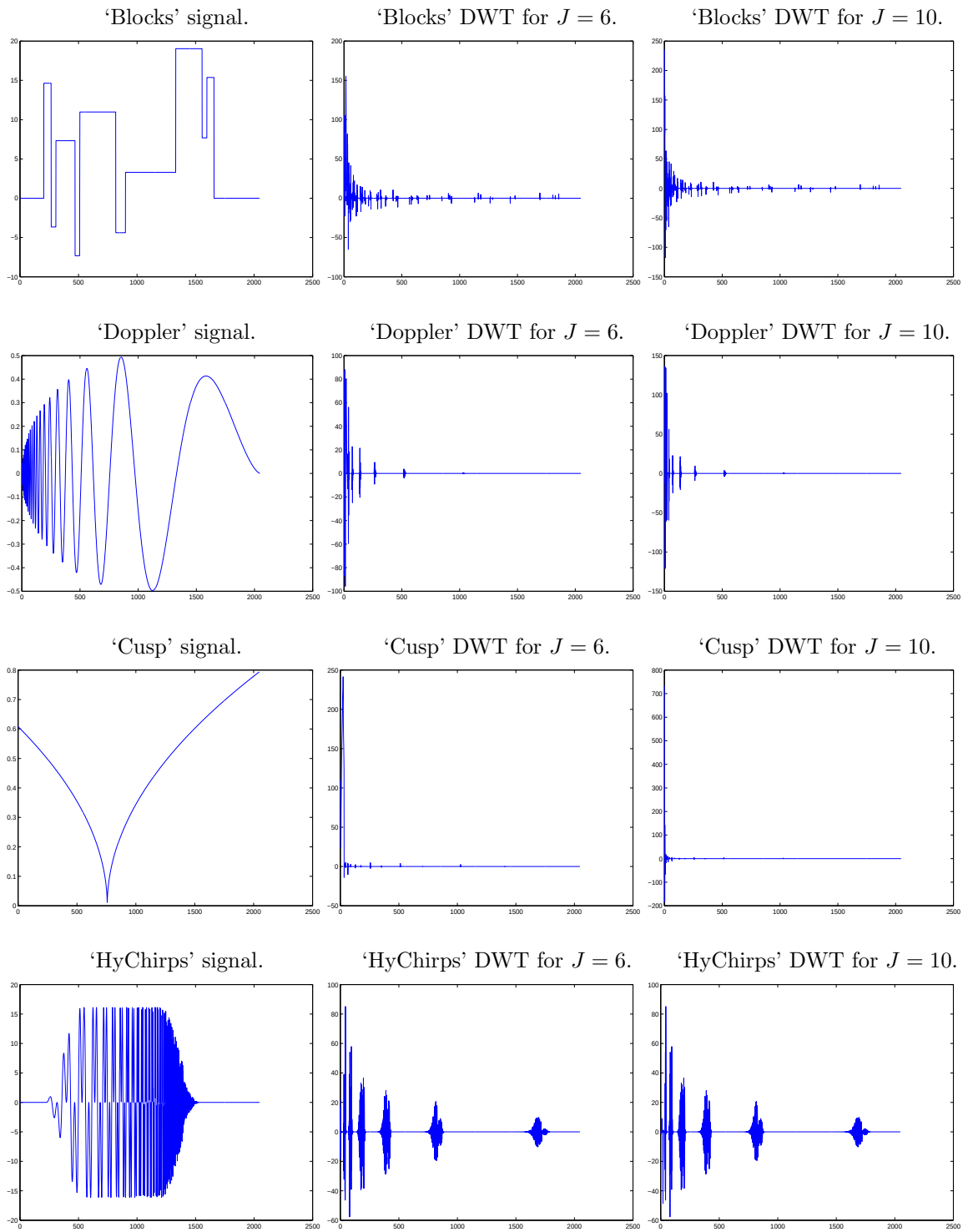


Figure 1: Examples of signals tested, with their DWT representations. The DWT concerns the resolution levels $j = 1, 2, \ldots, J$ where J is either 6 or 10.

decomposition level J, the minimax and universal thresholds outperform the detection threshold for estimating this signal. This is not very surprising for the following reason.

For every given SNR $= 1, 3, 5, 7$, about half of the signals under consideration are in fact elements of $\mathbf{\Theta}_N$. For instance, in table 3, signals whose names are written in boldface belong to this class. In particular, as far as the four signals considered in figure 1 are concerned, 'Blocks' belongs to this class only for SNR $= 5, 7$, 'Doppler' and 'Cusp' are not elements of this class, and 'HyChirps' belongs to this class for every SNR tested.

From the experimental results of this section, we can conclude that for every signal tested that belongs to $\mathbf{\Theta}_N$, the detection threshold performs better than the universal and the minimax thresholds. In addition, the detection threshold generally performs better even when the signal is not an element of $\mathbf{\Theta}_N$. As mentioned above, an exception occurs for the 'Cusp' signal, for which the universal and minimax thresholds lead to smaller risks.

4.2 Risk evaluation on standard images

We consider the standard images 'House' and 'Peppers' with size 256×256 as well as the usual 'Barbara', 'Lena', 'Finger', and 'Boat' images with size 512×512. These images are decomposed via the standard two-dimensional DWT. As in section 4.1, we use the 'sym8' wavelet for the decomposition. The decomposition levels are $j = 1, 2, \ldots, J$ where J is now chosen equal to 4. The DWT representations of the images under consideration are given in figure 3. These DWT representations are not very sparse in the sense given in [7]. In fact, most real-world signals and images are non-smooth. However, if $\boldsymbol{\theta}$ represents the coefficients returned by the DWT for a given image, it turns out (see table 4) that $r_0(\boldsymbol{\theta}) \geqslant \sigma^2\zeta(N)$ so that $\boldsymbol{\theta} \in \mathbf{\Theta}_N$ for every image mentioned above and every tested standard deviation value $\sigma = 9, 18, 27, 36$. This is consistent with the discussion, in section 3, about the signals pertaining to $\mathbf{\Theta}_N$. Indeed, the images considered in the present section, as well as most images encountered in practice, are non-smooth in the sense that they present many singularities due, for instance, to contours and texture. Note that $r_0(\boldsymbol{\theta})$ is generally much larger than $\sigma^2\zeta(N)$.

Table 5 presents the risks obtained with detection, minimax and universal thresholds, when estimation by soft thresholding is applied to the detail coefficients obtained at decomposition levels $j = 1, 2, \ldots, J$ so as to denoise the tested images additively corrupted by independent WGN. For every tested $\sigma = 9, 18, 27$, and 36, and for every threshold, the risk given in table 5 is the average value obtained over 10 trials. As can be seen in table 5, the risks obtained by using the detection threshold are smaller than those achieved with the minimax and universal thresholds. This confirms that the detection threshold performs better for elements of $\mathbf{\Theta}_N$.

4.3 Denoising by using the stationary wavelet transform

The transform is now the Stationary Wavelet Transform (SWT). This transform is particularly suitable for denoising because it is translation invariant and redundant [4, 13]. The 'sym8' wavelet was again used to perform the SWT. Soft thresholding is applied to the detail coefficients at decomposition levels $j = 1, 2, \ldots, J$ where $J = 6$ for signals and $J = 4$ for images.

Table 3: Risks r_λ for detection, minimax, and universal thresholds. Soft thresholding is applied to the detail coefficients at decomposition levels $j = 1, 2, \ldots, J$ where J is either 6 or 10. Signals with names in boldface are elements of $\mathbf{\Theta}_N$ for the SNRs tested. The value of $\zeta(N)$ is 0.1035 since $N = 2048$.

