Wavelet shrinkage using adaptive structured sparsity constraints

Diego Tomassi[a,c,d,*], Diego Milone[d], James D. B. Nelson[b]

[a] *Instituto de Matemática Aplicada del Litoral, UNL–CONICET, Argentina*
[b] *Department of Statistical Science, University College London, United Kingdom*
[c] *Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral*
[d] *Centro de Investigación en Señales, Sistemas e Inteligencia Computacional, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Argentina*

Abstract

Structured sparsity approaches have recently received much attention in the statistics, machine learning, and signal processing communities. A common strategy is to exploit or assume prior information about structural dependencies inherent in the data; the solution is encouraged to behave as such by the inclusion of an appropriate regularization term which enforces structured sparsity constraints over sub-groups of data. An important variant of this idea considers the tree-like dependency structures often apparent in wavelet decompositions. However, both the constituent groups and their associated weights in the regularization term are typically defined a priori. We here introduce an adaptive wavelet denoising framework whereby a sparsity-inducing regularizer is modified based on information extracted from the signal itself. In particular, we use the same wavelet decomposition to detect the location of salient features in the signal, such as jumps or sharp bumps. Given these locations, the weights in the regularizer associated to the groups of coefficients that cover these time locations are modified in order to favour retention of those coefficients. Denoising experiments show that, not only does the adaptive method preserve the salient features better than the non-adaptive constraints, but it also delivers significantly better shrinkage over the signal as a whole.

Keywords: structured sparsity, regularized regression, denoising, dual-tree complex wavelet transform

*Corresponding author. Colectora RN168 km. 472, Paraje El Pozo, 3000 Santa Fe, Argentina. TE: +54 342 4511370

Email address: diegot@santafe-conicet.gov.ar (Diego Tomassi)

1. Introduction

A key attraction of wavelets is their compressive representation of data. This is fundamental to powerful nonlinear processing methods such as wavelet shrinkage [1–3]. Early approaches often regarded wavelet coefficients as statistically independent. Further developments, however, showed that for many applications involving real-world signals and images, performance improved when the dependencies between coefficients were taken into account [4–9]. Most of such methods typically focussed on the persistency property which is often apparent across wavelet scales. The simplest models account for such statistical dependencies between parent coefficients at a given level of the decomposition and their child coefficients at the following level of finer resolution. Although methods based on these models proved successful in many applications such as denoising, compression, and classification, some concerns remained about the preservation of salient features in the signal, such as jumps or sharp bumps [10]. In applications such as denoising or deconvolution these features are typically over-smoothed which compromises the quality of the estimates. Some attempts to improve performance under these conditions explore total variation filtering [11, 12], combined Tikhonov and total variation regularization [10] and decompositions based on footprints of the discontinuities in the signal [13].

In this work we take advantage of the latest developments in regularized least-squares regression to promote tree-structured sparsity on the denoised estimates. Unlike previous tree-structured estimators [5, 8, 14–16], the method proposed here uses a lasso-like algorithm with a mixed-norm regularizer that induces structured sparsity over an overcomplete representation. A novel signal-driven approach is introduced to adapt the weights of the regularizer. The ability of shift-invariant complex wavelet transforms to detect salient features in the signal is exploited to design a penalization term which favours estimated jumps or sharp bumps during the optimization process. We show that this results in a denoising approach with better preservation of salient features.

The manuscript is organized as follows. In the remainder of the current section we provide motivation and discuss the specific contributions of our work in the context of the

current literature. In Section 2 we offer an overview of structured sparsity approaches and the dual-tree complex wavelet transform. The proposed method is introduced in Section 3. This considers both an oracle and a practical approach to account for the occurrence of salient features. In Section 4, results obtained in denoising experiments show the advantage of the proposed adaptive scheme over structured sparsity estimates set a priori. We then close with the main conclusions and a discussion of further work.

1.1. Motivation

Sparse representations have been at the core of many signal processing methods in recent years [17, 18]. Early algorithms such as basis pursuit [19] and matching pursuit [20] regarded coefficients as mutually independent, meaning that each atom in the decomposition is selected or discarded independently of its neighbours. In the signal processing community, efforts to introduce structured sparsity constraints were spurred by the compressed sensing paradigm [21, 22] which used prior knowledge to reconstruct signals with fewer samples than classical sampling theorems allowed. Model-based compressed sensing has showed promise in this context [23–25]. These early attempts, however, were based on non-convex or greedy optimization approaches. To achieve scalability without compromising consistency, non-greedy convex approaches are often desirable. To this end, regularized approaches using mixed-norms have proven successful in obtaining sparse estimates that retain an assumed dependence structure [26, 27].

It is important to note that most of the existing wavelet/structured models deal with the persistency property of the coefficients without taking into account any additional information provided by the specific choice of transformation or dictionary used to obtain the representation [27–29]. Moreover, all of these dependence structures are set a priori, and no further information from the signal is used to adapt them. In denoising applications, features with strong local high frequency content are often over-smoothed by such methods. This is due to the erroneous shrinkage or elimination of coefficients at finer scale levels. On the other hand, when regularization parameters are set to favour data-fitting much more than sparsity, the resulting estimates often retain too many fine-scale coefficients and remain noisy.

1.2. Contribution

In this work, a new signal processing method is developed that uses additional information, extracted directly from the signal, to reinforce the apriori structured sparsity constraints. To do so, we use the dual-tree complex wavelet transform (DTCWT) as the sparsity inducing transform together with a hierarchical mixed norm regularizer. The weights in the regularizer are adaptively modified in order to help preserve salient features of the analyzed signal. This adaptive modification is driven by a detection stage which aggregates information from the different scales of the wavelet decomposition to infer the locations of salient features in the signal. In this way, the mixed norm regularizer, defined a priori, is tailored to the observed signal.

1.3. Related work

Tree-structured estimators have been proposed earlier for wavelet decompositions, both in the signal processing and statistics communities [5, 8, 14–16]. They often rely on orthonormal transforms and hard-thresholding approaches. Following their success in machine learning and statistics, generalized lasso-type algorithms have received recent and growing attention for signal processing applications. The closest works are [29] and [30]. In [29], the parent-child dependence of wavelet coefficients is coded into overlapping groups, each of which comprises a parent-child pair. A variable replication approach is taken to account for different instances of a given coefficient appearing in different groups and a regularization term is added to account for the dissimilarity of the replicates of the same variable. Unlike the present work, their approach uses the standard discrete wavelet transform (DWT) without adding any additional information onto the structure assumed a priori. In [30], a chain structure is assumed to model the spectrogram of audio signals obtained from their short-time Fourier transform representation. This simple structure gives rise to a regularization term that is bounded above by a quantity which is simpler to compute, allowing for an efficient minimization-majorization algorithm. It should be noted, however, that it is suited for signals with emphasized band structures in their spectrogram. On the other hand, edge information has been used to aid image denoising [31] and reconstruction under compressed sensing applications [32]. To the best of our

knowledge, however, such information has not been used to adapt a structured regularizer as proposed here.

2. Background

This work builds upon two main ideas: the use of mixed-norm regularizers to obtain structured sparse estimates and the use of the DTCWT as a signal analysis tool to induce sparsity and locate salient features in the signal, while retaining desirable properties such as shift-invariance and low redundancy. We briefly examine these two ingredients in the following sections.

2.1. Regularized regression using structured sparsity constraints

Sparse linear models have become very popular in signal processing, machine learning and statistics. Their purpose is to predict an output by using linear combinations of only a small subset of the potential features that could describe the data. In this context, ℓ_1 regularization has become a widely-used tool to obtain estimation and feature selection simultaneously. The strategy can be stated as the solution of the convex optimization problem

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1. \tag{1}$$

In signal processing, the method is known as Basis Pursuit [19], and the aim is to find a sparse representation $\widehat{\boldsymbol{\theta}}$ of a signal $\mathbf{y}$ in terms of the columns (or atoms) of an overcomplete dictionary $\mathbf{A}$. In statistics, the method is known as the *Lasso* [33]. In this case, $\mathbf{A}$ is a data matrix comprising more variables than observations and the aim is to get a sparse regression of the observations $\mathbf{y}$ on the measurements $\mathbf{A}$.

The popularity of ℓ_1 regularization is largely due to the existence of efficient algorithms to solve (1) and a large body of supporting theory [33–39]. Nevertheless, in this formulation every variable or feature is regarded independent of the others. In practical situations, however, estimation can benefit from additional a priori knowledge regarding dependencies between sets of variables. Because of this, attention has been given in recent years to regularization problems that can accommodate such knowledge.

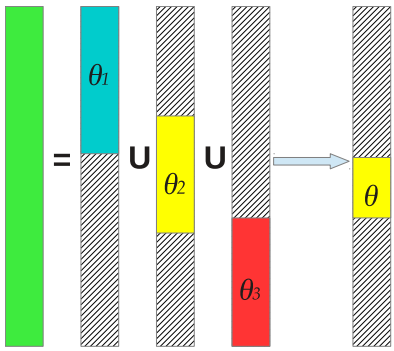


Figure 1: Group lasso with overlapping groups and its relationship with the zero and non-zero patterns of the obtained estimates. The non-zero pattern is obtained as the complement of the union of groups pushed to zero, in this case $\{1, 3\}$.

The most simple structure is the case in which we know, a priori, that sets of variables should be considered or discarded jointly from the linear model. Such a set of variables can be regarded as a group. Let $\mathcal{G}$ denote the set of groups, let $\boldsymbol{\theta}_g$ refer to the subset of variables of $\boldsymbol{\theta}$ in group g and let w_g be associated positive scalars acting as weights. The optimization problem reads [40]

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{g \in \mathcal{G}} w_g \|\boldsymbol{\theta}_g\|_q. \tag{2}$$

Common choices for q are $\{2, \infty\}$. The regularizer $\Omega_{1,q}(\boldsymbol{\theta}) = \sum_{g \in \mathcal{G}} w_g \|\boldsymbol{\theta}_g\|_q$ is often referred to as a mixed-norm regularizer and it can indeed be verified that it induces sparsity by deleting all the variables within a given group simultaneously. When the collection $\mathcal{G}$ forms a partition of the set of variables, the method is known as *group lasso* [40] and it is easy to see that the subset of coefficients shrunk to zero during estimation gives rise to a zero pattern that is the union of some groups in $\mathcal{G}$.

