mathematics

MDPI

Article

Anomaly Detection in Multichannel Data Using Sparse Representation in RADWT Frames

Daniela De Canditiis [1,*] and Italia De Feis [2]

[1] Istituto per le Applicazioni del Calcolo, CNR-Rome, 00185 Rome, Italy
[2] Istituto per le Applicazioni del Calcolo, CNR-Naples, 80131 Naples, Italy; i.defeis@iac.cnr.it
* Correspondence: d.decanditiis@iac.cnr.it

Abstract: We introduce a new methodology for anomaly detection (AD) in multichannel fast oscillating signals based on nonparametric penalized regression. Assuming the signals share similar shapes and characteristics, the estimation procedures are based on the use of the Rational-Dilation Wavelet Transform (RADWT), equipped with a tunable Q-factor able to provide sparse representations of functions with different oscillations persistence. Under the standard hypothesis of Gaussian additive noise, we model the signals by the RADWT and the anomalies as additive in each signal. Then we perform AD imposing a double penalty on the multiple regression model we obtained, promoting group sparsity both on the regression coefficients and on the anomalies. The first constraint preserves a common structure on the underlying signal components; the second one aims to identify the presence/absence of anomalies. Numerical experiments show the performance of the proposed method in different synthetic scenarios as well as in a real case.

Keywords: anomaly detection; RADWT; variable selection; multichannel; thresholding

Citation: De Canditiis, D.; De Feis, I. Anomaly Detection in Multichannel Data Using Sparse Representation in RADWT Frames. *Mathematics* **2021**, *9*, 1288. https://doi.org/10.3390/math9111288

Academic Editors: Francesca Pitolli and Annalisa Pascarella

Received: 20 April 2021
Accepted: 29 May 2021
Published: 3 June 2021

1. Introduction

Anomaly detection (AD) is an important problem that has received much attention in recent years. It consists of establishing whether a given data deviates from nominal shape or form. It is not possible to establish a single mathematical framework to deal with AD because it depends on the type of application. There are a lot of surveys in the literature, some of them specific to certain applications and some others are broader and application-free. See, for example, the review by [1] regarding the Intrusion Detection Systems, or the paper by [2] providing an overview on AD, analysis, and prediction in Internet of Things (IoT) systems or the paper by [3] considering the application of AD techniques to aviation and how they improve the safety and service of flight operations, just to cite some. However, a milestone of the literature is the paper by Chandola et al. [4], where the authors offer a very good understanding of the subject, defining anomalies, describing its challenges, depicting a relevant taxonomy of the different techniques, and illustrating the various domains of applications. As well explained by the authors, the AD problem depends on the nature of input data (points, sequence, functions, graphs, objects of different nature), on the type of anomaly (point anomalies, contextual or behavioral anomalies, or their combination), on the availability of labeled data for training/validation of the AD techniques (leading to unsupervised AD and supervised AD), and on the type of output of AD (scores or label). Two more recent reviews are the ones by Pimentel et al. [5] and by Ahmed et al. [6], where the authors display a huge and comprehensive number of references on the topic. In particular, reference [5] is devoted to novelty detection, where the distinction with the word anomaly is that the former considers, as normal, the data after their detection. Another important and very recent paper is the one by Thudumu et al. [7], which reviews AD in the context of big data where a curse of dimensionality occurs, bringing the failure of different methodologies.

We can broadly divide the AD techniques into four categories as very well represented by the taxonomy of Figure 3 in [6]: classification techniques, statistical techniques, clustering techniques, and information theory techniques.

In this paper, we focus on statistical techniques, whose recipe is to fit a statistical model to the given data and then check if some of them do not fit this model. We think that a good survey on statistical methods is given by Rousseeuw and Hubert in [8], where they consider robust regression techniques, signal processing techniques, and robust principal component analysis (PCA) techniques.

Our paper is placed in the context of statistical signal processing techniques, where the data are indeed time series, signals, or functions. It is worth observing that signal processing is synonymous with functional data analysis in more classical statistics nomenclature, as well as outliers and anomalies. In particular, we can distinguish between isolated and persistent outliers, see [9]. The isolated outliers insist on a small part of the signal, while the persistent outliers insist on most of their domain. It is important to stress that the literature on AD in functional data is quite recent and is mainly devoted to univariate functions.