SNR= 1

$J = 6$					$J = 10$				
Signal	HeaviSine	Bumps	Blocks	Doppler	Signal	HeaviSine	Bumps	Blocks	Doppler
$r_{\lambda_d(N)}$	0.0270	0.1758	0.0988	0.0661	$r_{\lambda_d(N)}$	0.0423	0.2275	0.1408	0.0967
$r_{\lambda_m(N)}$	0.0227	0.1950	0.1027	0.0677	$r_{\lambda_m(N)}$	0.0439	0.2575	0.1558	0.1065
$r_{\lambda_u(N)}$	0.0196	0.3239	0.1386	0.0924	$r_{\lambda_u(N)}$	0.0801	0.4461	0.2624	0.1872
Signal	Ramp	Cusp	Sing	**HiSine**	Signal	Ramp	Cusp	Sing	**HiSine**
$r_{\lambda_d(N)}$	0.0383	0.0231	0.07	0.8421	$r_{\lambda_d(N)}$	0.0506	0.0310	0.0794	0.8324
$r_{\lambda_m(N)}$	0.0357	0.0187	0.0743	0.8963	$r_{\lambda_m(N)}$	0.0530	0.0301	0.0873	0.8875
$r_{\lambda_u(N)}$	0.0404	0.0157	0.1272	1.0090	$r_{\lambda_u(N)}$	0.0917	0.0456	0.1645	1.0059
Signal	**LoSine**	**LinChirp**	Piece-Poly	**QuadChirp**	Signal	**LoSine**	**LinChirp**	Piece-Poly	**QuadChirp**
$r_{\lambda_d(N)}$	0.7009	0.7320	0.0727	0.5950	$r_{\lambda_d(N)}$	0.6930	0.7273	0.1213	0.5977
$r_{\lambda_m(N)}$	0.7769	0.7877	0.0730	0.6483	$r_{\lambda_m(N)}$	0.7692	0.7830	0.1332	0.6537
$r_{\lambda_u(N)}$	0.9941	0.9455	0.09	0.8296	$r_{\lambda_u(N)}$	0.9879	0.9472	0.2241	0.8530
Signal	**MishMash**	**Werner**	Leopold	Piece-Reg	Signal	**MishMash**	**Werner**	Leopold	Piece-Reg
$r_{\lambda_d(N)}$	0.7462	0.7247	0.0594	0.0672	$r_{\lambda_d(N)}$	0.7465	0.7466	0.0567	0.1158
$r_{\lambda_m(N)}$	0.7912	0.7670	0.0620	0.0672	$r_{\lambda_m(N)}$	0.7963	0.7950	0.06	0.1286
$r_{\lambda_u(N)}$	0.9128	0.8758	0.1024	0.0833	$r_{\lambda_u(N)}$	0.9392	0.9310	0.1029	0.2262
Signal	**Riemann**	**HypChirps**	sineoverx	**Chirps**	Signal	**Riemann**	**HypChirps**	sineoverx	**Chirps**
$r_{\lambda_d(N)}$	0.2744	0.3535	0.0994	0.6976	$r_{\lambda_d(N)}$	0.3274	0.3525	0.1252	0.6956
$r_{\lambda_m(N)}$	0.2883	0.3978	0.1077	0.7524	$r_{\lambda_m(N)}$	0.3558	0.3978	0.1401	0.7546
$r_{\lambda_u(N)}$	0.3477	0.6320	0.1505	0.9290	$r_{\lambda_u(N)}$	0.5048	0.6492	0.2291	0.9428

SNR= 7

$J = 6$					$J = 10$				
Signal	HeaviSine	**Bumps**	**Blocks**	Doppler	Signal	HeaviSine	**Bumps**	**Blocks**	Doppler
$r_{\lambda_d(N)}$	0.0693	0.4593	0.3644	0.1519	$r_{\lambda_d(N)}$	0.1140	0.5373	0.4420	0.1944
$r_{\lambda_m(N)}$	0.0716	0.5390	0.4236	0.1716	$r_{\lambda_m(N)}$	0.1291	0.6354	0.5208	0.2268
$r_{\lambda_u(N)}$	0.1032	1.0882	0.8331	0.3284	$r_{\lambda_u(N)}$	0.2475	1.3096	1.0692	0.4715
Signal	Ramp	Cusp	Sing	**HiSine**	Signal	Ramp	Cusp	Sing	**HiSine**
$r_{\lambda_d(N)}$	0.0726	0.0357	0.1185	2.8721	$r_{\lambda_d(N)}$	0.1084	0.0629	0.1435	2.8724
$r_{\lambda_m(N)}$	0.0773	0.0334	0.1339	3.4361	$r_{\lambda_m(N)}$	0.1237	0.0685	0.1674	3.4382
$r_{\lambda_u(N)}$	0.1353	0.0434	0.2659	7.4357	$r_{\lambda_u(N)}$	0.2539	0.1357	0.3567	7.4345
Signal	**LoSine**	**LinChirp**	Piece-Poly	**QuadChirp**	Signal	**LoSine**	**LinChirp**	Piece-Poly	**QuadChirp**
$r_{\lambda_d(N)}$	2.2457	2.4482	0.3114	1.8587	$r_{\lambda_d(N)}$	2.2193	2.4580	0.3777	1.8845
$r_{\lambda_m(N)}$	2.5826	2.9456	0.3592	2.2344	$r_{\lambda_m(N)}$	2.5575	2.9591	0.4446	2.2680
$r_{\lambda_u(N)}$	4.5711	6.4238	0.6736	4.8663	$r_{\lambda_u(N)}$	4.5570	6.4559	0.9005	4.9504
Signal	**MishMash**	**Werner**	Leopold	Piece-Reg	Signal	**MishMash**	**Werner**	Leopold	Piece-Reg
$r_{\lambda_d(N)}$	3.6411	3.7349	0.1092	0.2457	$r_{\lambda_d(N)}$	3.6485	3.7975	0.1243	0.3204
$r_{\lambda_m(N)}$	4.3826	4.4775	0.1220	0.2803	$r_{\lambda_m(N)}$	4.3948	4.5593	0.1422	0.3739
$r_{\lambda_u(N)}$	9.4769	9.4820	0.2326	0.5141	$r_{\lambda_u(N)}$	9.5406	9.6995	0.2892	0.7471
Signal	**Riemann**	**HypChirps**	sineoverx	**Chirps**	Signal	**Riemann**	**HypChirps**	sineoverx	**Chirps**
$r_{\lambda_d(N)}$	2.6154	0.9497	0.3157	3.1823	$r_{\lambda_d(N)}$	2.6693	0.9525	0.3574	3.1960
$r_{\lambda_m(N)}$	3.0237	1.1304	0.3756	3.7977	$r_{\lambda_m(N)}$	3.0936	1.1374	0.4289	3.8171
$r_{\lambda_u(N)}$	5.1670	2.3892	0.8042	7.8705	$r_{\lambda_u(N)}$	5.3815	2.4288	0.9404	7.9318