One way to generalise this formulation is to allow for more flexible grouping schemes whereby groups are allowed to overlap (i.e. such that $\mathcal{G}$ does not need to be a partition of the set of variables) [26–28, 41, 42]. In this case, the regularizer $\Omega_{1,q}(\boldsymbol{\theta}) = \sum_{g \in \mathcal{G}} w_g \|\boldsymbol{\theta}_g\|_q$ is still a norm, provided all covariates belong to at least one group, and it still induces complete groups of covariates to be set to zero. Moreover, a variable in a group that is set to zero during optimization will be pushed to zero even if it belongs also to other

groups that are not set to zero. More formally, it is shown in [27] that under very mild assumptions, the support of the solution $\widehat{\boldsymbol{\theta}}$ almost surely is the complement of the union of some groups in $\mathcal{G}$; that is,

$$\operatorname{supp}(\widehat{\boldsymbol{\theta}}) = \left(\bigcup_{g \in \mathcal{G}_0} g \right)^c = \bigcap_{g \in \mathcal{G}_0} g^c, \tag{3}$$

for some $\mathcal{G}_0 \subset \mathcal{G}$. This situation is illustrated in Figure 2.1. Sets of non-zero patterns that can be represented as in (3) are referred to as *intersection-closed*. There are some expected structures that cannot be described appropriately by an intersection-closed grouping scheme but which can be modeled by union-closed families of supports [41, 42].

Practical solution of the optimization problem involving overlapping groups is more challenging computationally and dedicated algorithms have been developed to address this task [43–46]. In addition, the selection of weights w_g has a greater impact on the estimate in the case of overlapping groups, since they have to mitigate not only the unbalanced size of the groups but also the over-penalization of variables appearing in a greater number of groups. Currently, principled rules to set the weights optimally for these regularizers are not available.

2.2. Dual-tree complex wavelet transform

Despite its widespread use in applications, the standard (real) DWT suffers from some important shortcomings such as a lack of shift-invariance and the substantial aliasing due to critical downsampling that affects perfect reconstruction if some processing is applied to the coefficients. For image analysis, the standard DWT with real wavelets also lacks directionality due to the standard tensor product construction of multidimensional wavelets. The DTCWT provides a better alternative to deal with those problems, while retaining simple computation and low redundancy [47]. It seeks to provide a nearly analytic wavelet transform using two real filter-bank trees, one for the real part and other for the imaginary part of the transform. These two real wavelet transforms use two different set of filters, each of compact support and satisfying perfect reconstruction. The choice of filters for each tree is not arbitrary, but they are designed to form an

approximate Hilbert pair, so that the resulting complex wavelet transform is as close as possible to analytic. Indeed, the design of the pair of filters is the key point to the success of the transform, and several efforts have addressed this topic [48–55]. The resulting transform is near shift-invariant [56] and, of particular importance to structured sparsity approaches, it has also been shown that the magnitude of the coefficients are more strongly dependent in inter-scale and intra-scale neighborhoods [47, 57, 58] than those of the DWT. Refer to [47] and references there in for a comprehensive introduction to the DTCWT and its properties.

3. Proposed method

In this section we describe the overlapping-group lasso approach for tree-structured wavelet estimators. We then propose a way to estimate the salient features directly from the signal and show that this affords the opportunity to adaptively choose the weights. We then conclude the section with a treatment of how these ideas can be incorporated into an optimization framework which solves the structured sparse estimation problem for overlapping groups of variables.

3.1. Overlapping-groups lasso with adaptive weights

The multiresolution nature of wavelet decompositions allows one to think of parent coefficients at a given scale level and child coefficients at the next finer scale level. Wavelet decompositions possess two properties of great interest to structural sparsity approaches: firstly, they are typically sparse and secondly, coefficients in the same time interval usually show a persistency property across scale [5]. The persistency property means that the magnitude of the child coefficients depends strongly on that of their parent— a large/small parent usually implies a large/small child. Such a dependence can be modelled as a set of trees, each rooted at a wavelet coefficient from the coarsest scale of the decomposition [28, 59]. Let I be the set of indexes for the coefficients $\boldsymbol{\theta}\colon I \mapsto \mathbb{C}$ and let $I_0 \subset I$ be the subset of indexes indicating the wavelet coefficients at the coarsest scale. Furthermore, let $g(i)$ be a subset of indexes of I organized as a tree

rooted at some index i, and let $\boldsymbol{\theta}_{g(i)}$ be the corresponding subset of coefficients in this tree. Consider the collection of groups

$$\mathcal{G} := \{g(i) \colon i \in I\}. \tag{4}$$

We assume the inter-scale dependence structure between wavelet coefficients determines a forest-like hierarchical structure $\mathcal{M}(\boldsymbol{\theta})$, so that for all $i, j \in I$, $g(i) \subset g(j)$, $g(j) \subset g(i)$, or $g(i) \cap g(j) = \emptyset$, and

$$\mathcal{M}(\boldsymbol{\theta}) := \bigoplus_{i \in I_0} \boldsymbol{\theta}_{g(i)}. \tag{5}$$

This grouping scheme is illustrated in Figure 2-(a), where circles represent coefficients of the wavelet representation and rectangles represent the groups. For simplicity, only one tree of the forest is shown in the figure. This nested structure can be understood as an example of the composite absolute penalties family and it allows one to obtain estimates whose supports show the desired persistency property. See [60] for a discussion of these type of penalties. An example is shown in Figure 2-(b). Highlighted rectangles show the groups pushed to zero during optimization and the shaded circles show the variables set to zero as a consequence. The support of the estimate is then represented by the set of white circles, which is the complement of the union of the groups set to zero during the optimization. This example illustrates that the hierarchical structure defines a intersection-closed set of supports, which can be then obtained solving a regularized regression problem like (2); in particular,

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{i \in I} w_i \|\boldsymbol{\theta}_{g(i)}\|_2. \tag{6}$$

Whilst the use of this structure and the associated mixed-norm regularizer has already been proposed [27, 28], we note that the introduction of this structured regularizer also requires the specification of the weights w_i. Indeed, although much less attention has been given to this topic, it is well-known that the values of the weights in fact have an important effect on determining the support of the resulting estimates [26, 27].

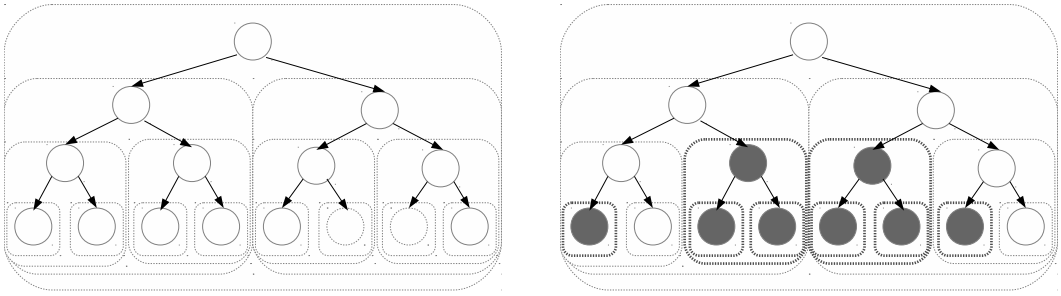


Figure 2: (a): Grouping scheme for the proposed mixed-norm regularization. (b): Example of estimate induced by the adopted grouping scheme; its support is the complement of the union of the sets pushed to zero during optimization which in turn sets to zero the coefficients represented by shaded circles.

Motivated by this dependence on the choice of the weights, we propose to adapt them from the a priori specification by taking into account the salient features of the analyzed signal. In particular, we modify the weights w_g with the aim of achieving a better reconstruction of jumps and bumps, which are often over-smoothed by denoising algorithms.

To formalize the idea, let $k \in \mathrm{supp}(f)$ be a location of interest, where the signal f presents a salient feature such as a jump or a sharp bump. Let $\check{\mathcal{G}}_k \subset \mathcal{G}$ be the set of groups containing coefficients $\theta_{j,n}$ related to k and let $\check{\mathbf{w}}_k$ be the vector of associated weights. We define a shrinkage operator on the weights $P : w \in \mathbb{R}_+ \to \mathbb{R}_+$ by

$$P(w_i) \equiv \tilde{w}_i = \begin{cases} \alpha_i w_i & \text{if } g(i) \in \check{\mathcal{G}}_k, \\ w_i & \text{otherwise,} \end{cases} \tag{7}$$

where $\alpha_i \in [0; 1]$. For a given group $g(i)$, $\tilde{w}_i \leq w_i$ and then the shrunk weights allows the respective groups to be penalized more weakly during the optimization process. In this way, coefficients related to salient features of the signal can be favoured in the regularized estimation process so that they are surely retained in the final estimate.

3.2. Edge and ridge detection via DTCWT

We now focus on the practicalities of detecting and locating these salient features. It is shown in [61] that analytic complex wavelets (with a non-negative frequency spec-

trum) can be used to extract edge-like and ridge-like features from noise-free signals. Theoretically, at the limit of the finest resolution, all the coefficients are zero except those corresponding to the location of the features. Moreover, the phase of these limiting coefficients can be used to distinguish between an edge-like or a ridge-like structure at a given position of interest.

In practice, the edges and bumps can be estimated by replacing the limiting process with an average of the normalized complex coefficients over a finite number of scales. This approach, nevertheless, has two important limitations. The first is that an undecimated or interpolated complex wavelet representation is required so that the coefficients included in the average are well localized in time in order for the the detected locations of interest to be accurate. The second limitation is perhaps more serious, namely that the approach as stated currently does not take into account noisy conditions. Thus, there is no guarantee on the attainable performance when using it in noisy scenarios.

For 1D signals, it is well-known that jumps and ridges give rise to strong persistency figures in their scalogram when using the continuous wavelet transform [62]. This behaviour is still noticeable when using a discrete wavelet transform. Motivated by this, we propose to average the magnitude of the wavelet coefficients throughout the different scales to get the signatures. Since the DTCWT provides a decimated representation, we interpolate the magnitudes of the wavelet coefficients to match the resolution of the finest scale.