In this paper, we deal with the AD problem for multichannel (i.e., multivariate) data affected by persistent outliers and approach it from the perspective of penalized multiple regression framework. We model the nominal shape as K multiple signals from K channels sharing a joint sparse representation into a special dictionary, and then we add a possible anomaly term of any form and shape to each signal component. This framework represents situations where we expect similar shapes for the components of the signal, but some components can undergo anomaly behavior for some unpredictable reason. In particular, we consider nominal signals with a fast oscillating characteristic that can be well represented only by using an appropriate dictionary, namely a RADWT, as done in the context of multiple regression in [10]. A RADWT is a modern and fast computational tool for analyzing signals which are a mixture of oscillatory and nonoscillatory transient behaviors. Such kinds of signals are not periodic and may describe many physiological and physical signals (for example, speech, stock-market, biomedical EEG, etc.) which could not be represented sparsely in the classic orthogonal bases such as Fourier, wavelet, etc. The request of joint sparse representation formalizes the hypothesis that the signal components are expected to have the same spectral characteristics while preserving numerical differences. The idea we develop in our work is the following: a linear model is applied to simultaneously estimate the signal by using its sparse representation in this special dictionary and detect potential anomalies modeled as residuals in this sparse representation.

A similar idea has been already applied in the literature under different model assumptions, for example in [11–13]. Specifically, in [11] the authors propose to learn the dictionary by using a set of nominal signals through PCA, as well as in [12] the authors learn the dictionary by using a training set of "normal" nominal signals, i.e., not affected by anomalies, under the specific hypothesis that they are multivariate mixtures of discrete and continuous signals and applying their sparse coding algorithm to select the training signals having the highest residuals. Subsequently, in both papers, the authors detect anomalies for a new signal occurrence by evaluating the residual of its sparse representation in this dictionary. On the other hand, in [13], the authors face the univariate AD problem considering a linear regression model where the univariate signal is expressed as a linear combination of columns of a given design matrix without the sparseness hypothesis. The authors propose detecting anomalies by simultaneously estimating both the nominal signal and its residual using nonconvex penalized regression with the constraint on the outliers vector.

Our paper stands in between these recent proposals [11–13] for two reasons. First, we use a dictionary for sparsely representing functions (as in [11,12]) but without requiring a training set of "normal" nominal signals for dictionary learning because we rely on an innovative instrument such as the RADWT transform to deal with fast oscillating signals. Second, we look for anomalies by simultaneously estimating both the nominal signal and its residual (as in [13]) but imposing a double constraint on the regression coefficients and the anomalies, taking into account that few RADWT coefficients are necessary to

represent the signals. To ensure that the residuals can preserve information about the outliers/anomalies, we must use robust regression techniques. As we will clarify later on, a good AD undergoes a good robust regression analysis, because the procedure aims simultaneously to detect the anomalies and to estimate the signal's components, the success of the first goal ensuring the success of the second and vice versa. Hence, our paper can be also considered as part of the literature on robust regression in the high-dimensional setting, see [14,15]. In this perspective, we also refer to [16], from which we were inspired to use robust losses, such as the Huber loss and the skipped mean loss.

The remainder of the paper is organized as follows. Section 2 describes the data model with the working hypothesis. Section 3 presents and discusses the proposed AD procedure within the paradigm of group-lasso regression, enlightening the connections with other existing procedures. Section 4 describes some important implementation issues. Section 5 shows numerical experiments on synthetic and real data and finally, the last section discusses conclusions.

2. The Data Model and Problem Setting

We will consider hereafter a functional dataset consisting of K curves $f^{(k)}(t)$, $k = 1, \ldots, K$ observed on a set of equispaced grid points $t_1, \ldots, t_n$ and resulting in observed K dimensional response vectors $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)}$, $\ldots$,$\mathbf{y}^{(K)}$. The observed curves are composed by nominal signals $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$,... $\mathbf{f}^{(K)}$ which may have some functional properties to which are eventually added anomaly signals $\mathbf{a}^{(1)}$, $\mathbf{a}^{(2)}$,..., $\mathbf{a}^{(K)}$ which may appear as functional outliers $a^k(t)$, $k = 1, \ldots, K$. Once the data is discretized our data model becomes

$$\begin{cases} \mathbf{y}^{(1)} &= \mathbf{f}^{(1)} + \mathbf{a}^{(1)} + \boldsymbol{\varepsilon}^{(1)} \\ \mathbf{y}^{(2)} &= \mathbf{f}^{(2)} + \mathbf{a}^{(2)} + \boldsymbol{\varepsilon}^{(2)} \\ & \vdots \\ \mathbf{y}^{(K)} &= \mathbf{f}^{(K)} + \mathbf{a}^{(K)} + \boldsymbol{\varepsilon}^{(K)} \end{cases} \tag{1}$$