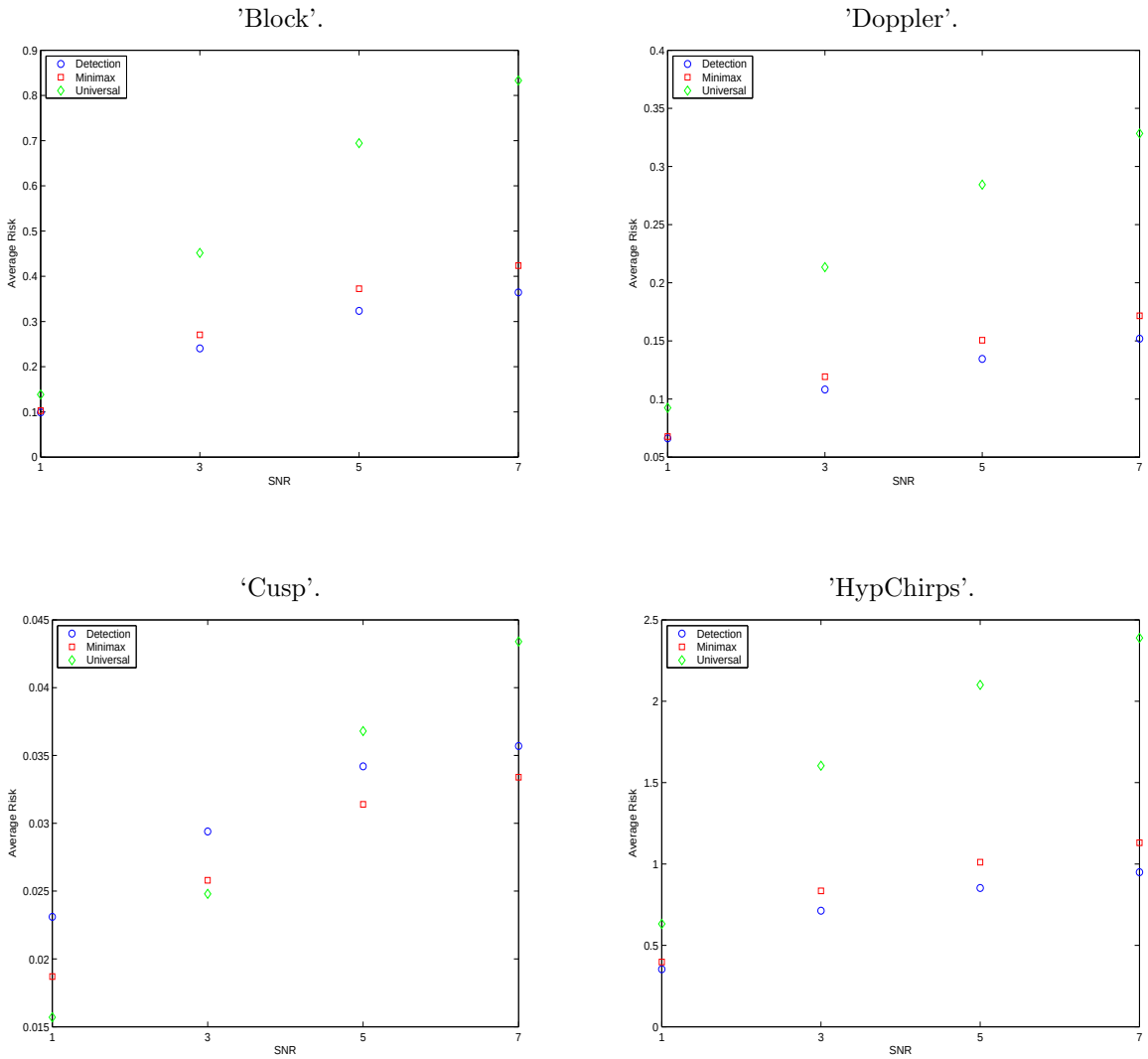


Figure 2: Average risk computed over 25 noise realisations versus the tested SNR $= 1, 3, 5, 7$, for the estimation of the 'Block', 'Doppler', 'Cusp' and 'HypChirps' signals by using soft thresholding with either detection, minimax or universal thresholds. Soft thresholding is applied to the detail coefficents returned by the 'sym8' DWT at decomposition levels $j = 1, 2, \ldots, 6$.

We begin with the standard 'Doppler' signal, additively corrupted by independent WGN with standard deviation $\sigma = 1$ and SNR $= 7$.

Figure 4 shows the noisy 'Doppler' signal in comparison to the three denoised 'Doppler' signals obtained by adjusting the soft thresholding estimation with either the detection, the minimax or the universal threshold. The original 'Doppler' is represented by a dotted line in each of the three figures presenting the denoised signals. In addition, figure 5 zooms on the first 50 and the last 50 coefficients of the several denoised versions of figure 4. These figures show that soft thresholding with the universal threshold achieves a smoother estimate of the original signal than soft thresholding with the minimax

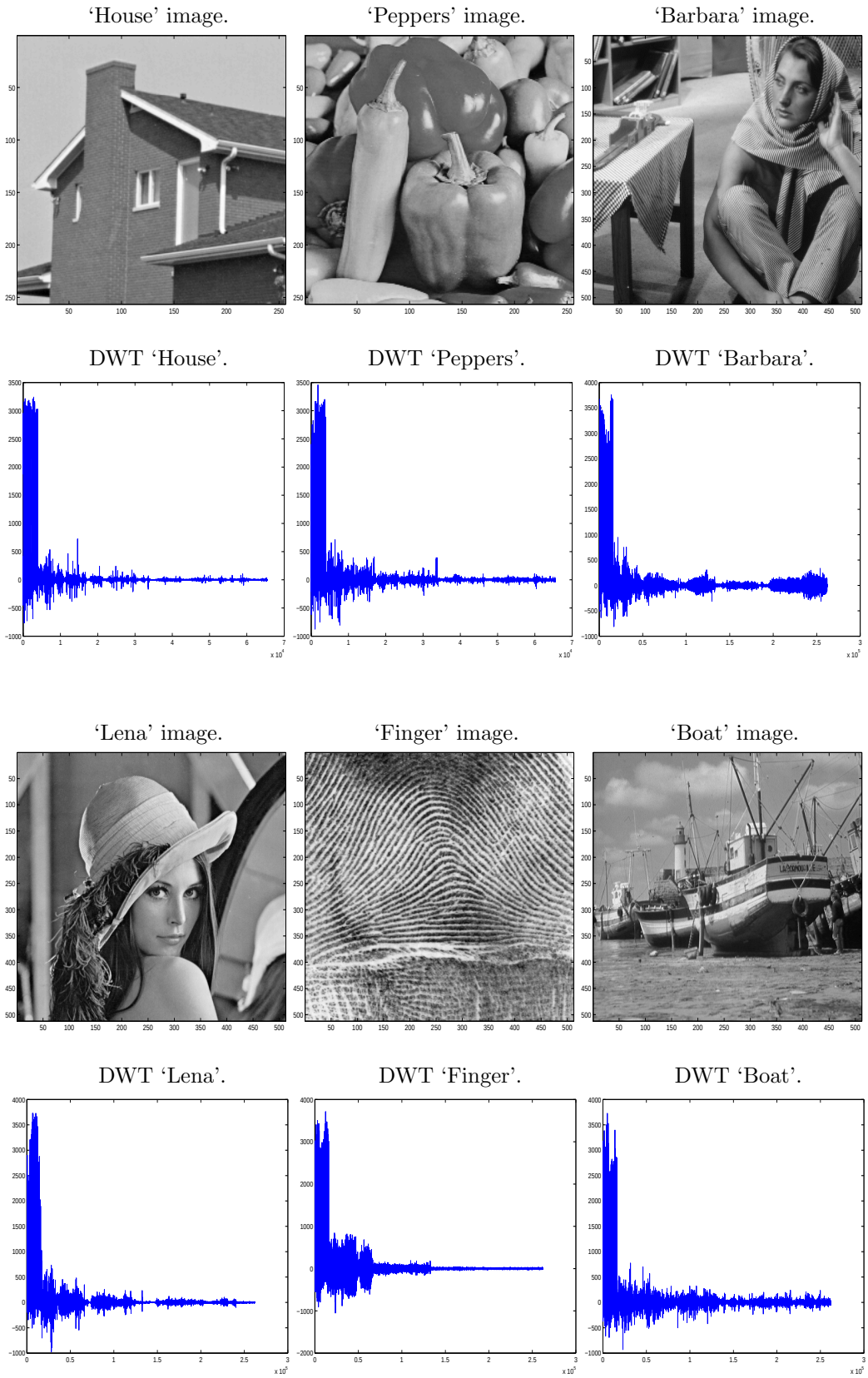


Figure 3: Some standard images, and their 'sym8' DWT representations for $J = 4$.

Table 4: Value of $r_0(\boldsymbol{\theta})$ for every image tested. The vector $\boldsymbol{\theta}$ is the 'sym8' DWT of a given image. The DWT concerns decomposition levels $j = 1, 2, 3, 4$.