3.2.1. A measure for relevant information

Let $\boldsymbol{\theta}$ be the DTCWT of discrete signal f and let $|\tilde{\theta}_{j,n}|$ be the interpolated magnitude of the coefficient corresponding to location n in time at the scale indexed by j. We measure the amount of important information on f at location n by a function $\delta : \mathrm{supp}(f) \to [0; 1]$

$$\delta(n) := \left(\prod_{j=J_0}^{J} \gamma\big(|\tilde{\theta}_{j,n}|\big) \right)^{\frac{1}{J-J_0+1}}, \tag{8}$$

with $\gamma(\cdot)$ a normalization function given by

$$\gamma\big(\theta_{j,n}\big) := \frac{\theta_{j,n} - \min \theta_{j,\cdot}}{\max \theta_{j,\cdot} - \min \theta_{j,\cdot}} \,. \tag{9}$$

At locations where the signal has a jump or a bump, δ is close to 1. For noise-free scenarios, $\delta \to 0$ at intervals where the signal is smooth. In this case, J_0 should be set to 1, in order to include the coefficients at the finest scale of the decomposition. In noisy scenarios, coefficients at the finest scale can be noisy and averaging from scale $J_0 = 2$ can prove useful. In addition, when the decomposition runs up to a very coarse level, very little information can be extracted from the coarsest scales about the location of salient features. Indeed, for very smooth signals the interpolated coefficients can become nearly constant at the coarsest scales. The geometric mean given by Equation (8) shows small sensitivity to this effect because it depends mostly on the coefficients whose magnitude is close to zero [63].

3.2.2. Detection of points of interest

Given a threshold $\tau \in (0; 1)$ and a minimum length of interval r, we define the sets $\mathrm{K}_{\tau,r}(1), \mathrm{K}_{\tau,r}(2), \ldots, \mathrm{K}_{\tau,r}(S)$ to be the collection of non-empty, non-intersecting, and non-neighbouring intervals of length greater than r such that $\delta(n) > \tau$ and where each set is simply connected (i.e. for each s, the set $\mathrm{K}_{\tau,r}(s)$ does not contain any gaps between elements). Then, the set of S-many locations $\mathcal{K}_{\tau,r}$ of the salient features in f is given by the points where $\delta(n)$ attains its maximum value within each interval of $\mathrm{K}_{\tau,r}$; that is,

$$\mathcal{K}_{\tau,r} := \left\{ \max_{n \in K_{\tau,r}(s)} \delta(n) \right\}_{s=1}^{S} .$$

The procedure to estimate the set of locations $\mathcal{K}_{\tau,r}$ is summarized in Algorithm 1. The detection of points of interest as described above depends on the choice of parameters τ and r. For fixed r, moving τ from 0 to 1 changes the pattern of connected sets $\mathrm{K}_{\tau,r}$. In particular, for $\tau_1 < \tau_2$, $\mathrm{K}_{\tau_1,r} \supset \mathrm{K}_{\tau_2,r}$ but the set of detected locations $\mathcal{K}_{\tau_2,r}$ can add a new element with respect to $\mathcal{K}_{\tau_1,r}$ if a local minimum in δ is passed-through when going

Algorithm 1 Detection of points of interest

Inputs: f; r, n_0.
Outputs: $\mathcal{K}_{r,n_0}$
procedure DETECTPOI(f; r, n_0)
  $\boldsymbol{\theta} \leftarrow \text{DTCWT}(f)$
  $|\tilde{\boldsymbol{\theta}}| \leftarrow \text{Interpolate}(|\boldsymbol{\theta}|)$
  for all $n \in \text{supp}(f)$ **do**
    $\delta(n) \leftarrow$ Compute the geometric mean of $|\tilde{\boldsymbol{\theta}}|_{\cdot,n}$ ▷ see (8)
  end for
  $\{\tau_q\} \leftarrow$ Compute quantiles of δ
  for all τ_q **do**
    $K_{\tau_q,r} \leftarrow \{n : \delta(n) > \tau_q\}$
    for all $K_{r,\tau_q}(i)$ **do**
      $\mathcal{K}_{\tau_q,r}(i) \leftarrow \arg\max_{n \in K_{\tau_q,r}(i)} \delta(n)$
    end for
  end for
  for all $\{\mathcal{K}_{\tau_q,r}(i)\}$ **do**
    $\kappa_i \leftarrow$ Check persistence $\{\mathcal{K}_{\tau_q,r}(i)\} \geq n_0$
  end for
  $\mathcal{K}_{r,n_0} \leftarrow \{\kappa_i\}$
end procedure

from τ_1 to τ_2. If the new extremum is a local maximum due to noise or a irrelevant fluctuation in δ, it will disappear as τ is increased. On the other hand, if the added extremum represents a location of a salient feature, it will persist over a wide range of $\tau > \tau_2$. As such, persistence of detected locations on the sequence $\{\mathcal{K}_{\tau_n,r}\}$, obtained using a sequence of threshold values $\{\tau_n\}$, is an indicative of a true salient feature in f. For this procedure to be useful in general, we should define the sequence $\{\tau_n\}$ to be adaptive to the data, since the peaks in δ at the locations of interest are less emphasized when the noise increases. To do so, we propose to match $\{\tau_n\}$ to a sequence of quantiles of δ and look for the locations that persist in $\{\mathcal{K}_{\tau_n,r}\}$ at least n_0 times.

3.3. Implementation

In this section we describe an algorithm to solve (6) and a criterion to select the regularization parameter λ.

3.3.1. Optimization algorithm

Several numerical approaches have been proposed to deal with the $\ell_{1,2}$ minimization problem involving overlapping groups [43–46, 59]. In this work we solve (6) using a proximal method based on Mureau-Yosida regularization [64]. Our presentation follows

[64]. Let $\Omega_{1,2} = \sum_{i\in I} w_i \|\boldsymbol{\theta}_{g(i)}\|_2$ be the regularization term in (6). Under the Mureau-Yosida regularization framework, the regularization associated with $\Omega_{1,2}$ for a given $\mathbf{v} \in \mathbb{R}^p$ is given by:

$$\phi_\lambda(\mathbf{v}) = \arg\min_{\mathbf{v}} \left\{ \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{v}\|_2^2 + \lambda\Omega_{1,2} \right\}, \tag{10}$$

for some $\lambda > 0$. Let $\pi_\lambda(\cdot)$ be a minimizer of (10) and let $\widehat{\boldsymbol{\theta}}$ be an optimal solution to (6). Then, $\widehat{\boldsymbol{\theta}}$ satisfies:

$$\widehat{\boldsymbol{\theta}} = \pi_{\lambda_\tau}(\widehat{\boldsymbol{\theta}} + \tau\mathbf{A}^\dagger(\mathbf{y} - \mathbf{A}\widehat{\boldsymbol{\theta}})), \ \ \forall\tau > 0, \tag{11}$$

with $\mathbf{A}^\dagger$ and $\mathbf{A}$ here denoting the DTCWT and its inverse transformation, respectively. Equation (11) affords the opportunity to apply an accelerated gradient descent for solving (6). The key point of the algorithm is the solution of (10). It is shown that this minimizer has indeed an analytical solution that can be found with Algorithm 2 (see [64] for details).

Algorithm 2 Mureau-Yosida regularization

Inputs: $\mathbf{v} \in \mathbb{R}^p$; grouping structure $\mathcal{G}$, related weights $\{w_i\}$ and $\lambda > 0$.
Outputs: $\pi_\lambda(\mathbf{v})$
procedure SolveMyTree($\mathbf{v}, \lambda; \mathcal{G}, \{w_i\}$)
  $\lambda_i \leftarrow \lambda w_i$
  $\mathbf{u}^{L_0+1} \leftarrow \mathbf{v}$
  for $i = L_0$ to 1 **do** ▷ Iteration runs from finest to coarsest scale.
    for all $g(j)$ at scale i **do**

$$\mathbf{u}^i_{g(j)} \leftarrow \begin{cases} \mathbf{0}, & \text{if } \|\mathbf{u}^{i+1}_{g(j)}\|_2 \le \lambda_j \\ \frac{\|\mathbf{u}^{i+1}_{g(j)}\|_2 - \lambda_j}{\|\mathbf{u}^{i+1}_{g(j)}\|_2}\mathbf{u}^{i+1}_{g(j))}, & \text{if } \|\mathbf{u}^{i+1}_{g(j)}\|_2 > \lambda_j, \end{cases} \tag{12}$$

    end for
  end for
  $\pi_\lambda(\mathbf{v}) \leftarrow \mathbf{u}^1$
end procedure

3.3.2. Selection of the regularization parameter

For practical applications, an important aspect to be specified is the value of the regularization parameter λ. Although there exist well-stablished criteria for selecting λ for the standard lasso and related problems when $\mathbf{A}$ is orthonormal, such criteria do not apply for the case of overlapping groups. In recent papers involving structured regularizers with overlapping groups, selection of the regularization parameter has been done empirically using simulations. For instance, in [30] a numerical study is carried out

to relate the value of λ with the expected reduction of noise variance for uncorrelated white Gaussian noise. The offered values depend on the choice of block structure in the regularizer. In [10], a grid of values for a pair of regularization parameters is evaluated and the best pair is selected graphically.

The approach proposed in this paper is based on theoretical results for regularized M-estimators with decomposable regularizers [65]. It is easy to see that the highly structured regularizer used in (6) is decomposable in the sense defined in [65]. Let $\widehat{\boldsymbol{\theta}}_\lambda$ denote the solution to (6) for a given value of λ. It is shown in that paper that theoretical guarantees on $\|\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^o\|$ can be found provided

$$\lambda \geq 2\ \Omega^*_{1,2}(\nabla \mathcal{L}(\boldsymbol{\theta}^o)), \tag{13}$$

where $\Omega^*_{1,2}$ stands for the dual of the regularizer $\Omega_{1,2}$, $\boldsymbol{\theta}^o$ is the true vector of coefficients under the tree-structured model, and $\mathcal{L}$ is the loss function, which in this paper is the squared-error loss. Following this, we require:

$$\lambda \geq 2\ \left\| w_i \|\boldsymbol{\psi}_{g(i)}\|_2 \right\|_\infty, \tag{14}$$

with $\boldsymbol{\psi}$ the DTCWT of the noise $\boldsymbol{\eta}$ in the model $\mathbf{y} = \mathbf{A}\boldsymbol{\theta}^o + \boldsymbol{\eta}$.

Expression (14) is impossible to compute, since it involves the DTCWT of the *true* noise $\boldsymbol{\eta}$. Nevertheless, we can use simulations to find a bound for λ according to (14). The procedure involves generating a noise signal $\boldsymbol{\eta}_i$ with variance σ^2_η and picking the maximum of $S = \max\{w_i \|\boldsymbol{\psi}_{g(i)}\|_2\}$ using the non-adapted weights $\{w_i\}$. After repeating the steps to obtain a large number of replicates of S, the distribution of $\tilde{S} = 2S/\sigma$ is found to be invariant. Thus, we can use a quantile of the distribution of S to set a bound for λ that holds with high probability. For uncorrelated white Gussian noise and using the .95 quantile of the empirical distribution, we set $\lambda = 0.568\widehat{\sigma}$, with $\widehat{\sigma}$ an estimate of the noise variance.