where, for each $k = 1, \cdots, K$, $\mathbf{f}^{(k)} \in \mathbb{R}^{n\times 1}$ is the underlying unknown nominal signal, $\mathbf{a}^{(k)} \in \mathbb{R}^{n\times 1}$ is the potential anomaly, and $\boldsymbol{\varepsilon}^{(k)} \sim N(\mathbf{0}, \sigma^2 I_n)$ is the additive white noise. Note that if the Gaussian noise in some channel is correlated with known covariance structure, we can easily manage this situation reparametrizing the data.

The mathematical hypothesis is that, while the K components have a nominal behavior which results in a joint sparse representation into a RADWT dictionary, the anomalies can be of any shape, so they have no sparse representation in the same dictionary, and they are independent of each other, so an anomaly can be detected in any, some, or all components.

We do not hypothesize functions $f^{(k)}(t)$ belong to some functional Sobolev space $H^s_{p,q}[a, b]$ as it is usually done in functional nonparametric regression setting, instead we let these functions be much more general and we restrict our attention to their finite-dimensional representation. Since many physiological and physical signals exhibit a mixture of oscillatory and nonoscillatory behaviors (for example, speech, stock-market, biomedical, EEG, etc.), we suppose that each component is a "high resonance" signal or a "low resonance" signal or a sum of both types of signal. By a high-resonance component, we mean a signal consisting of multiple simultaneous sustained oscillations; in contrast, by a low-resonance component, we mean a signal consisting of nonoscillatory transients of unspecified shape and duration. We stress that the high and low resonance components of a signal can not be extracted from its high and low frequencies components in a time-scale decomposition, but they can be well represented by a high-Q factor RADWT and a low-Q factor RADWT respectively, see [17]. The RADWT is a normalized tight dictionary of $L_2(R)$ defined as $\left\{ (\frac{q}{p})^{k/2} \psi\left((\frac{q}{p})^k t + \frac{sp}{q} l \right) \right\}_{k,l\in Z}$ where ψ is a wavelet function and (p, q, s) is a triplet of parameters that gives the time-scale characteristic of the dictionary. In particular, the ratio $q/p > 1$ is closely related to the scale (or frequency) dilatation factor, the parameter s is closely related to the time dilatation factor, and $\frac{p}{s(q-p)}$ is the redundant factor. The Q-

factor depends on these parameters although there is not an explicit formula, in particular setting the dilatation factor q/p between 1 and 2 and $s > 1$ gives a RADWT with high Q-factor, while setting $s = 1$ we obtain a low Q-factor RADWT with time-scale characteristic similar to the dyadic wavelet transform. In particular, when $q = 2$, $p = 1$ and $s = 1$, the dictionary reduces to the classical wavelet basis. Given a finite energy signal $\mathbf{x}$ of length n and $J \in \mathbb{N}$ levels of decomposition, the RADWT transform is obtained by a sequence of proper downsampling operations and fast Fourier transforms; it ends up with $\lceil \frac{np^J}{q^J} \rceil$ scaling coefficients (low-pass filtering) and $\lceil \frac{np^j}{q^j s} \rceil$ wavelet coefficients (high-pass filtering) at each level $j = 1, ..J$. See [18] for details on fast analysis and synthesis schemes.

In the following, we will use these signal processing results in order to formulate our working setup. Let $\mathbf{\Psi} \in \mathbb{R}^{n \times d1}$ be the finite matrix representation of the low Q-factor analysis filter, where $d1$ indicates the cardinality of the low Q-factor RADWT, and let $\mathbf{\Phi} \in \mathbb{R}^{n \times d2}$ be the finite matrix representation of the high Q-factor analysis filter (the synthesis operators being just the transpose matrices), where $d2$ indicates the cardinality of the high Q-factor RADWT. Let us define the dictionary $\boldsymbol{W} = [\mathbf{\Psi}\ \mathbf{\Phi}] \in \mathbb{R}^{n \times d}$, with $d = d1 + d2$, then our working hypothesis is the following:

(H1) signals $\mathbf{f}^{(k)}$ have a **jointly** sparse representation in $\boldsymbol{W}$, i.e., setting $\mathbf{f}^{(k)} = \boldsymbol{W}\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\beta}^{(k)} \in \mathbb{R}^{d \times 1}$, the coefficients matrix

$$\mathbf{B} = \left[\begin{array}{cccc} \boldsymbol{\beta}^{(1)}, & \boldsymbol{\beta}^{(2)}, & \cdots, & \boldsymbol{\beta}^{(K)} \end{array} \right] \in \mathbb{R}^{d \times K}$$

has a minimal number of rows different from zero. This means considering the following constraint

$$\|\mathbf{B}\|_{0,2} = \sum_{i=1}^{d} \mathcal{I}(\|\mathbf{b}_{i,.}\|_2) = \sum_{i=1}^{d} \mathcal{I}\left(\sqrt{\sum_{k=1}^{K} \left(\beta_i^{(k)}\right)^2} \right) \leq T \tag{2}$$

where $\|\mathbf{B}\|_{2,0}$ counts the number of non zero rows, T is the a priori parameter indicating the expected degree of sparsity level, and $\mathcal{I}(a)$ is the indicator function:

$$\mathcal{I}(a) = \begin{cases} 1 & \text{if} \quad x \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

It is worth observing that the role of dictionary matrix $\boldsymbol{W}$ could be played by any other dictionary in which the nominal signals have a sparse joint representation. What we will expose in the following is not related to the nature of the dictionary $\boldsymbol{W}$ and can therefore be considered a valid method for other types of signals as long as there is a dictionary that sparsely represents them. For example, $\boldsymbol{W}$ could be made up of only $\mathbf{\Phi}$ if it is known that the f signals have only a high resonance component, or it could be made up of the union of other types of basis, frames, or dictionaries. However, when $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are the finite representation of RADWT, they constitute thigh frames, i.e., $\mathbf{\Psi}\ \mathbf{\Psi}' = \boldsymbol{I}_n$ and $\mathbf{\Phi}\ \mathbf{\Phi}' = \boldsymbol{I}_n$, so $\boldsymbol{W}\ \boldsymbol{W}' = 2\boldsymbol{I}_n$ and then, once we obtain the estimate of the coefficients it is straightforward to obtain the estimate of the signal. When the choice of the dictionary change, $\mathbf{\Psi}\ \mathbf{\Psi}' = \boldsymbol{I}_n$ and $\mathbf{\Phi}\ \mathbf{\Phi}' = \boldsymbol{I}_n$ is not necessarily satisfied, and hence signal reconstruction is not obvious.

3. Proposed Procedures

The linear model in (1) can be rewritten in terms of dictionary coefficients as follows

$$\begin{cases} \mathbf{y}^{(1)} &= \boldsymbol{W}\boldsymbol{\beta}^{(1)} + \mathbf{a}^{(1)} + \boldsymbol{\varepsilon}^{(1)} \\ \mathbf{y}^{(2)} &= \boldsymbol{W}\boldsymbol{\beta}^{(2)} + \mathbf{a}^{(2)} + \boldsymbol{\varepsilon}^{(2)} \\ & \vdots \\ \mathbf{y}^{(K)} &= \boldsymbol{W}\boldsymbol{\beta}^{(K)} + \mathbf{a}^{(K)} + \boldsymbol{\varepsilon}^{(K)} \end{cases} \tag{4}$$

which turns out to be a linear multiple regression model with a special common design matrix. A first and somewhat naive approach, but sub-optimal, would consist of treating each channel separately, ignoring the underlying common structure. This is the reason why such kind of problem is reformulated in terms of a unique regression problem in the following form:

$$\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \vdots \\ \mathbf{y}^{(K)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{W} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{W} & \cdots & \mathbf{0} \\ \cdots & & & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{W} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \\ \vdots \\ \boldsymbol{\beta}^{(K)} \end{bmatrix} + \begin{bmatrix} \mathbf{a}^{(1)} \\ \mathbf{a}^{(2)} \\ \vdots \\ \mathbf{a}^{(K)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}^{(1)} \\ \boldsymbol{\varepsilon}^{(2)} \\ \vdots \\ \boldsymbol{\varepsilon}^{(K)} \end{bmatrix} \tag{5}$$

$$\mathbf{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \mathbf{a} + \boldsymbol{\varepsilon}$$