N	256×256		512×512			
Images	'House'	'Peppers'	'Barbara'	'Lena'	'Finger'	'Boat'
$\sigma = 9$						
$\sigma^2\zeta(N)$	3.5412	3.5412	2.5070	2.5070	2.5070	2.5070
$r_0(\boldsymbol{\theta})$	17.2998	24.4012	25.0542	17.9539	36.0245	26.5791
$\sigma = 18$						
$\sigma^2\zeta(N)$	14.1647	14.1647	10.0280	10.0280	10.0280	10.0280
$r_0(\boldsymbol{\theta})$	37.9459	56.1343	60.4723	36.4248	87.0434	55.7944
$\sigma = 27$						
$\sigma^2\zeta(N)$	31.8705	31.8705	22.5630	22.5630	22.5630	22.5630
$r_0(\boldsymbol{\theta})$	58.3006	89.8592	98.1831	54.9842	139.9741	84.0878
$\sigma = 36$						
$\sigma^2\zeta(N)$	56.6587	56.6587	40.1120	40.1120	40.1120	40.1120
$r_0(\boldsymbol{\theta})$	78.6185	124.2232	136.0478	73.5691	193.8070	111.5801

or detection thresholds; however, this smooth estimate generally fits the original signal less well than the estimate obtained by using either the detection or the minimax threshold. This oversmoothing obtained with the universal threshold explains why, as illustrated below, images denoised by soft thresholding with the universal threshold are more blurred than images denoised by soft thresholding with the minimax threshold or detection threshold.

Consider now the standard 512×512 'Lena' image additively corrupted by independent WGN. Table 6 presents the risks obtained with the detection, minimax and universal thresholds, when estimation by soft thresholding is used to denoise this image. For every tested noise standard deviation σ and every threshold, each risk given in table 6 is the average value obtained over 10 trials. This table also displays the corresponding Peak Signal-to-Noise Ratio (PSNR), in dB, achieved by the denoising. For a threshold height λ, this PSNR is

$$\mathrm{PSNR}(\lambda) = 10 \log_{10} \left(255^2 / r_\lambda \right). \tag{17}$$

By using the detection threshold, the gain in PSNR is about one to two dB with respect to the PSNRs achieved with the minimax threshold.

An example of 'Lena' image denoising is given in figure 6. The noise standard deviation is $\sigma = 25$. As can be seen, the image denoised by soft thresholding with the detection threshold is sharper than that obtained by soft thresholding with minimax or universal thresholds. Moreover, the contours of the original image are better restored in the image returned by soft thresholding with the detection threshold than in the other two.

Table 5: Risks r_λ obtained with detection, minimax, and universal thresholds for 'Lena' with size 512×512. Soft thresholding is applied to the detail coefficients returnd by the 'sym8' DWT at decomposition levels $j = 1, 2, \ldots, 4$.

N	256×256		512×512			
Images	'House'	'Peppers'	'Barbara'	'Lena'	'Finger'	'Boat'
$\sigma = 9$						
$r_{\lambda_d(N)}$	46.5527	71.4106	89.4982	45.8152	119.5679	74.5194
$r_{\lambda_m(N)}$	60.7369	97.2327	130.4604	62.7722	173.3831	103.3656
$r_{\lambda_u(N)}$	84.3010	141.5590	180.4511	82.9846	240.7068	137.0883
$\sigma = 18$						
$r_{\lambda_d(N)}$	89.3033	150.7902	189.6661	87.2403	255.6930	143.1684
$r_{\lambda_m(N)}$	114.4277	199.3795	257.5655	116.3257	358.1196	189.5936
$r_{\lambda_u(N)}$	153.6769	279.7856	325.7944	149.0234	481.0768	238.7614
$\sigma = 27$						
$r_{\lambda_d(N)}$	128.7033	225.4500	269.5180	124.0678	384.6643	200.0341
$r_{\lambda_m(N)}$	160.9144	294.0141	343.2329	160.7170	529.2084	255.7203
$r_{\lambda_u(N)}$	211.2654	403.9108	404.9466	199.0735	699.8262	311.7397
$\sigma = 36$						
$r_{\lambda_d(N)}$	164.7825	297.3454	328.7292	156.6675	507.7654	247.9232
$r_{\lambda_m(N)}$	202.1105	380.8619	399.0873	197.1790	689.4484	308.4268
$r_{\lambda_u(N)}$	262.7180	508.8113	452.3670	237.0470	898.5787	363.9042

Table 6: Risks r_λ obtained with detection, minimax, and universal thresholds for 'Lena' with size 512×512. Soft thresholding is applied to the detail coefficients of decomposition levels $j = 1, 2, \ldots, 4$.

σ	9	18	27	36
$r_{\lambda_u(N)}$	76.0	141.6	192.1	231.3
$r_{\lambda_m(N)}$	56.2	108.7	152.8	189.3
$r_{\lambda_d(N)}$	40.0	79.4	115.1	146.5
PSNR [initial]	29.0	23.0	19.5	17.0
PSNR[$\lambda_u(N)$]	29.3	26.6	25.3	24.5
PSNR[$\lambda_m(N)$]	30.6	27.8	26.3	25.4
PSNR[$\lambda_d(N)$]	32.1	29.1	27.5	26.5

5 Conclusions and extensions

In this work, the thresholds proposed in [16] have been used for non-parametric estimation by soft thresholding [6, 7]. We have proposed a new threshold, the so-called detection threshold, which is