3.3.3. Overall algorithm

An overall algorithm to implement the proposed method is shown in Algorithm 3, which joins the optization steps and the procedure to detect the point of salient features discussed in the current section. The evolution of the objective function as the number of iterations increases within the main loop in the denosing algorithm is shown in Figure 3. Shown curve corresponds to the piecewise-polynomial signal from WaveLab [66], corrupted with white Gaussian noise of variance $\sigma_\eta^2 = 4$. Similar evolutions are obtained for other types of signals and noise levels. It can be seen that convergence is monotone and little improvement is gained after 20 iterations. In Appendix A, the performance obtained with λ chosen as explained in the previous section is compared against the best performance obtained with Algorithm 3 using a fine grid of values of λ. The influence of r and n_0 on the overall denoisng procedure is analyzed in Appendix B.

Algorithm 3 Adaptive tree-structured wavelet shrinkage

Inputs: $\mathbf{y}$; grouping structure $\mathcal{G}$ and related prior weights $\{w_i\}$.
Outputs: Denoised signal $\widehat{\mathbf{z}}$
procedure DENOISING($\mathbf{y}$; $\mathcal{G}$, $\{w_i\}$,r, n_0)
  $\mathcal{K}_{r,n_0} \leftarrow$ DETECTPOI($\mathbf{y}; r, n_0$) ▷ see Algorithm 1.
  $\check{\mathcal{G}}_k \leftarrow$ Find groups in $\mathcal{G}$ related to $\mathcal{K}_{r,n_0}$
  Set $\mathbf{x} \leftarrow$ Compute DTCWT($\mathbf{y}$)
  $\widehat{\sigma}^2 \leftarrow$ Estimate noise variance from $\mathbf{x}$
  $\alpha \leftarrow$ Set shrinkage factor according to $\widehat{\sigma}$
  $\{w_i\} \leftarrow$ Adapt weights $\{w_i\}$, from $(\check{\mathcal{G}}_k, \alpha)$ ▷ see (7).
  $\lambda \leftarrow$ Set λ according to $\widehat{\sigma}$ ▷ see (14) and comments below it.
  Set $L = 1,\ \zeta = 1,\ \zeta_p = 0; \mathbf{x}_p = \mathbf{x}$
  repeat
    $\beta \leftarrow (\zeta_p - 1)/\zeta$
    $\mathbf{s} \leftarrow \mathbf{x} + \beta(\mathbf{x} - \mathbf{x}_p)$
    $G \leftarrow$ Compute gradient from $\mathbf{y}, \mathbf{s}$
    while $\|\mathbf{A}(\mathbf{x} - \mathbf{s})\|_2^2 > L\|\mathbf{x} - \mathbf{s}\|_2^2$ **do**
      $\mathbf{v} \leftarrow \mathbf{s} - G/L$
      $\mathbf{x} \leftarrow$ SOLVEMYTREE($\mathbf{v}, \lambda/L; \mathcal{G}, \{w_i\}$) ▷ see Algorithm 3
      $L \leftarrow \max(2L, \|(\mathbf{x} - \mathbf{s})\|_2^2/\|\mathbf{A}(\mathbf{x} - \mathbf{s})\|_2^2)$
    end while
    $(\zeta, \zeta_p) \leftarrow$ Update(ζ, ζ_p)
  until convergence
  $\widehat{\boldsymbol{\theta}} \leftarrow \mathbf{x}$
  $\widehat{\mathbf{z}} \leftarrow$ Compute IDTCWT($\widehat{\boldsymbol{\theta}}$)
end procedure

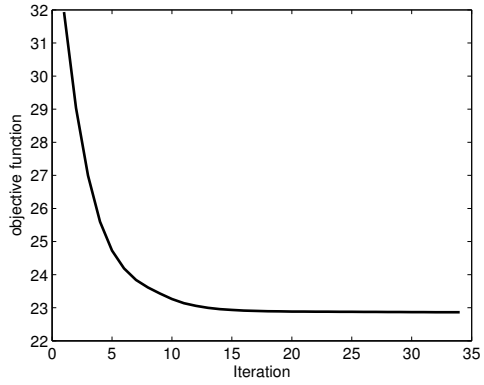


Figure 3: Convergence of the proposed algorithm. Shown figure corresponds to the PP signal corrupted with white Gaussian noise of variance $\sigma_\eta^2 = 4$.

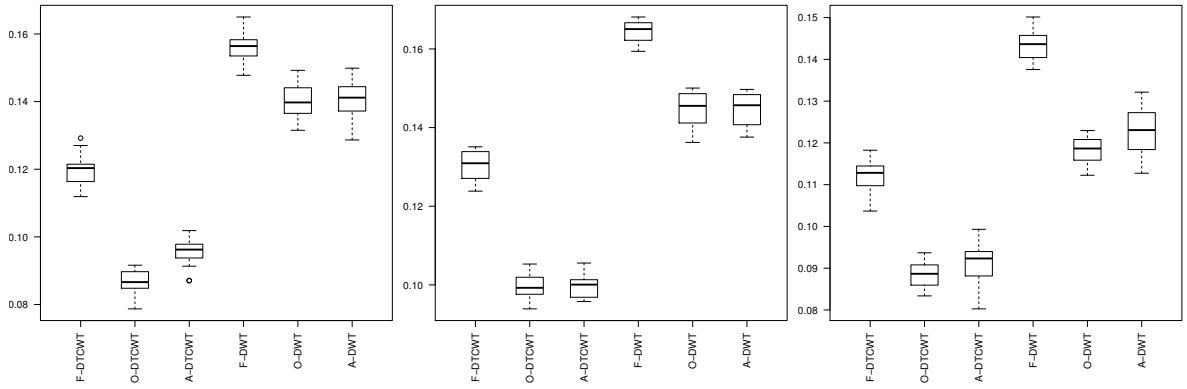
4. Experiments and results

To assess the performance of the proposed method, simulation studies were carried out using synthetic signals from the WaveLab Toolbox [66], which have been used extensively for benchmarking wavelet-based denoising methods. The Blocks, Bumps, Piecewise-Regular (PR) and Piecewise-Polynomial (PP) signals were chosen, since they comprise smooth segments with jumps and bumps and serve to illustrate the kind of signals targeted to by the proposed method. All test signals were generated with $N = 1024$ sample points. Independent and identically distributed white Gaussian noise was added to the signals, at different SNR determined by the variance σ_η^2 of the noise. Near-symmetric (13,19)-tap filters were used to compute the first stage of the decomposition, while Q-Shift (14,14)-tap filters were used for the rest of the scales. Preliminary experiments with other combinations of filters showed that the effect of this choice on the performance of the proposed method was neglible for most of the tested conditions. Decomposition was carried out up to level $L_0 = 7$. Implementation of Algorithm 3 was done in MATLAB. The optimization code was adapted from [67] to deal with complex variables. For the detection of salient features, interpolation of coefficient magnitudes at each scale was carried out using a cubic interpolator, although the proposed method was not found sensitive to the type of interpolator chosen.

4.1. DTCWT vs DWT

A key characteristic of the proposed method is the use of the DTCWT instead of the DWT. In this subsection we illustrate the benefits of this choice. A simulation was run to assess the performance of the proposed tree-structured estimators for both DTCWT and DWT decompositions, using the same grouping structure for both of them. Both the adaptive methodology exploiting detection of salient features (A-DTCWT and A-DWT) and the alternative methods with the weights of the goups fixed a priori were assessed (F-DTCWT and F-DWT). In addition, for the adaptive structured estimators, results obtained using the oracle locations of salient features (O-DTCWT and O-DWT) were included in order to explore whether the choice of transformation affects the detection of salient features, the denoising process or both of them. The results of 100 experiments were averaged to assess the performance of each method. In each run, all the methods processed the same signal so that noise variability between realizations do not contribute to a difference in performance between the methods. To avoid the influence of the choice of the regularization parameter for the adaptive methods, a set of 1000 uniformly spaced samples of λ in the interval $[0, \lambda_{max}]$ were tried out for each method at each run. λ_{max} is such that for $\lambda > \lambda_{max}$, the resulting estimate is a zero vector. λ_{max} was estimated in each run using the algorithm proposed in [64]. Only the best estimate across the differente values of λ was picked as the result corresponding to the run, both for DTCWT and DWT-based methods.

Obtained results are shown in Figure 4. Reconstruction errors for $\sigma_\eta^2 = 4$ only are reported, since very similar figures are obtained for other signal to noise ratios. It can be seen that methods using the DTCWT significantly outperforms alternatives of the same algorithms that use the DWT. Furthermore, it can be seen that the difference in performance between O-DTCWT and A-DTCWT is roughly the same as between O-DWT and A-DWT, albeit the O-DTCWT clearly outperforms O-DWT. Thus, the main benefit of the DTCWT over the DWT lies in the signal denoising step and not in the detection of salient features. These results, together with the fact that the magnitude of the (near shift-invariant) DTCWT coefficients have stronger dependence in inter-scale



(a) (b) (c)

Figure 4: Comparison of results obtained when using the DTCWT or the standard DWT. (a) Blocks; (b) Bumps; (c) PP signal. In all cases, variance of the noise is $\sigma_\eta^2 = 4$.

and intra-scale neighborhoods than those of the DWT, suggest that the assumed tree-structured models are better suited to DTCWT decompositions than to DWT.

4.2. Accuracy of the edge/ridge detection method

Examples of obtained results are shown in Figure 5 and Figure 6. The variance of noise increases from left to right in each figure, starting with the noise-free condition in panel (a). In each panel, the upper box shows the analyzed signal, the middle-box shows the estimated δ and the box at the bottom shows the detected locations. The choice of the minimum length of interval r and minimum persistence n_0 for each signal remained fixed across the different noise level. In particular, values $r = 4$ and $n_0 = 7$ where used for both signals, and fifteen different values of τ where used to assess persistence. It can be seen that for the noise-free conditions and the first noisy conditions, the algorithm achieves perfect detection of the points of interest. For the most severe noisy condition, the algorithm introduces some false positives, two in the case of the PR signal and only one for the PP signal. The rate of false positives can be controlled by modifying the values of r and the minimum persistence n_0. An increase in any of them, most notably in n_0, reduces the number of detections. As a trade-off, some points of interest will be missed.