with obvious correspondence between elements of the two expressions. So, $\mathbf{y}$ is a column vector of $n \cdot K$ response variables, $\boldsymbol{X}$ a design matrix of dimension $n \cdot K \times K \cdot d$, $\boldsymbol{\beta}$ an unknown regression coefficients column vector of length $K \cdot d$, $\mathbf{a}$ is an unknown anomalies column vector of length $n \cdot K$, and, finally without loss of generality, we let $\boldsymbol{\varepsilon}$ be a $n \cdot K$-variate Gaussian random column vector with zero mean and covariance matrix $\sigma^2 \boldsymbol{I}_{n \cdot K}$. Our regression problem (5) has $d \cdot K + n \cdot K$ regression parameters and only $n \cdot K$ data point, so it clearly falls into the class of high-dimensional regression problems. Under the working hypothesis (H1), we expect $\boldsymbol{\beta}$ to be group sparse, i.e., for many $j = 1, ..., d$ we expect that $\beta_j^{(k)} = 0$, for all $k = 1, .., K$. This provides the following nonoverlapping group structure for the whole vector

$$\{1, 2..., K \cdot d\} = G_1 \cup \cdots \cup G_d, \tag{6}$$

with

$$G_j = \{j, j + d, j + 2d \ldots, j + (K - 1)d\}, j = 1, \ldots, d$$

group of size K. Let us define the following $l_{2,1}$ norm

$$\|\boldsymbol{\beta}\|_{2,1} = \sum_{j=1}^{d} \left\| \boldsymbol{\beta}_{G_j} \right\|_2 \quad \text{and} \quad \|\mathbf{a}\|_{2,1} = \sum_{j=1}^{K} \left\| \mathbf{a}^{(k)} \right\|_2$$

with $\boldsymbol{\beta}_{G_j}$ denoting the reduction of vector $\boldsymbol{\beta}$ to the subset of index G_j. The problem we are dealing with is not a classic regression problem, because our interest is to simultaneously estimate the vector $\mathbf{a}$, more specifically the $\mathbf{a}^{(1)}$, .., $\mathbf{a}^{(K)}$ anomalies that compose it, and the $\boldsymbol{\beta}$ vector. For example, if an anomaly is present in the first channel, the vector $\mathbf{a}^{(1)}$, representing the residual of the linear regression of $\mathbf{y}^{(1)}$ on $\boldsymbol{W}$ will have a l_2-norm much higher than the residual in another channel with no anomaly. While the anomalies vectors are not connected, because the anomalies can appear in any, some, or all channels, the underlying signals are not independent since by hypothesis they share the same spectral characteristic. For these reasons, we propose solving the following convex minimization problem

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{a}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{K\cdot d\times 1},\, \mathbf{a}\in\mathbb{R}^{K\cdot n\times 1}}{argmin} \underbrace{\frac{1}{2}\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta} - \mathbf{a}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_{2,1} + \lambda_2\|\mathbf{a}\|_{2,1}}_{F(\boldsymbol{\beta},\mathbf{a})} \tag{7}$$

where $\lambda_1 > 0$ is a regularization parameter that controls the group sparsity of vector $\boldsymbol{\beta}$, while $\lambda_2 > 0$ is a regularization parameter that controls the number of expected anomalies. Both the regularization parameters represent a trade-off between the fit and the complexity of the model. Being λ_2 fixed, $\lambda_1 \to 0$ implies a non sparse $\boldsymbol{\beta}$, generalizing the model proposed in [13] to the multichannel setting, on the contrary $\lambda_1 \to \infty$ implies a sparser $\boldsymbol{\beta}$ vector, increasing the energy of the anomalies. On the other hand, being λ_1 fixed, $\lambda_2 \to 0$ detects anomalies in each channel, while $\lambda_2 \to \infty$ brings to a model without anomalies. Their choice is a delicate point and we will discuss it in the Implementation, Section 4. Finally, for each $k \in \{1, 2, ..., K\}$ we detect an anomaly if the obtained estimate $\left\|\hat{\mathbf{a}}^{(k)}\right\|_2 \neq 0$, where $\hat{\mathbf{a}} = \left(\hat{\mathbf{a}}^{(1)}, ..., \hat{\mathbf{a}}^{(K)}\right)$ is the solution of the optimization problem (7). Moreover, the intensity of the detected anomaly, measured as $\|\hat{\mathbf{a}}^{(k)}\|_2$, can be considered as a score for the anomaly and treated as an incipient level of fault according to the specific application which defines an alarm threshold.