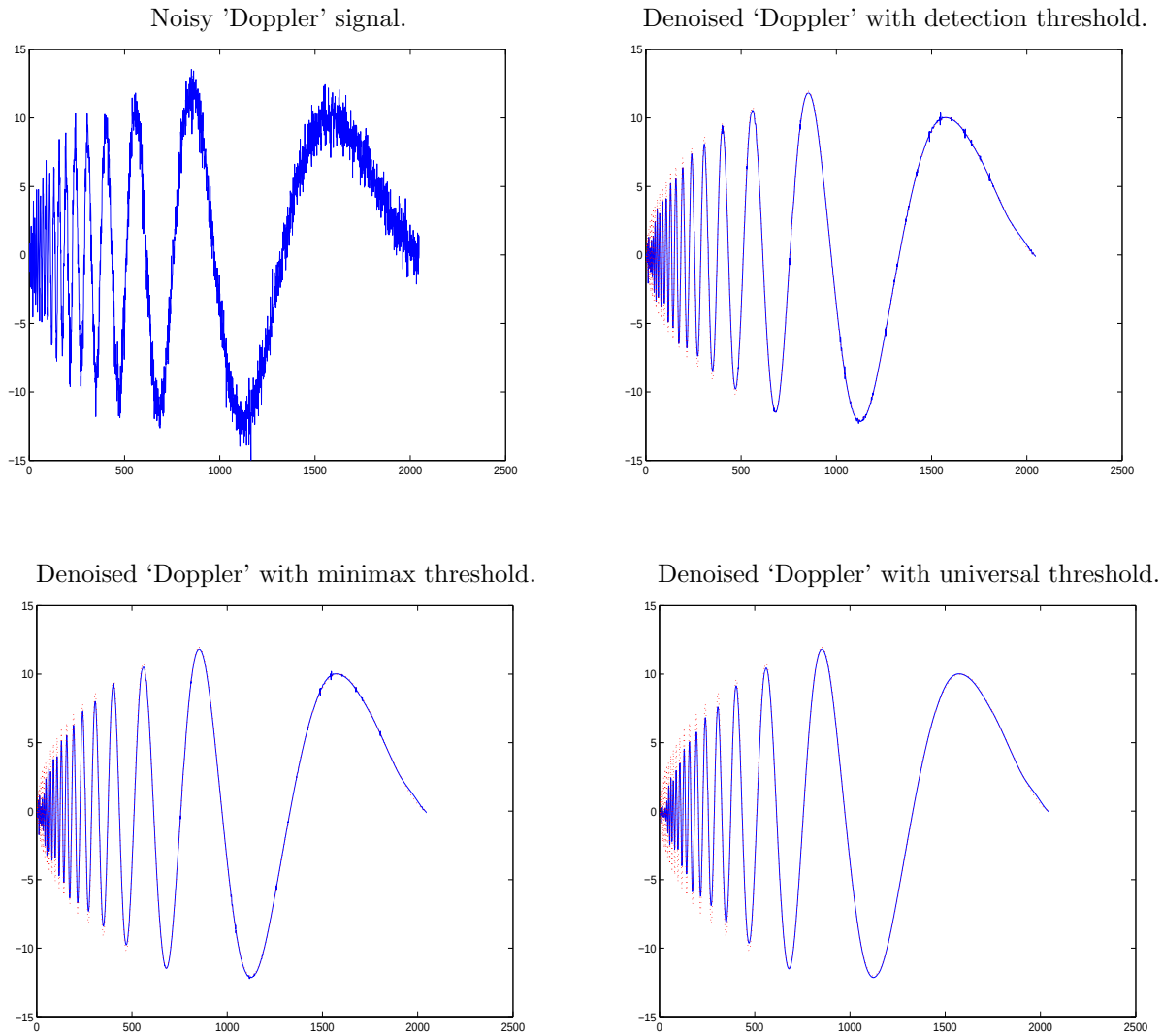
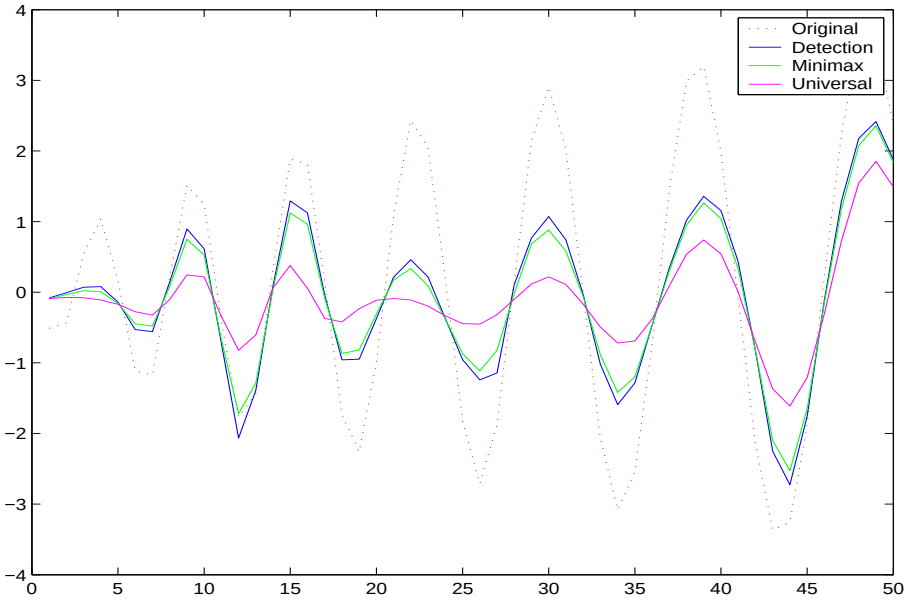


Figure 4: Noisy 'Doppler' signal and denoising of the noisy 'Doppler' signal. The wavelet transform used is a discrete stationary wavelet transform based on the 'sym8' wavelet. The thresholding is applied to the detail coefficients at decomposition levels $j = 1, 2, \ldots, 6$.

relevant to deciding which coefficients, returned by a sparse transform such as the wavelet transform, will be used to estimate the signal. When the sample size N is large, the bound for the risk of the soft thresholding estimation is smaller with the detection threshold than with the minimax or the universal threshold, for a certain class of signals and images. Experiments on standard signals and images show that most of these signals belong to this class and that smaller risks are generally obtained by using the detection threshold instead of the minimax or the universal threshold, even for signals that are not elements of this class. Therefore, for the non-parametric estimation of a signal by sparse transform and soft thresholding, we recommend using the detection threshold instead of the universal and minimax thresholds.

(a) First 50 denoised coefficients.



(b) Last 50 denoised coefficients.

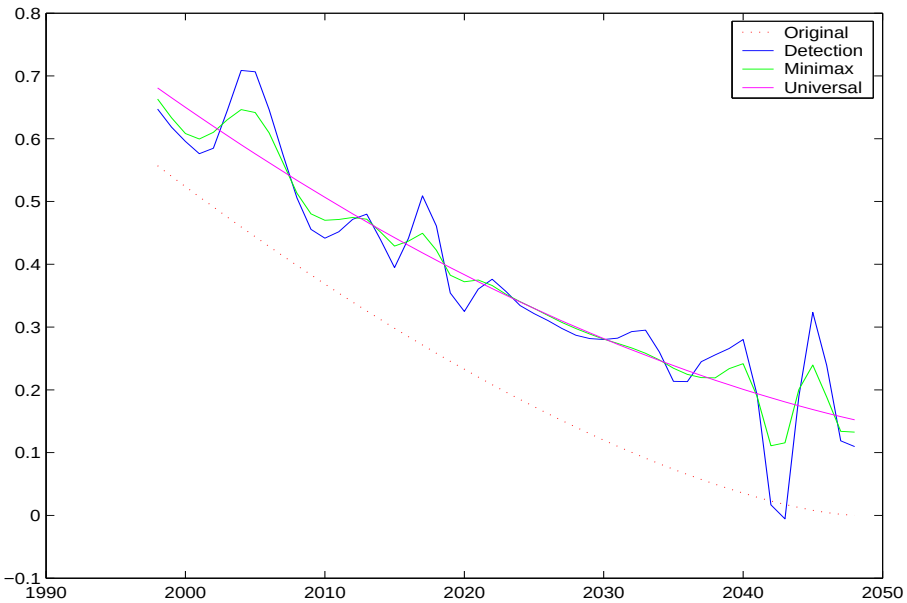


Figure 5: Zooms on (a) the first 50 wavelet coefficients and (b) the last 50 coefficients of the original 'Doppler' signal and its denoised versions via the detection, the minimax and the univeral thresholds.

From a general point of view, the results presented in this paper suggest some extensions in non-parametric estimation, among which are the following.

To begin with, we are interested in studying to what extent the theoretical contents of this paper can be connected with results - such as those stated in [8, 9], among others - about sparsity and Besov spaces. In particular, and as an extension of the discussion following proposition 2 above, such a study

Figure 6: Noisy 'Lena' image and denoised images by soft thresholding with detection, minimax and universal thresholds. The noise standard deviation is $\sigma = 25$. The wavelet transform used is a discrete stationary wavelet transform based on the 'sym8' wavelet. The thresholding is applied to the detail coefficients at decomposition levels $j = 1, 2, 3, 4$.

could refine our knowledge about the class of those signals for which the detection threshold is actually preferable than the minimax or the universal threshold. It is also expected that this study makes it possible to derive detection thresholds that are adapted to the smoothness, in the Besov sense, of the function to recover.

Another possible extension is the following one. The detection threshold of Eq. (11) is derived by bounding our lack of prior knowledge about the signal since we assume that this signal is less present than absent and that this signal is relatively large in the sense that its amplitude exceeds some reasonable value. In fact, consider a dyadic wavelet decomposition based on convolution and downsampling.

It is known that for smooth or piecewise regular signals, the proportion of significant coefficients, which play a role similar to the probability of presence of the signal, increases with the decomposition level [13, Section 10.2.4, p. 460]. Therefore if we can first give an upper-bound $\mathrm{p}_j^* < 1/2$ for the probability of presence of the signal at every given decomposition level $j = 1, 2, \ldots, J$ so that the sequence p_j^* increases with j and, second, a lower-bound a_j for the amplitudes of the wavelet coefficients of the signal, Eq. (10) suggests using the detection threshold $\lambda_d(\mathsf{a}_j, \mathrm{p}_j^*) = \sigma\xi(\mathsf{a}_j/\sigma, \mathrm{p}_j^*)$ for $j = 1, 2, \ldots, J$. By proceeding thus, the detection threshold will be adjusted according to each decomposition level j. This approach will be investigated in further work. The use of detection thresholds adapted to the decomposition levels is expected to yield performance measurements comparable to those obtained with the BLS-GSM introduced in [18] - and considered so far as the best parametric method - and the latest SURE (Stein Unbiased Risk of Estimation) approach, described in [12].