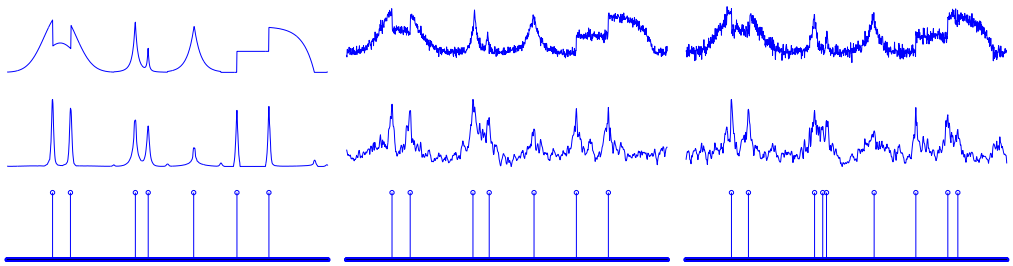


Figure 5: Detection of salient features for the Piecewise-Regular signal from WaveLab. (a): noise-free condition; (b) $\sigma_\eta^2 = 4$; (c) $\sigma_\eta^2 = 9$.

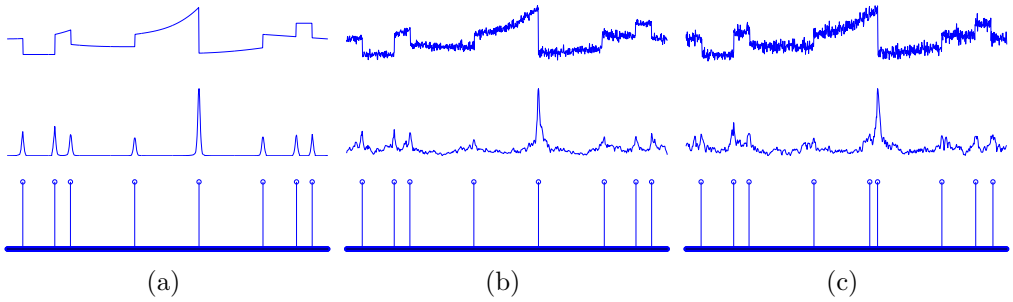


Figure 6: Detection of salient features for the PP from WaveLab. (a): noise-free condition; (b) $\sigma_\eta^2 = 4$; (c) $\sigma_\eta^2 = 9$.

ROC curves were constructed to facilitate assessment of the performance salient feature detector. For a given noise level, the detection algorithm was run with r taking values in $\{1, 3, 5, 7, 9\}$ and n_0 taking integer values from 1 to 50. The set of thresholds $\{\tau_n\}$ used to evaluate persistence of candidate locations was determined using a regular grid of fifty points in the interval $[Q_{0.50}, Q_{0.95}]$. Each point in the ROC curve gives the true positives rate (TPR) and the false positives rate (FPR) for a given value of the persistence parameter n_0, averaged over 500 replicates of the experiment. Figure 7 shows the ROC curves obtained for PR and PP signals, for three different noise levels. For both signals, it can be seen that the detection ability degrades with increasing noise levels, as showed by the maximum attained values of TPR. It is important to clarify that these low TPR are in fact due mostly to detection of salient features in locations close to the true ones. Computation of the TPR is sensitive to these displacements of the detected points

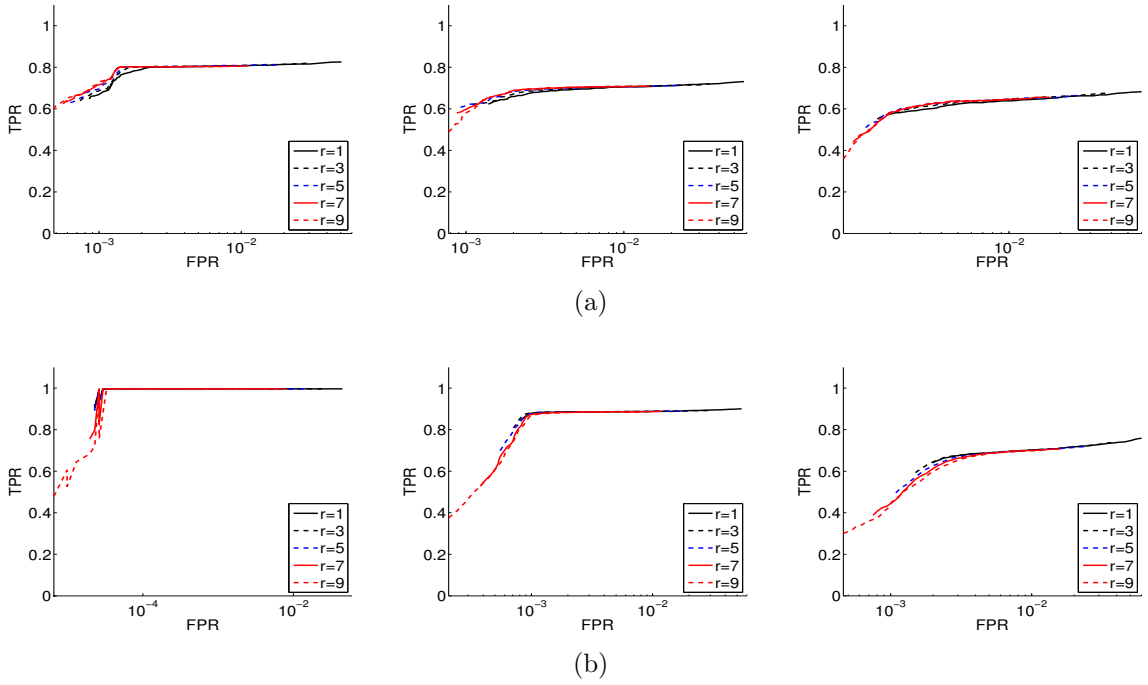


Figure 7: ROC curves for the detector of salient features at different noise levels.(a): PR signal; (b): PP signal. Noise levels increase from left to right, at $\sigma_\eta^2 = 1$, $\sigma_\eta^2 = 4$, $\sigma_\eta^2 = 9$.

relative to the true ones. Nevertheless, when plotting such detections like in Figure 5 and Figure 6, they are usually not noticeable. It can be seen also that the different choices of the parameter r lead to very similar ROC curves. Comparing the figures obtained for both signals shows that better performance is achieved for the PP signal; this suggests that jumps are easier to detect than bumps.

A study of the influence of the choice of parameters r and n_0 on the peformance of the overall denoising process as proposed in Algorithm 3 is left to Appendix B.

4.3. Denoising

To illustrate the effect of weight adaptation in the mixed-norm regularizer, denoising results were obtained using the following methods: (i) proposed method with locations of salient features given by an oracle (O-DTCWT); (ii) proposed method with locations of salient features estimated by the proposed detection algorithm (A-DTCWT); (iii) regularized method using the same structured prior but with weights fixed in advance (F-

DTCWT); and (iv) standard soft-thresholding algorithm as proposed in [2] (ST). Results for ST were included as a baseline, to help appreciate whether the added complexity of the proposed method was worth the gain in efficiency. Performance is measured as both the $\|\cdot\|_\infty$ and $\|\cdot\|_2$ of the reconstruction errors, averaged over 100 replicates of the experiment. Informed by the ROC analysis in Section 4.1, the parameters that control the detection of salient features were set to $r = 4$ and $n_0 = 7$, with persistence measured using 15 different values of τ sampled regularly on the interval specified by the quantiles $Q_{0.5}$ and $Q_{0.95}$. This choice for $n_0 = 7$ proved convenient for all signals and all noise levels, representing a good compromise between TPR and FPR. Other combinations of (r, n_0) are similarly capable, as shown in Appendix B.

For all the methods except ST, initial values of the weights were set according to the cardinality of each group, using[1] $w_g = |\boldsymbol{\theta}_g|^{1/4}$. These weights remained fixed for F-DTCWT but were modified using information from the location of salient features for O-DTCWT and A-DTCWT. For the adaptive methods, given the locations of the salient features, the weights of the corresponding groups in the regularizer were shrunk using $\alpha_g = \min(1, \widehat{\sigma}/\sigma_0)$, with $\sigma_0 = 4$, $\widehat{\sigma} = \text{MAD}/0.6745$, and MAD the median absolute value of the appropriately normalized wavelet coefficients at the finest resolution, as proposed for the standard ST method [2]. With this choice of α_g, in a noise-free scenario the weights w_g related to the locations of salient features are set close to zero, while in very noisy scenarios the weights will remain unchanged[2].

Results are shown in Table 1. For Blocks, Bumps and PP signals, it can be seen that the proposed method involving an adaptive regularizer outperforms the other alternatives for all the tested conditions, both when using $\|\cdot\|_\infty$ or $\|\cdot\|_2$ as the measure of performance. It is important to note that these conclusions are valid for the adaptive regularizer based on oracle information as well as for the adaptive approach using the proposed algorithm to detect the locations of the salient features of the signal. Indeed, results show that

[1] $|\cdot|$ here denotes cardinality.

[2] Note that α_g should be such that it favors retention of involved groups, but not force it. In this sense, for vey mild noisy conditions, it might be appropiate to lower bound the value of α_g.

Table 1: Performance of denoising algorithms for synthetic signals.