From a statistical perspective, the problem given in Equation (7) falls into the class of group lasso regression problem applied to the augmented parameter vector $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}^t; \mathbf{a}^t)^t$ and augmented design matrix $\widetilde{\boldsymbol{X}} = [\boldsymbol{X}\ I_{n\cdot K}]$. Indeed, the problem in Equation (7) can be rewritten as

$$\hat{\widetilde{\boldsymbol{\beta}}} = \underset{\widetilde{\boldsymbol{\beta}}\in\mathbb{R}^{K\cdot d+K\cdot n\times 1}}{argmin} \frac{1}{2}\|\mathbf{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}\|_2^2 + \lambda_1\mathcal{P}(\widetilde{\boldsymbol{\beta}}).$$

with group Lasso penalty $\mathcal{P}(\widetilde{\boldsymbol{\beta}}) = \sum_{j=1}^{d}\left\|\boldsymbol{\beta}_{G_j}\right\|_2 + \frac{\lambda_2}{\lambda_1}\sum_{j=1}^{K}\left\|\mathbf{a}^{(k)}\right\|_2$. If the unknown vector $\widetilde{\boldsymbol{\beta}}$ is sparse, under appropriate hypothesis on the design matrix $\widetilde{\boldsymbol{X}}$, consistency of the proposed estimator is guaranteed, see [19–21].

From a numerical perspective, problem (7) is jointly convex in $\boldsymbol{\beta}$ and $\mathbf{a}$, and its simple form suggests that we can apply an alternating optimization given initial estimates for $\boldsymbol{\beta}$ and $\mathbf{a}$. Of course, we need an estimate for $\boldsymbol{\beta}$ that takes into account the potential presence of the anomaly, i.e., robust.

Given an initial robust estimate for $\boldsymbol{\beta}$, i.e., $\hat{\boldsymbol{\beta}}^0$, we define $\hat{\mathbf{a}}^0 = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^0$, then at each iteration the procedure solves the following two sub-problems,

given $\mathbf{a}$, estimate $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{Kd}}{argmin} \underbrace{\left\{\frac{1}{2}||\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta} - \mathbf{a}||_2^2 + \lambda_1||\boldsymbol{\beta}||_{2,1} + \lambda_2\|\mathbf{a}\|_{2,1}\right\}}_{F_2(\boldsymbol{\beta})} \tag{8}$$

given $\boldsymbol{\beta}$, estimate $\mathbf{a}$

$$\hat{\mathbf{a}} = \underset{\mathbf{a}\in\mathbb{R}^{Kn}}{argmin} \underbrace{\left\{\frac{1}{2}||\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta} - \mathbf{a}||_2^2 + \lambda_1||\boldsymbol{\beta}||_{2,1} + \lambda_2\|\mathbf{a}\|_{2,1}\right\}}_{F_1(\mathbf{a})}. \tag{9}$$

Iteration stops when a given number of maximum iteration is reached or when the relative error on the anomaly vector $\mathbf{a}$ is below a fixed threshold. Let us give a closer look at the two sub-problems. Function $F_1(\mathbf{a}) = F_1\left(\mathbf{a}^{(1)}, ..., \mathbf{a}^{(K)}\right)$ is separable in each channel, i.e.,

$$F_1(\mathbf{a}) = \sum_{k=1}^{K} \frac{1}{2}\left\|\mathbf{r}^{(k)} - \mathbf{a}^{(k)}\right\|_2^2 + \lambda_2\left\|\mathbf{a}^{(k)}\right\|_2 + const,$$

where $\mathbf{r}^{(k)} = \mathbf{y}^{(k)} - \boldsymbol{W}\boldsymbol{\beta}^{(k)}$ is the current residual of the regression of $\mathbf{y}^{(k)}$ on $\boldsymbol{W}$; so the solution of sub-problem (9) is obtained by the multivariate-Soft thresholding rule

$$\hat{\mathbf{a}}^{(k)} = \frac{\mathbf{r}^{(k)}}{\left\|\mathbf{r}^{(k)}\right\|_2}\left(\left\|\mathbf{r}^{(k)}\right\|_2 - \lambda_2\right)_+, \tag{10}$$

where the directional vector $v/||v||$ plays the rule of $sign()$ function in classical one dimensional setting, see [22]. In Equation (10), the Soft thresholding operator $(.)_+$ acts on the vector $\mathbf{r}^{(k)}$ by shortening it towards 0 by an amount λ_2 if its norm is greater than λ_2, by setting it to zero if its norm is less than λ_2. For the seek of completeness, we sketch below how to derive the above statement. Vector in Equation (10) is the solution of the following convex problem

$$\hat{\mathbf{a}}^{(k)} = \underset{\mathbf{a}\in\mathbb{R}^{n\times 1}}{argmin}\left\{\sum_{k=1}^{k} \frac{1}{2}\left\|\mathbf{r}^{(k)} - \mathbf{a}^{(k)}\right\|_2^2 + \lambda_2\left\|\mathbf{a}^{(k)}\right\|_2\right\}.$$