In forthcoming work, we also plan to address the case of an unknown standard deviation. This is a topic of practical interest. According to [7, p. 446] (see also [13, p. 459]), a robust estimate of the noise standard deviation can be computed on the basis of the Median Absolute Deviation (MAD) of the detail wavelet coefficients at the first decomposition level. The robustness of the MAD estimator is due to the fact that the median value is not very affected by a few large coefficients among those used to perform the estimation. However, for non-regular signals or textured images, the detail wavelet coefficients of the first decomposition level may still contain too many coefficients pertaining to the signals and, in such a case, the MAD estimator can fail to achieve a good estimation of the noise standard deviation. It is then interesting to study the behaviour of the estimation by sparse transform and soft thresholding when the detection threshold is adjusted with an estimate of the noise standard deviation derived from the results presented in [14] and [15]. In fact, these papers propose estimators of the noise standard deviation that are computed given non signal-free observations where the signals have unknown probability distributions and are less present than absent in the sense of assumption **(F)**. On the basis of [14] and [15], we expect to propose a new estimator for the estimation of the noise standard deviation, an estimator that remains robust even when coefficients pertaining to the signal are not necessarily few.

6 Acknowledgement

The authors are very grateful to the reviewers for their insightful comments and sound suggestions that helped to significantly improve this paper. In particular, the authors are very thankful to the reviewer who discussed the conditions under which the detection threshold outperforms standard thresholds, suggesting interesting prospects in connection with Besov spaces.

Proof of proposition 1:

For the sake of simplicity, and without loss of generality, we can assume that $\sigma = 1$. We carry out the proof in several steps.

[Step 1]: For $\rho \geqslant 0$, $V(\rho, p)$ is strictly concave for $0 < p < \widetilde{\mathrm{p}}(\rho)$ with $\widetilde{\mathrm{p}}(\rho) = e^{\rho^2/2}/(1 + e^{\rho^2/2})$.

Without resorting to general results such as those given in [17, Chapter II, section C], we can proceed as follows to prove this assertion.

Let $\rho \geqslant 0$. For any p such that $0 < p < \widetilde{\mathrm{p}}(\rho)$, with $\widetilde{\mathrm{p}}(\rho) = e^{\rho^2/2}/(1 + e^{\rho^2/2})$, some algebra shows that $\xi(\rho, p)$ is, in fact, the unique solution in u to the equation

$$\cosh(\rho u) = \frac{1-p}{p} e^{\rho^2/2}. \tag{18}$$

Therefore, for $0 < p < \widetilde{\mathrm{p}}(\rho)$,

$$\frac{\partial V}{\partial p}(\rho, p) = \mathcal{R}(\rho, \xi(\rho, p)) + \mathcal{R}(0, \xi(\rho, p)) - 1, \tag{19}$$

where $\mathcal{R} : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ is the map defined for every $(u, v) \in \mathbb{R}^+ \times \mathbb{R}^+$ by

$$\mathcal{R}(u, v) = \int_{u-v}^{u+v} \Phi(x) dx = \mathrm{F}(u+v) - \mathrm{F}(u-v), \tag{20}$$

with $\Phi(x) = (1/\sqrt{2\pi}) e^{-x^2/2}$.

As a consequence, the sign of $\frac{\partial^2 V}{\partial p^2}(\rho, p)$ for $0 < p < \widetilde{\mathrm{p}}(\rho)$ is exactly that of $\frac{\partial \xi(\rho,p)}{\partial p}$. By differentiating Eq. (18) with respect to $p > 0$, we straightforwardly obtain $\frac{\partial \xi(\rho,p)}{\partial p} < 0$. Hence, $p \longmapsto \xi(\rho, p)$ is decreasing and $p \longmapsto V(\rho, p)$ is strictly concave for $0 < p < \widetilde{\mathrm{p}}(\rho)$.

[Step 2]: The least favourable prior is strictly above $1/2$.

According to **[Step 1]**, there exists only one value $\mathrm{p_L}(\mathsf{a})$ such that $0 < \mathrm{p_L}(\mathsf{a}) < \widetilde{\mathrm{p}}(\mathsf{a})$, the so-called least favourable prior, that maximizes the function $V(\mathsf{a}, \cdot)$. We now establish a strict inequality on the value of this least favourable prior by mimicking the reasoning followed to prove [16, Proposition VI.2]. However, in the monodimensional case, the proof is easier.

Since $\mathrm{p_L}(\mathsf{a})$ is the point where the strictly concave function $p \longmapsto V(\mathsf{a}, p)$ attains its maximum for $0 < p < \widetilde{\mathrm{p}}(\mathsf{a})$ and since $\widetilde{\mathrm{p}}(\mathsf{a})$ strictly exceeds $\frac{1}{2}$, a necessary and sufficient condition for $\mathrm{p_L}(\mathsf{a}) > \frac{1}{2}$ is that

$$\frac{\partial V}{\partial p}(\mathsf{a}, 1/2) > 0. \tag{21}$$

The latter inequality will be a consequence of the following two facts:

(i) $\displaystyle\lim_{\rho \to +\infty} \frac{\partial V}{\partial p}(\rho, 1/2) = 0,$

(ii) $\displaystyle\rho \longmapsto \frac{\partial V}{\partial p}(\rho, 1/2)$ is a decreasing function for $\rho > 0$.

When ρ tends to ∞, the asymptotic behaviour of $\rho \longmapsto \xi(\rho) = \xi(\rho, 1/2)$ is easily seen to be $\xi(\rho) = \frac{\rho}{2} + \frac{\log 2}{\rho}(1 + \delta(\rho))$ with $\lim_{\rho \to +\infty} \delta(\rho) = 0$. Equality **(i)** above then follows from this asymptotic behaviour, the expression of $V(\rho, 1/2)$ and Eq. (20). To establish **(ii)**, we prove that the derivative of $\dfrac{\partial V}{\partial p}(\cdot, 1/2)$ is negative. According to Eqs. (19) and (20), we have

$$\frac{\partial V}{\partial p}(\rho, 1/2) = \int_{-\xi(\rho)-\rho}^{\xi(\rho)-\rho} \Phi(t) dt + \int_{-\xi(\rho)}^{\xi(\rho)} \Phi(t) dt - 1. \tag{22}$$

With some easy algebra and by taking into account Eq. (18), it follows from Eq. (22) that the sign of the derivative of $\frac{\partial V}{\partial p}(\cdot, 1/2)$ is that of $\rho \longmapsto \mathcal{J}(\rho) = 2\xi'(\rho) - \tanh(\rho\xi(\rho))$. By differentiating Eq. (18) to obtain an equation satisfied by the derivative $\xi'(\cdot)$ of $\xi(\cdot)$, taking again Eq. (18) into account and noting that Eq. (18) can be re-written in the form $\log(\cosh(t)) = \rho^2/2$ when p = 1/2, we now obtain that

$$\mathcal{J}(\rho) = \frac{2}{\tanh(\rho\xi(\rho))} - \frac{\rho\xi(\rho)}{\log(\cosh(\rho\xi(\rho)))} - \tanh(\rho\xi(\rho)).$$

To prove that $\frac{\partial V}{\partial p}(\cdot, 1/2)$ is decreasing, it thus suffices to show that the map

$$t \in [0, \infty) \mapsto \frac{2}{\tanh(t)} - \frac{t}{\log(\cosh(t))} - \tanh(t)$$

is negative. Therefore, by setting, similarly to [16], $g(t) = \tanh(t)$ and $G(t) = \log(\cosh(t))/t$, a sufficient condition for **(ii)** to be true is that

$$Q(t) = \frac{g(t)}{G(t)(2 - g(t)^2)} > 1 \tag{23}$$

for $t > 0$. This will be established by showing that $Q(t) > 1$ for positive large (resp. small) values of t and that any stationnary point t_0 of Q is such that $Q(t_0) > 1$.