Error	σ_η^2	ST mean(sd)	F-DTCWT mean(sd)	O-DTCWT mean(sd)	A-DTCWT mean(sd)
Blocks signal					
$\|\mathbf{z}-\widehat{\mathbf{z}}\|_\infty/\|\mathbf{z}\|_\infty$	1	0.3238 (0.0224)	0.1742 (0.0235)	0.1411 (0.0213)	0.1475 (0.0235)
	4	0.4594 (0.0356)	0.3182 (0.0425)	0.2666 (0.0412)	0.2811 (0.0391)
	9	0.5621 (0.0351)	0.4362 (0.0593)	0.4029 (0.0617)	0.4218 (0.0709)
$\|\mathbf{z}-\widehat{\mathbf{z}}\|_2/\|\mathbf{z}\|_2$	1	0.1011 (0.0031)	0.0666 (0.0019)	0.0581 (0.0017)	0.0588 (0.0022)
	4	0.1479 (0.0052)	0.1182 (0.0035)	0.0887 (0.0028)	0.0941 (0.0049)
	9	0.1881 (0.0064)	0.1650 (0.0062)	0.1219 (0.0057)	0.1310 (0.0080)
Bumps signal					
$\|\mathbf{z}-\widehat{\mathbf{z}}\|_\infty/\|\mathbf{z}\|_\infty$	1	0.1509 (0.0105)	0.0631 (0.0077)	0.0434 (0.0067)	0.0423 (.0104)
	4	0.2206 (0.0185)	0.1194 (0.0165)	0.0812 (0.0127)	0.0871 (0.0048)
	9	0.2876 (0.0263)	0.1717 (0.0205)	0.1189 (0.0174)	0.1282 (0.0256)
$\|\mathbf{z}-\widehat{\mathbf{z}}\|_2/\|\mathbf{z}\|_2$	1	0.1249 (0.033)	0.0714 (0.0018)	0.0661 (0.0017)	0.0656 (0.0020)
	4	0.1742 (0.0060)	0.1309 (0.0039)	0.1009 (0.0038)	0.1015 (0.0048)
	9	0.2255 (0.0090)	0.1857 (0.0060)	0.1353 (0.0060)	0.1384 (0.0075)
Piecewise-Regular signal					
$\|\mathbf{z}-\widehat{\mathbf{z}}\|_\infty/\|\mathbf{z}\|_\infty$	1	0.1989 (0.0194)	0.1435 (0.0371)	0.1333 (0.0086)	0.1333 (0.0086)
	4	0.2953 (0.0220)	0.2476 (0.0281)	0.2059 (0.0314)	0.2101 (0.0357)
	9	0.3477 (0.0245)	0.3165 (0.0470)	0.2617 (0.0042)	0.2669 (0.0196)
$\|\mathbf{z}-\widehat{\mathbf{z}}\|_2/\|\mathbf{z}\|_2$	1	0.0644 (0.0026)	0.0614 (0.0037)	0.0617 (0.0014)	0.0621 (0.0021)
	4	0.1045 (0.0041)	0.1044 (0.0052)	0.0941 (0.0033)	0.0945(0.0031)
	9	0.1355 (0.0056)	0.1224 (0.0042)	0.1112 (0.0061)	0.1151 (0.0083)
Piecewise-Polynomial signal					
$\|\mathbf{z}-\widehat{\mathbf{z}}\|_\infty/\|\mathbf{z}\|_\infty$	1	0.1897 (0.0176)	0.1091 (0.0871)	0.0871 (0.0151)	0.0871 (0.0151)
	4	0.3010 (0.0471)	0.1934 (0.0256)	0.1663 (0.0267)	0.1786 (0.0290)
	9	0.3953 (0.0685)	0.2619 (0.0433)	0.2436 (0.0410)	0.2516 (0.0491)
$\|\mathbf{z}-\widehat{\mathbf{z}}\|_2/\|\mathbf{z}\|_2$	1	0.0812 (0.0025)	0.0642 (0.0019)	0.0572 (0.0016)	0.0573 (0.0017)
	4	0.1207 (0.0043)	0.1135 (0.0040)	0.0891 (0.0041)	0.0918 (0.0057)
	9	0.1519 (0.0062)	0.1319 (0.0056)	0.1181 (0.0050)	0.1219 (0.0074)

the practical method achieves scores very similar to those for the oracle version and the difference between them is not significant for moderate levels of noise. This can be visualized in Figure 8, which shows boxplots of the obtained reconstruction errors when the variance of the noise is $\sigma_\eta^2 = 4$. It is clear from the figure that results obtained with the practical implementation are almost identical to those achieved when the locations of the salient features are known. In addition, variance of the obtained errors is similar for the proposed method and standard soft-thresholding for all tested conditions. It is interesting to note that, although experiments have considered Gaussian noise only, neither the overall method nor the strategy to select λ assume this condition.

4.4. Example with real data

Figure 9 shows the performance of the proposed method in supressing noise from a real electromyographic (EMG) signal of a healthy person. The signal was taken from the

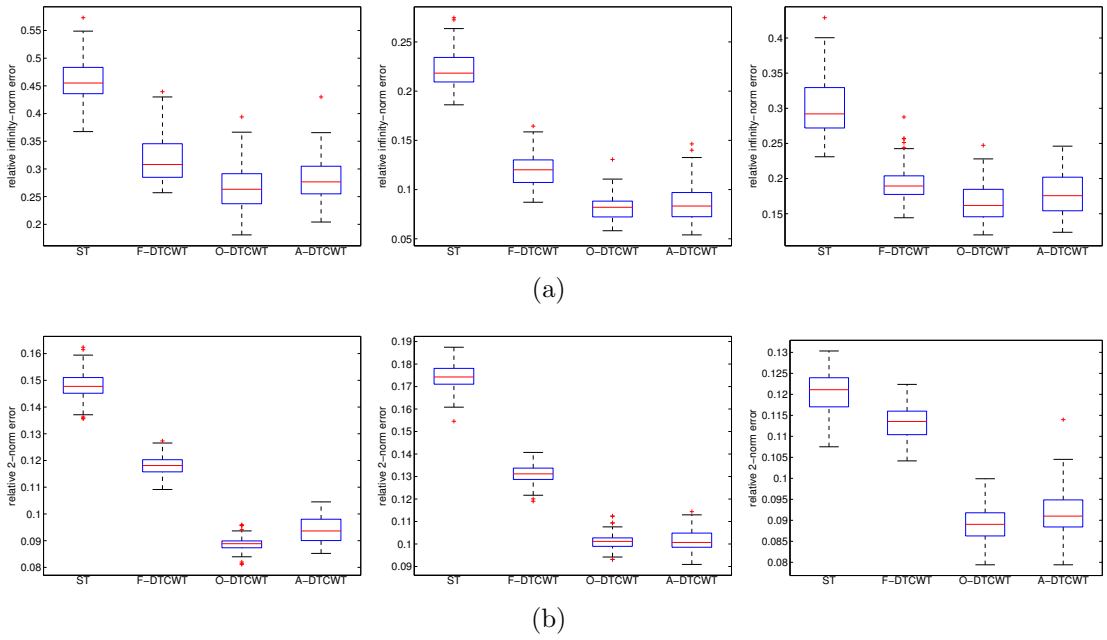


Figure 8: Box-plots of the obtained reconstruction errors over 100 runs of the experiment, for a noisy scenario with $\sigma^2 = 4$. (a) $\|\mathbf{z} - \widehat{\mathbf{z}}\|_\infty$ for for Blocks, Bumps and PP signal; (b) $\|\mathbf{z} - \widehat{\mathbf{z}}\|_2$ for Blocks, Bumps and PP signal.

Physionet data bank[3]. Uncorrelated white Gaussian noise of variance $\sigma_\eta^2 = 4$ was added to the standardized EMG. To aid visualization, only a segment cut at random from the whole signal was used in the experiment. It can be seen that ST oversmoothes the signal significantly, thus loosing many details. The estimated signal using F-DTCWT preserves more details than ST but a reduced dynamic range, especially in the peaks and bumps, is still evident. On the contrary, the A-DTCWT estimate better preserves details and dynamic range of the main features of the signal. For this example, the relative ℓ_2-norm of the reconstruction error is 0.421 for ST, 0.319 for F-DTCWT and 0.172 for A-DTCWT. For $\|\mathbf{z} - \widehat{\mathbf{z}}\|_\infty / \|\mathbf{z}\|_\infty$, obtained reasults for the shown example are 0.425 for ST, 0.289 for F-DTCWT and 0.169 for A-DTCWT. Both measures shown the superiority of the proposed method for this denoising task. To check that the obtained results are not a consequence of a favorable choice of fragment of the EMG signal, 100 replicates of the experiment were run, each with a segment of length $N = 1024$ cut at random from the

[3]http://physionet.org/physiobank/database/emgdb/

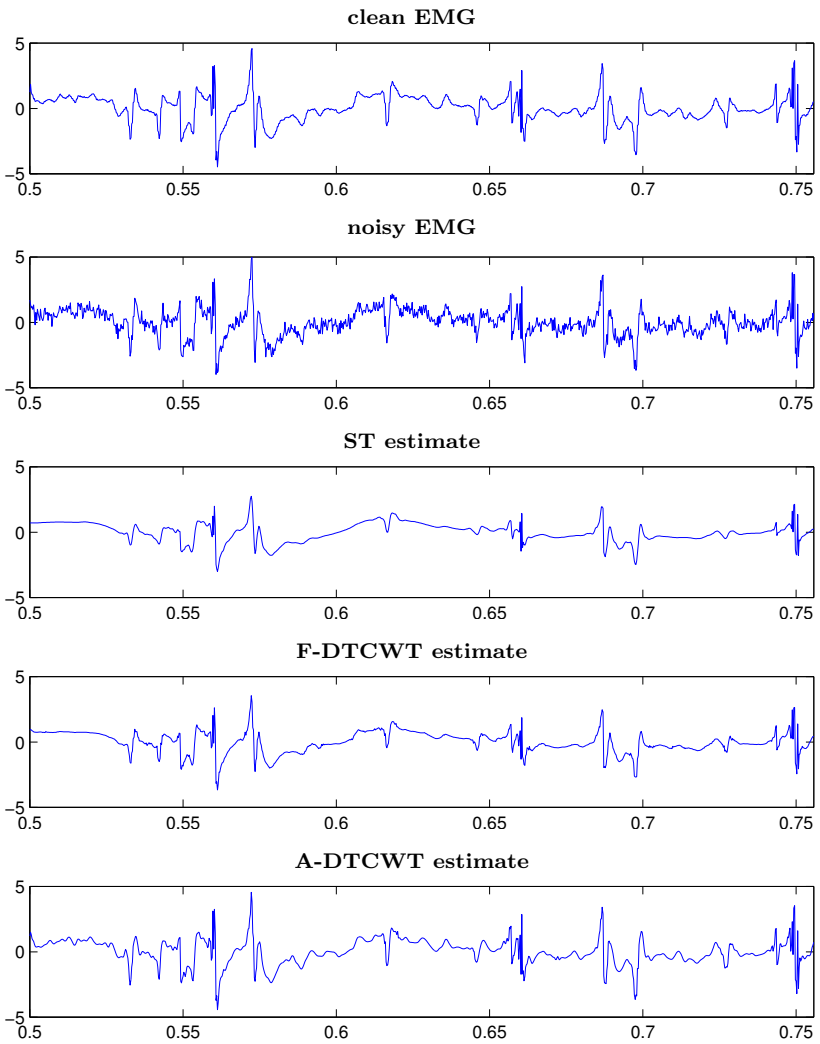


Figure 9: Example with real EMG signal

Table 2: Average denoising results for 100 fragments of EMG picked at random

$\sigma_\eta^2 = 4$	ST	F-DTCWT	A-DTCWT
Error measure	MEAN(SD)	MEAN(SD)	MEAN(SD)
$\|\widehat{\mathbf{z}} - \mathbf{z}\|_2 / \|\mathbf{z}\|_2$	0.347 (0.062)	0.260 (0.017)	0.173 (0.017)
$\|\widehat{\mathbf{z}} - \mathbf{z}\|_\infty / \|\mathbf{z}\|_\infty$	0.345 (0.059)	0.266 (0.054)	0.212 (0.076)

whole signal. Averaged results can be seen in Figure 2, showing the same trend as for the realization shown in Figure 9.