Considering the case $\mathbf{a}^{(k)} \neq 0$ and setting the derivative with respect to $\mathbf{a}^{(k)}$ to zero, we obtain

$$-\left(\mathbf{r}^{(k)} - \mathbf{a}^{(k)}\right) + \lambda_2\frac{\mathbf{a}^{(k)}}{\left\|\mathbf{a}^{(k)}\right\|_2} = 0,$$

from which the following two equations derive:

$$\mathbf{a}^{(k)} = \left(1 + \frac{\lambda_2}{\left\|\mathbf{a}^{(k)}\right\|_2}\right)^{-1}\mathbf{r}^{(k)} \qquad \text{and} \qquad \left\|\mathbf{a}^{(k)}\right\|_2 = \left(1 + \frac{\lambda_2}{\left\|\mathbf{a}^{(k)}\right\|_2}\right)^{-1}\left\|\mathbf{r}^{(k)}\right\|_2,$$

by solving the second equation in $\left\|\mathbf{a}^{(k)}\right\|$ and substituting into the first we get,

$$\hat{\mathbf{a}}^{(k)} = \frac{\mathbf{r}^{(k)}}{\left\|\mathbf{r}^{(k)}\right\|_2}\left(\left\|\mathbf{r}^{(k)}\right\|_2 - \lambda_2\right). \tag{11}$$

Now, consider the case $\mathbf{a}^{(k)} = 0$. We need to impose that the $\mathbf{0}$ vector belongs to the pseudogradient evaluated at $\mathbf{a}^{(k)}$, i.e., there exists a vector $\boldsymbol{v} \in \mathbb{R}^{n\times 1}$, $\boldsymbol{v} \neq 0$ and $\|\boldsymbol{v}\|_2 \leq 1$ such that $-\mathbf{r}^{(k)} + \lambda_2\boldsymbol{v} = 0$. This is possible only if $\left\|\mathbf{r}^{(k)}\right\|_2 \leq \lambda_2$. Finally, combining both cases into a single equation, we get exactly the expression in Equation (10).

Let us now consider sub-problem (8). It is possible to prove that minimization of $F_2(\boldsymbol{\beta}) = F(\boldsymbol{\beta}, \hat{\mathbf{a}})$, where $\hat{\mathbf{a}}$ is the solution of Equation (9), i.e., vectors $\hat{\mathbf{a}}^{(k)}$ given by Equation (10), is equivalent to solve the following Huber's cost functional with parameter λ_2 plus a group lasso penalty. i.e.,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{K\cdot d\times 1}}{argmin}\left\{\sum_{k=1}^{K} \rho_{\lambda_2}^{H}\left(\mathbf{y}^{(k)} - \boldsymbol{W}\boldsymbol{\beta}^{(k)}\right) + \lambda_1||\boldsymbol{\beta}||_{2,1}\right\}.$$

Here the mutivariate Huber cost function is given by the following rule:

$$\rho_{\lambda_2}^H(v) = \begin{cases} \|v\|_2^2/2 & \text{if} \quad \|v\|_2 \leq \lambda_2 \\ \lambda_2\|v\|_2 - \lambda_2^2/2 & \text{if} \quad \|v\|_2 > \lambda_2, \end{cases} \tag{12}$$

for any $v \in \mathbb{R}^{n\times 1}$.

For the sake of completeness, we sketch below how to derive the above statement. Let $I = \left\{k : \hat{\mathbf{a}}^{(k)} = 0\right\} = \left\{k : \left\|\mathbf{r}^{(k)}\right\|_2 < \lambda_2\right\}$ and I^c be its complement.