It is easy to see that

$$Q(t) = \frac{t\sinh(2t)}{(3 + \cosh(2t))\log(\cosh(t))}.$$

It then follows that $Q(t) = 1 + \frac{t^2}{3} + O(t^4)$ when $t \to 0$ and that $Q(t) = 1 + \frac{\log 2}{t} + O(\frac{1}{t^2})$ when $t \to +\infty$. Therefore, for large (resp. small) values of t, $0 \leqslant t < \infty$, we have $Q(t) > 1$.

Consider now a stationary point t_0 of Q, that is a positive real number t_0 such that $Q'(t_0) = 0$. Since we have $G'(t) = \frac{g(t) - G(t)}{t}$, it follows from Eq. (23) that $Q'(t_0) = 0$ implies that

$$G(t_0) = \frac{g(t_0)^2(2 - g(t_0)^2)}{t_0 g'(t_0)(2 - g(t_0)^2) + g(t_0)(2 - g(t_0)^2) + 2g(t_0)^2 g'(t_0)t_0}.$$

Injecting this expression of $G(t_0)$ back into Eq. (23) and taking into account that $g'(t) = 1 - g(t)^2$, we obtain that

$$Q(t_0) = \frac{t_0(1 - g(t_0)^2)(2 + g(t_0)^2) + g(t_0)(2 - g(t_0)^2)}{g(t_0)(2 - g(t_0)^2)^2}.$$

For any $0 \leqslant y < 1$, it follows from [1, Eq. 4.1.33, p. 68] that

$$\frac{1}{2}\log\frac{1+y}{1-y} = \frac{1}{2}\log\left(1 + \frac{2y}{1-y}\right) > \frac{y}{y+1} \geq y\frac{2 - y^2}{2 + y^2}. \tag{24}$$

Since $y \mapsto \frac{1}{2}\log\frac{1+y}{1-y}$ is the inverse map of tanh, it suffices to apply inequality (24) to $y_0 = g(t_0) = \tanh(t_0)$ to obtain that $t_0(2 + g(t_0)^2) - g(t_0)(2 - g(t_0)^2) > 0$, which proves that $Q(t_0) > 1$. Statement **(ii)** above is thus established. As mentioned above, **(i)** and **(ii)** are sufficient to guarantee that inequality (21) holds true and, thus, that $\mathrm{p_L}(\mathsf{a}) > 1/2$.

[Step 3]: The probability of error of the thresholding test with threshold height $\xi(\mathsf{a}, \mathrm{p}^*)$ does not exceed $V(\mathsf{a}, \mathrm{p}^*)$.

The probability of error $\mathrm{P}_e(\mathcal{T}_\lambda)$ of any thresholding test $\mathcal{T}_\lambda$, $\lambda \geqslant 0$, is given by

$$\mathrm{P}_e\,[\,\mathcal{T}_\lambda\,] = \pi_0\,\mathrm{P}\,[\,|X| \geq \lambda\,] \,+\, \pi_1\,\mathrm{P}\,[\,|S+X| \leq \lambda\,], \tag{25}$$

where π_0 (resp. π_1) henceforth stands for the *a priori* probability of occurrence of hypothesis $\mathcal{H}_0$ (resp. hypothesis $\mathcal{H}_1$).

Because Φ is even, we have $\mathrm{P}\,[\,|s+X| \leq \xi\,] = \mathcal{R}(|s|,\xi)$ for every $s \in \mathbb{R}$. Therefore,

$$\mathrm{P}\,[\,|S+X| \leq \lambda\,] = \int \mathcal{R}(|s|,\lambda)\mathrm{P}_S(ds),$$

where P_S denotes the probability distribution of S and $\mathcal{R}$ given by Eq. (20). We also have $\mathrm{P}\,[\,|X| \leq \lambda\,] = \mathcal{R}(0,\lambda)$, and, thus,

$$\mathrm{P}_e\,[\,\mathcal{T}_\lambda\,] = \pi_0(1-\mathcal{R}(0,\lambda)) + \pi_1 \int \mathcal{R}(|s|,\lambda)\mathrm{P}_S(ds). \tag{26}$$

We now set

$$C(s,p,t) = p\,\mathcal{R}(s,t) \,+\, (1-p)\,(1-\mathcal{R}(0,t))\,, \tag{27}$$

with $s,t \geqslant 0$ and $0 \leqslant p \leqslant 1$. We have

$$\mathrm{P}_e\,[\,\mathcal{T}_\lambda\,] = \int C(|s|,\pi_1,\lambda)\mathrm{P}_S(ds). \tag{28}$$

Given any non-negative real number v, $\mathcal{R}(\cdot,v)$ is a non-decreasing function. Therefore, since $|S| \geq \mathsf{a}$ (a-s), the first integral on the right hand side (rhs) of Eq. (26) is less than or equal to $\mathcal{R}(\mathsf{a},\lambda)$ for any $0 \leqslant \lambda < \infty$. It then follows from Eq. (26) with $\lambda = \xi(\mathsf{a},\mathrm{p}^*)$ that

$$\mathrm{P}_e\,[\mathcal{T}_{\xi(\mathsf{a},\mathrm{p}^*)}\,] \leqslant C(\mathsf{a},\pi_1,\xi(\mathsf{a},\mathrm{p}^*)). \tag{29}$$

Since $\pi_0 + \pi_1 = 1$, the rhs in the inequality above can now be written in the form

$$C(\mathsf{a},\pi_1,\xi(\mathsf{a},\mathrm{p}^*)) = 1 - \mathcal{R}(0,\xi(\mathsf{a},\mathrm{p}^*)) + \pi_1\,(\mathcal{R}(\mathsf{a},\xi(\mathsf{a},\mathrm{p}^*)) + \mathcal{R}(0,\xi(\mathsf{a},\mathrm{p}^*)-1)\,. \tag{30}$$

Since $\mathrm{p}^* \leq 1/2$, it follows from **[Step 2]** that $\mathrm{p}^* \leq \mathrm{p}_{\mathrm{L}}(\mathsf{a})$. Now, according to **[Step 1]** and Eq. (19), the coefficient of π_1 on the rhs of Eq. (30) is positive. Taking into account that π_1 is assumed to be less than or equal to p^*, we derive from the foregoing and (30) that

$$C(\mathsf{a},\pi_1,\xi(\mathsf{a},\mathrm{p}^*)) \leqslant 1 - \mathcal{R}(0,\xi(\mathsf{a},\mathrm{p}^*)) + \mathrm{p}^*\,(\mathcal{R}(\mathsf{a},\xi(\mathsf{a},\mathrm{p}^*)) + \mathcal{R}(0,\xi(\mathsf{a},\mathrm{p}^*)-1)\,.$$

According to Eq. (9), the rhs in the inequality above is $V(\mathsf{a},\mathrm{p}^*)$. Since the MPE test $\mathcal{L}$ yields the smallest possible probability of error among all possible tests, we derive from Eq. (29) that

$$\mathrm{P}_e\,[\,\mathcal{L}\,] \leq \mathrm{P}_e\,[\,\mathcal{T}_{\xi(\mathsf{a},\mathrm{p}^*)}\,] \leq V(\mathsf{a},\mathrm{p}^*). \tag{31}$$

[Step 4]: End of the proof. Consider the specific case where $S \in \{-\mathsf{a},\mathsf{a}\}$ with $\mathrm{P}\,[\,S=\mathsf{a}\,] = \mathrm{P}\,[\,S=-\mathsf{a}\,] = 1/2$ and $\pi_1 = \mathrm{p}^*$. We must prove that the inequalities in Eq. (31) above become equalities.