5. Conclusions and further work

In this work we have introduced an adaptive structured wavelet shrinkage estimator. In the proposed method, the weights in a hierarchical structured regularizer are modified in order to favour the retention of coefficients related to the locations of salient features in the signal. The detection of such locations is carried out using information extracted from the wavelet decomposition. Denoising experiments with synthetic and EMG signals showed that the adaptive scheme outperforms the non-adaptive structured estimators. These results encourage the extension of the adaptive scheme to images, in order to improve preservation of edges in denoising applications. Further extensions could also involve relaxing the hierarchical structure to allow for more general dictionaries to replace the DTCWT used in the present work. Alternatively a union-of-basis-framework could be introduced to catch different features in the signal.

Acknowledgements

Authors would like to thank Professor Nick Kingsbury for providing the code to compute the DTCWT. Most of this work was completed during a visit of D.T. to the Department of Statistical Science, University College London. D.T. was also supported by CONICET under grant PIP 11220110100742; ANPCyT under grants PICT 2012-2590 and 2011-2440; and UNL under grant CAI+D 201101-00062LI.

[1] D. L. Donoho, J. M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (3) (1994) 425–455.

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

[3] M. Jansen, Noise reduction by wavelet thresholding, Lecture notes in statistics, Springer, New York, 2001.

[4] L. Sendur, I. Selesnick, Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency, Signal Processing, IEEE Transactions on 50 (11) (2002) 2744–2756.

[5] M. Crouse, R. Nowak, R. Baraniuk, Wavelet-based statistical signal processing using hidden Markov models, Signal Processing, IEEE Transactions on 46 (4) (1998) 886–902.

[6] M. Vanucci, F. Corradi, Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (4) (1999) 971–986.

[7] A. M. Achim, Bivariate wavelet shrinkage using alpha-stable distributions, in: Optics & Photonics 2005, International Society for Optics and Photonics, 2005, pp. 59141J–59141J.

[8] L. Evers, T. J. Heaton, Locally adaptive tree-based thresholding, Journal of Computational and Graphical Statistics 18 (4) (2009) 961–977.

[9] S. Yin, L. Cao, Y. Ling, G. Jin, Image denoising with anisotropic bivariate shrinkage, Signal Processing 91 (8) (2011) 2078 – 2090.

[10] A. Gholami, S. M. Hosseini, A balanced combination of Tikhonov and total variation regularizations for reconstruction of piecewise-smooth signals, Signal Processing 93 (7) (2013) 1945 – 1960.

[11] T. F. Chan, H.-M. Zhou, Total variation wavelet thresholding, Journal of Scientific Computing 32 (2) (2007) 315–341.

[12] S. Durand, J. Froment, Reconstruction of wavelet coefficients using total variation minimization, SIAM Journal on Scientific Computing 24 (5) (2003) 1754–1767.

[13] P.-L. Dragotti, M. Vetterli, Wavelet footprints: theory, algorithms, and applications, Signal Processing, IEEE Transactions on 51 (5) (2003) 1306–1323.

[14] R. G. Baraniuk, Optimal tree approximation with wavelets, Vol. 3813, 1999, pp. 196–207.

[15] T. C. Lee, Tree-based wavelet regression for correlated data using the minimum description length principle, Australian and New Zealand Journal of Statistics 44 (1) (2002) 23–39.

[16] F. Autin, J.-M. Freyermuth, R. von Sachs, Ideal denoising within a family of tree-structured wavelet estimators, Electronic Journal of Statistics 5 (2011) 829–855.

[17] M. Elad, M. A. T. Figueiredo, Y. Ma, On the role of sparse and redundant representations in image processing, Proceedings of the IEEE 98 (6) (2010) 972–982.

[18] H. Cheng, Z. Liu, L. Yang, X. Chen, Sparse representation and learning in visual recognition: Theory and applications, Signal Processing 93 (6) (2013) 1408 – 1425.

[19] S. Chen, D. Donoho, M. Saunders, Atomic decomposition by basis pursuit, SIAM Journal on Scientific Computing 20 (1) (1998) 33–61.

[20] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, Signal Processing, IEEE Transactions on 41 (12) (1993) 3397–3415.

[21] D. Donoho, Compressed sensing, Information Theory, IEEE Transactions on 52 (4) (2006) 1289–1306.

[22] E. Candes, M. Wakin, An introduction to compressive sampling, Signal Processing Magazine, IEEE 25 (2) (2008) 21–30.

[23] R. Baraniuk, V. Cevher, M. Duarte, C. Hegde, Model-based compressive sensing, Information Theory, IEEE Transactions on 56 (4) (2010) 1982–2001.

[24] L. He, L. Carin, Exploiting structure in wavelet-based Bayesian compressive sensing, Signal Processing, IEEE Transactions on 57 (9) (2009) 3488–3497.

[25] C. Chen, J. Huang, Compressive sensing MRI with wavelet tree sparsity, in: NIPS, 2012, pp. 1124–1132.

[26] F. Bach, J. Jenatton, J. Mairal, G. Obozinski, Structured sparsity through convex optimization, Statistical Science 27 (4) (2012) 450–468.

[27] J. Jenatton, J. Audibert, F. Bach, Structured variable selection with sparsity-inducing norms, Journal of Machine Learning Research 12 (2011) 2777–2824.

[28] J. Huang, T. Zhang, D. Metaxas, Learning with structured sparsity, Journal of Machine Learning Research 12 (2011) 3371–3412.

[29] N. Rao, R. Nowak, S. Wright, N. Kingsbury, Convex approaches to model wavelet sparsity patterns, in: Image Processing (ICIP), 2011 18th IEEE International Conference on, 2011, pp. 1917–1920.

[30] P.-Y. Chen, I. W. Selesnick, Translation-invariant shrinkage/thresholding of group sparse signals, Signal Processing 94 (0) (2014) 476 – 489.

[31] C. Ni, Q. Li, L. Z. Xia, A novel method of infrared image denoising and edge enhancement, Signal Processing 88 (6) (2008) 1606 – 1614.

[32] J. Wu, F. Liu, L. Jiao, X. Wang, Multivariate pursuit image reconstruction using prior information beyond sparsity, Signal Processing 93 (6) (2013) 1662 – 1672.

[33] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B 58 (1994) 267–288.

[34] M. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso), Information Theory, IEEE Transactions on 55 (5) (2009) 2183–2202.

[35] P. Zhao, B. Yu, On model selection consistency of lasso, Journal of Machine Learning Research 7 (2006) 2541–2563.

[36] P. Bühlmann, S. van de Geer, Statistics for High-Dimensional Data, Springer Series in Statistics, Springer Berlin Heidelberg, 2011.

[37] T. Zhang, Some sharp performance bounds for least squares regression with L1 regularization, Annals of Statistics 37 (5A) (2009) 2109–2144.

[38] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, Pathwise coordinate optimization, Annals of Applied Statistics 1 (2007) 302–332.

[39] S. Wright, R. Nowak, M. A. T. Figueiredo, Sparse reconstruction by separable approximation, Signal Processing, IEEE Transactions on 57 (7) (2009) 2479–2493.

[40] M. Yuan, M. Yuan, Y. Lin, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society, Series B 68 (2006) 49–67.

[41] L. Jacob, G. Obozinski, J.-P. Vert, Group lasso with overlap and graph lasso, in: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, New York, NY, USA, 2009, pp. 433–440.

[42] G. Obozinski, L. Jacob, J.-P. Vert, Group lasso with overlaps: the latent group lasso approach, Tech. rep., submitted (2011).

[43] F. Bach, R. Jenatton, J. Mairal, Optimization with Sparsity-Inducing Penalties, Foundations and Trends in Machine Learning, Now Publishers Inc., Hanover, USA, 2011.

[44] L. Yuan, J. Liu, J. Ye, Efficient methods for overlapping group lasso, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (9) (2013) 2104–2116.

[45] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning 3 (1) (2011) 1–122.

[46] W. Deng, W. Yin, Y. Zhang, Group sparse optimization by alternating direction method, Tech. rep., Rice University CAAM Technical Report TR11-06 (2011).

[47] I. Selesnick, R. Baraniuk, N. Kingsbury, The dual-tree complex wavelet transform, Signal Processing Magazine, IEEE 22 (6) (2005) 123–151.

[48] N. Kingsbury, Complex wavelets for shift invariant analysis and filtering of signals, Applied and Computational Harmonic Analysis 10 (3) (2001) 234 – 253.

[49] I. Selesnick, The design of approximate Hilbert transform pairs of wavelet bases, Signal Processing, IEEE Transactions on 50 (5) (2002) 1144–1152.

[50] H. Shi, B. Hu, J. Q. Zhang, A novel scheme for the design of approximate Hilbert transform pairs of orthonormal wavelet bases, Signal Processing, IEEE Transactions on 56 (6) (2008) 2289–2297.

[51] D. B. H. Tay, M. Palaniswami, Design of approximate Hilbert transform pair of wavelets with exact symmetry, in: Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on, Vol. 2, 2004, pp. ii–921–4 vol.2.

[52] D. B. H. Tay, A new approach to the common-factor design technique for Hilbert-pair of wavelets, Signal Processing Letters, IEEE 17 (11) (2010) 969–972.

[53] N. Kingsbury, Design of q-shift complex wavelets for image processing using frequency domain energy minimization, in: Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, Vol. 1, 2003, pp. I–1013–16 vol.1.

[54] N. Kingsbury, A dual-tree complex wavelet transform with improved orthogonality and symmetry properties, in: Image Processing, 2000. Proceedings. 2000 International Conference on, Vol. 2, 2000, pp. 375–378 vol.2.

[55] X. Zhang, H. Morihara, Design of q-shift filters with flat group delay, in: Circuits and Systems (ISCAS), 2012 IEEE International Symposium on, 2012, pp. 2337–2340.

[56] K. Chaudhury, M. Unser, On the shiftability of dual-tree complex wavelet transforms, Signal Processing, IEEE Transactions on 58 (1) (2010) 221–232.

[57] J. Romberg, H. Choi, R. Baraniuk, Multiscale edge grammars for complex wavelet transforms, in: Image Processing, 2001. Proceedings. 2001 International Conference on, Vol. 1, 2001, pp. 614–617 vol.1.

[58] H. Choi, J. Romberg, R. Baraniuk, N. Kingsbury, Hidden Markov tree modeling of complex wavelet transforms, in: Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on, Vol. 1, 2000, pp. 133–136 vol.1.

[59] R. Jenatton, J. Mairal, G. Obozinski, F. Bach, Proximal methods for hierarchical sparse coding, Journal of Machine Learning Research 12 (2011) 2297–2334.

[60] P. Zhao, G. Rocha, B. Yu, The composite absolute penalties family for grouped and hierarchical variable selection, The Annals of Statistics 37 (6A) (2009) 3468–3497.

[61] L. Demaret, P. Massopust, M. Storath, Signal analysis based on complex wavelet signs, Preprint, http://arxiv.org/abs/1208.4578.

[62] S. Mallat, A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way, 3rd Edition, Academic Press, 2008.

[63] J. Fauqueur, N. G. Kingsbury, R. Anderson, Multiscale keypoint detection using the dual-tree complex wavelet transform., in: International Conference on Image Processing, IEEE, 2006, pp. 1625–1628.

[64] J. Liu, J. Ye, Moreau-Yosida regularization for grouped tree structure learning, in: NIPS, 2010, pp. 1459–1467.

[65] S. N. Negahban, P. Ravikumar, M. J. Wainwright, B. Yu, A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers, Statistical Science 27 (4) (2012) 538–557.

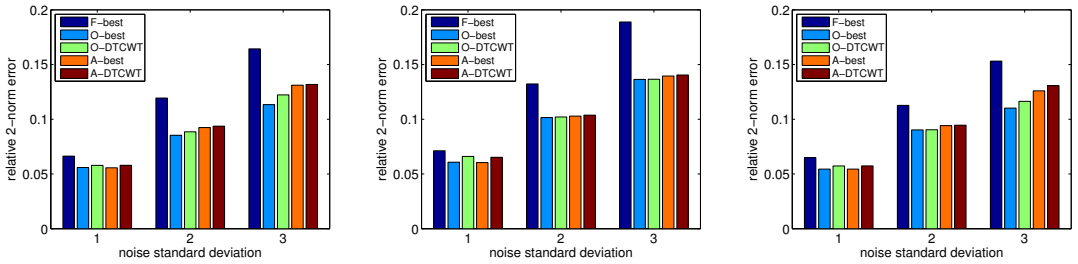
[66] J. Buckheit, D. Donoho, Wavelab and reproducible research, Springer-Verlag, 1995, pp. 55–81.

[67] J. Liu, S. Ji, J. Ye, SLEP: Sparse Learning with Efficient Projections, Arizona State University (2009).
URL `http://www.public.asu.edu/ jye02/Software/SLEP`

AppendixA. Influence of the selection of λ

The goodness of the proposed procedure to select the value of the regularization parameter λ is assessed. The reconstruction error obtained with the proposed λ is compared to the minimum reconstruction error achieved using a fine grid of values of λ. The rest of the parameters were set as described in Section 4.3. The ℓ_2-norm of the error is used for comparison. Results of 50 experiments were averaged to get the reported results. A set of 1000 uniformly spaced samples of λ in the interval $[0, \lambda_{max}]$ were tried out for each method at each run. λ_{max} is such that for $\lambda > \lambda_{max} > 0$, the resulting estimate is a zero vector. λ_{max} was estimated in each case using the Algorithm proposed in [64].

Obtained results are shown in Figure A.10. O-best and A-best denote results corresponding to the best choice of the regularization parameter when using the true location of the salient features or estimated ones, respectively. For methods O-best and A-best, only the best estimate across the differente values of λ was picked as the result corresponding to the run. Results for the non-adaptive tree-grouped estimator is also included for comparison. It can be seen that the difference in performance between the best selection of λ and the one proposed here is always significantly smaller than the difference in performance between the non-adaptive alternative F-best and the adaptive one A-DTCWT. This result shows the advantage of adaptation beyond the fine tunning of the regularization parameter. Furthermore, for methods A-DTCWT and A-best that do not use oracle information about the location of salient features, it can be seen that the difference in performance between them is smaller than for O-best versus O-DTCWT. When noise increases, it can be seen also that the difference in performance between A-best and A-DTCWT is indeed smaller than the difference between O-best and A-best. This suggests that the performance of the denoising method is more sensitive to the correct detection of salient features than to the deviation of the proposed value for λ from its optimal choice. Moreover, the overall performance of the adaptive method due to practical implementation constraints is still significantly better than the performance of tree-structured estimators with the regularizer fixed a priori. These results indicate that, at a very modest cost to simplicity, the suggested procedure for selecting λ is suitable for



(a) (b) (c)

Figure A.10: Performance of the introduced denoising algorithm using the proposed selection method for the regularization parameter λ, compared with the best achievable performance estimated using a fine grid of 1000 values of λ in the interval $(0, \lambda_{max})$. (a) Blocks; (b) Bumps; (c) PP.

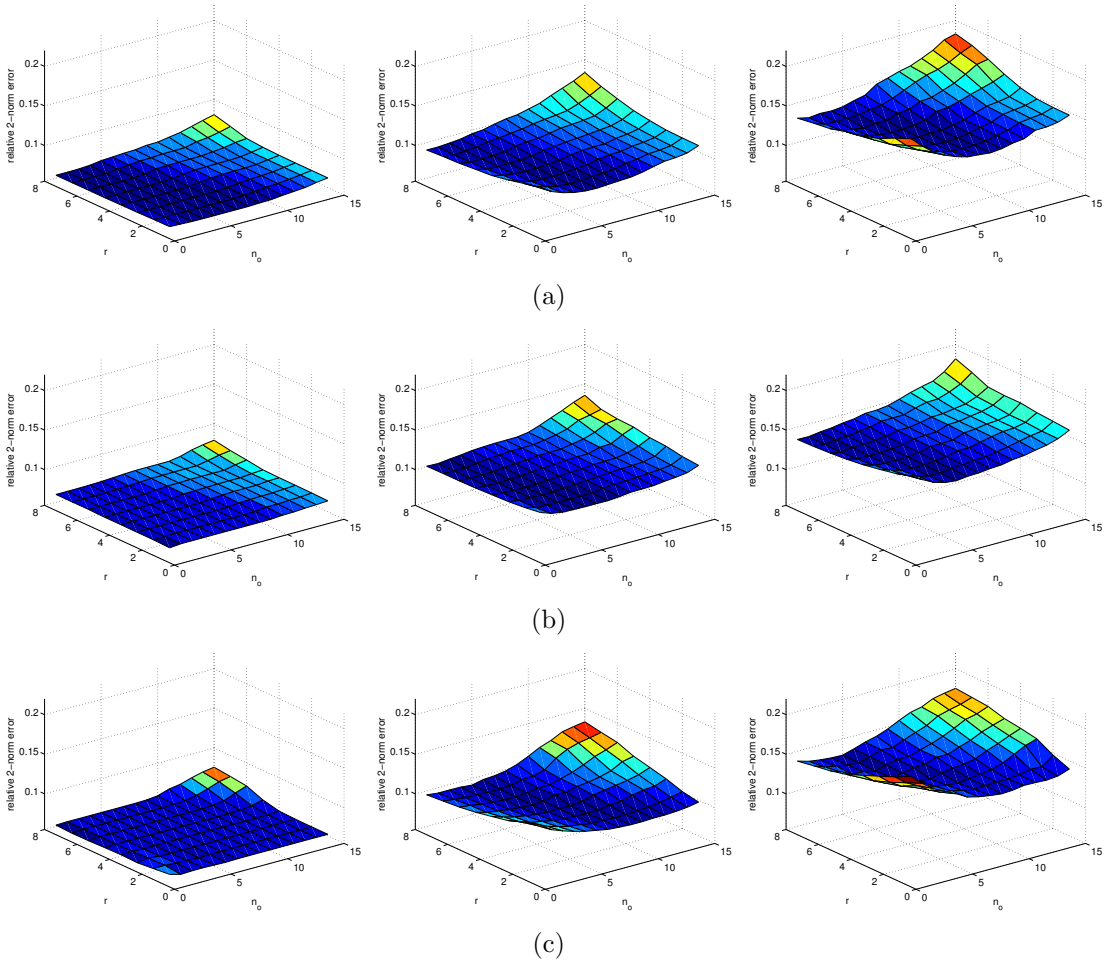


Figure B.11: Influence of parameters for detection of salient features on the overall proposed denoising method. (a) Blocks; (b) Bumps; (c) PP. For each signal, noise variance increases from left to right.

practical applications.

AppendixB. Influence of the parameters r and n_0

The effect of the parameters (r, n_0) on the accuracy of the detection stage of the method was assessed in Section 4.2. Here, the influence of these parameters on the final outcome of Algorithm 3 is addressed. A simulation was run using the synthetic signals considered in Section 4, corrupted with uncorrelated white Gaussian noise. Results of 50 experiments were used to get the reported results. For each run, the regularization parameter was set as discussed in Section 3.3.2 and a grid of values (r_i, n_j) was evaluated

for the detection stage. For each pair of parameters values, the resulting reconstruction error measured in the ℓ_2-norm was stored.

Averaged results over the 50 runs are shown in Figure B.11. It can be seen that for all signals and noise levels, denoising performance suffers for large values of the persistence parameter n_0 and radius r. These large values accounts for situations where only large isolated peaks in $\delta(n)$ are detected increasing the number of false negatives. When the noise is weak, there is little effect of the choice of (r, n_0) on the denoising result; a wide range of values of the parameters, except the largest ones, obtain very similar reconstruction errors. When the signal to noise ratio becomes small, as in the situation in the plots on the right of Figure B.11, small values of n_0 and r also affects negatively the denoising performance. This is expected, since this leads to an increase in the false positive rate of the detection stage, meaning that noisy coefficients are forced to be preserved during shrinkage. Nevertheless, figure shows that there is a wide range of values in the middle of the grid so that, for a fixed choice around $(r = 4, n_0 = 6)$, the performance is good and near the best one regardless of the signal and noise level.