Assume first that $\mathrm{p}^* = 0$. Clearly, the MPE test is then the thresholding test $\mathcal{T}_\infty$ with infinite threshold, that is, the thresholding test $\mathcal{T}_{\xi(\mathsf{a},\infty)}$ since $\xi(\mathsf{a},\infty) = \infty$.

If $\mathrm{p}^* \neq 0$, it now follows from the general form of the MPE test (see [17, 11, Sec. II.B], among others), that the MPE test $\mathcal{L}$ is given by

$$\mathcal{L}(u) = \begin{cases} 1 & \text{if} \quad \cosh(\mathsf{a}u) \geqslant \frac{1-\mathrm{p}^*}{\mathrm{p}^*} e^{\mathsf{a}^2/2}, \\ 0 & \text{if} \quad \cosh(\mathsf{a}u) < \frac{1-\mathrm{p}^*}{\mathrm{p}^*} e^{\mathsf{a}^2/2}. \end{cases} \tag{32}$$

Since $\mathrm{p}^* \leq 1/2 < \mathrm{p_L}(\mathsf{a})$, we derive from Eq. (18) that $\mathcal{L}$ coincides with the thresholding test $\mathcal{T}_{\xi(\mathsf{a},\mathrm{p}^*)}$, which concludes the proof. Note that, if $\mathrm{p}^* = 0$, we again find from Eq. (32) that the MPE test is the thresholding test with infinite threshold.

Proof of lemma 1:

The risk of the estimation by soft thresholding is $r_\lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\Big(\theta_i - \delta_\lambda(c_i)\Big)^2$. Since $\delta_\lambda(x) = \sigma\delta_{\lambda/\sigma}(x/\sigma)$ for every $x \in \mathbb{R}$, this risk can be written

$$r_\lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \frac{\sigma^2}{N}\sum_{i=1}^N \mathbb{E}\Big(\frac{\theta_i}{\sigma} - \delta_{\lambda/\sigma}(\frac{c_i}{\sigma})\Big)^2. \tag{33}$$

An easy extension of [13, lemma 10.1, p. 452] shows that for all $X \sim \mathcal{N}(\mu, 1)$ and for all $t \geqslant 0$

$$\mathbb{E}\Big(\mu - \delta_t(X)\Big)^2 \leqslant 2\int_0^{+\infty} x^2\Phi(x+t)dx + \min(\mu^2, 1+t^2) \tag{34}$$

where, as above, $\Phi(x) = (1/\sqrt{2\pi})e^{-x^2/2}$. For every $i = 1, 2, \ldots, N$, $c_i/\sigma \sim \mathcal{N}(\theta_i/\sigma, 1)$. Thus, according to Eq. (34), we have

$$\mathbb{E}\Big(\frac{\theta_i^2}{\sigma^2} - \delta_t(\frac{c_i}{\sigma})\Big)^2 \leqslant 2\int_0^{+\infty} x^2\Phi(x+t)dx + \min(\frac{\theta_i^2}{\sigma^2}, 1+t^2), \tag{35}$$

with $t = \lambda/\sigma$. From Eq. (33) and Eq. (35), we obtain

$$r_\lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \leq 2\sigma^2\int_0^{+\infty} x^2\Phi(x+t)dx + \frac{\sigma^2}{N}\sum_{i=1}^N \min(\frac{\theta_i^2}{\sigma^2}, 1+t^2). \tag{36}$$

In addition, we have

$$2\int_0^{+\infty} x^2\Phi(x+t)dx \leqslant (1+t^2)e^{-t^2/2}. \tag{37}$$

This inequality is derived as follows. For every $0 \leqslant t < \infty$, put

$$h(t) = \frac{2e^{t^2/2}}{1+t^2}\int_0^{+\infty} x^2\Phi(x+t)dx.$$

Clearly, $h(t) = \frac{2}{\sqrt{2\pi}(1+t^2)}\int_0^{+\infty} x^2 e^{-x^2/2}e^{-xt}dx$ and h is non-increasing. Therefore, $h(t) \leqslant g(0)$. Since

$$h(0) = (2/\sqrt{2\pi})\int_0^{+\infty} x^2 e^{-x^2/2}dx = 1,$$

we obtain $h(t) \leqslant 1$ for all $0 \leqslant t < \infty$, which proves inequality (37). From Eq. (36) and Eq. (37), we derive that

$$r_\lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \leq \sigma^2(1+t^2)e^{-t^2/2} + \frac{\sigma^2}{N}\sum_{i=1}^N \min\left(\theta_i^2/\sigma^2, 1+t^2\right). \tag{38}$$

In addition, $\min\left(\theta_i^2/\sigma^2, 1+t^2\right) \leq (1+t^2)\min\left(\theta_i^2/\sigma^2, 1\right)$ and $\sigma^2 \min\left(\theta_i^2/\sigma^2, 1\right) = \min\left(\theta_i^2, \sigma^2\right)$. Therefore, it follows from Eq. (38) that $r_\lambda(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \leqslant \sigma^2(1+t^2)e^{-t^2/2} + (1+t^2)\sum_{i=1}^N \min(\theta_i^2, \sigma^2)/N$ with $t = \lambda/\sigma$, which completes the proof.

References

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions. Ninth printing.* Dover Publications Inc., New York., 1972.

[2] S. M. Berman. *Sojourns and extremes of stochastic processes.* Wadsworth and Brooks/Cole, 1992.

[3] A. G. Bruce and H. Y. Gao. Understanding waveshrink: Variance and bias estimation. Research Report 36, StatSci, 1996.

[4] R. R. Coifman and D. L. Donoho. *Translation invariant de-noising*, pages 125–150. Number 103. Lecture Notes in Statistics, 1995.

[5] D. L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, 1(1):100–115, 1993.

[6] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, May 1995.

[7] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, Aug. 1994.

[8] D. L. Donoho and I. M. Johnstone. Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli 2*, (1):39–62, 1996.

[9] D. L. Donoho and I. M. Johnstone. Asymptotic minimaxity of wavelet estimators with sampled data. *Statistica Sinica*, 9(1):1–32, 1999.

[10] I. M. Johnstone. Wavelets and the theory of non-parametric function estimation. *Journal of the Royal Statistical Society*, A(357):2475–2493, 1999.

[11] H. Krim, D. Tucker, S. Mallat, and D. L. Donoho. On denoising and best signal representation. *IEEE Transactions on Information Theory*, 45(7):2225+, Nov. 1999.

[12] F. Luisier, T. Blu, and M. Unser. A new sure approach to image denoising: Interscale orthonormal wavelet thresholding. *IEEE Transactions on Image Processing*, 16(3):593–606, Mar. 2007.

[13] S. Mallat. *A wavelet tour of signal processing, second edition.* Academic Press, 1999.

[14] D. Pastor. A theoretical result for processing signals that have unknown distributions and priors in white gaussian noise. To appear in Computational Statistics & Data Analysis, CSDA, `http://dx.doi.org/10.1016/j.csda.2007.10.011`.

[15] D. Pastor and A. Amehraye. Algorithms and applications for estimating the standard deviation of awgn when observations are not signal-free. *Journal of Computers*, 2(7), Sept. 2007.

[16] D. Pastor, R. Gay, and A. Gronenboom. A sharp upper bound for the probability of error of likelihood ratio test for detecting signals in white gaussian noise. *IEEE Transactions on Information Theory*, 48(1):228–238, Jan. 2002.

[17] H. V. Poor. *An Introduction to Signal Detection and Estimation, 2nd Edition.* Springer-Verlag, New York, 1994.

[18] J. Portilla, V. Strela, M.J. Wainwright, and E.P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on Image processing*, 12(11):1338–1351, November 2003.

[19] R. J. Serfling. *Approximation theorems of mathematical statistics.* John Wiley and Sons, 1980.

[20] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 49(3):426–482, 1943.