$$\begin{aligned} F_1(\boldsymbol{\beta}) &= F(\boldsymbol{\beta}, \hat{\mathbf{a}}) \\ &= \frac{1}{2}\sum_{k=1}^{K}\left\|\mathbf{y}^{(k)} - \boldsymbol{W}\boldsymbol{\beta}^{(k)} - \hat{\mathbf{a}}^{(k)}\right\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_{2,1} + \lambda_2\sum_{k=1}^{K}\left\|\hat{\mathbf{a}}^{(k)}\right\|_2 \\ &= \frac{1}{2}\sum_{k\in I}\left\|\mathbf{y}^{(k)} - \boldsymbol{W}\boldsymbol{\beta}^{(k)}\right\|_2^2 + \frac{1}{2}\sum_{k\in I^c}\lambda_2^2 + \lambda_2\sum_{k\in I^c}\left(\left\|\mathbf{r}^{(k)}\right\|_2 - \lambda_2\right) + \lambda_1\|\boldsymbol{\beta}\|_{2,1} \\ &= \frac{1}{2}\sum_{k\in I}\left\|\mathbf{y}^{(k)} - \boldsymbol{W}\boldsymbol{\beta}^{(k)}\right\|_2^2 + \lambda_2\sum_{k\in I^c}\left\|\mathbf{r}^{(k)}\right\|_2 - \frac{1}{2}\sum_{k\in I^c}\lambda_2^2 + \lambda_1\|\boldsymbol{\beta}\|_{2,1}. \end{aligned}$$

In robust regression, i.e., regression in presence of outliers, it is well known that Huber loss is less sensitive to outliers than quadratic loss because it combines the squared loss and the absolute loss in an adaptive way. Intuitively, when the data points are not too big, Huber's loss function penalizes like the squared loss; otherwise, it penalizes like the absolute loss. However, it is not resistant to high leverage outliers (outliers in the predictors) because it never rejects gross outliers that have moderate or high leverage. Its breakdown point is $1/n$. Indeed, according to [23], a convex criterion is inherently incompatible with robustness.

The proposed procedure is summarized in Algorithm 1.

Algorithm 1 Calculate $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{a}}\right)$ with the Soft Thresholding operator

Input data:

- K curves $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(K)}$;
- parameters for the calculation of the low Q-factor and high Q-factor analysis filters p, q, s, J;
- maximum number of iterations max_iter;
- $\epsilon > 0$

(1) Build matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi} \rightarrow \boldsymbol{X}$, see Equation (5)
(2) Set initial estimate for $j = 0$, $\hat{\boldsymbol{\beta}}^0$ and $\hat{\mathbf{a}}^0 = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^0$
while no convergence **do**
$j = j + 1$
i. solve $\hat{\boldsymbol{\beta}}^j = \underset{\boldsymbol{\beta}\in\mathbb{R}^{Kd}}{argmin}\left\{\frac{1}{2}||\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta} - \hat{\mathbf{a}}^{j-1}||_2^2 + \lambda_1||\boldsymbol{\beta}||_{2,1} + \lambda_2\|\hat{\mathbf{a}}^{j-1}\|_{2,1}\right\}$
ii. evaluate residuals $\hat{\mathbf{r}}^j = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^j = \begin{bmatrix} \hat{\mathbf{r}}^{(1),j} \\ \hat{\mathbf{r}}^{(2),j} \\ \vdots \\ \hat{\mathbf{r}}^{(K),j} \end{bmatrix}$
iii. Soft threshold the residuals $\hat{\mathbf{a}}^{(k),j} = \frac{\hat{\mathbf{r}}^{(k),j}}{\left\|\hat{\mathbf{r}}^{(k),j}\right\|_2}\left(\left\|\hat{\mathbf{r}}^{(k),j}\right\|_2 - \lambda_2\right)_+$
iv. check convergence $\frac{\left\|\hat{\mathbf{a}}^j - \hat{\mathbf{a}}^{j-1}\right\|_2}{\left\|\hat{\mathbf{a}}^{j-1}\right\|_2} < \epsilon$ or $j =$max_iter
end while
Output: $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{a}}\right) = \left(\hat{\boldsymbol{\beta}}^j, \hat{\mathbf{a}}^j\right)$

Possible Improvements

In this section we explore the possibility to improve our estimator, adopting a different type of threshold operator. We are inspired by [13], where the authors deal with the outliers detection problem in the case of one channel. In the paper they propose a two step procedure: after a proper initialization for $\boldsymbol{\beta}$, in the first step, they apply a Soft thresholding operator to the residual, in the second step they use Ordinary Least Squares (OLS) to update vector $\boldsymbol{\beta}$. The main difference with our procedure is in the second step, where we perform a group lasso regression instead of an OLS to face the multitask regression. Although in [13] the authors deal with the outliers detection, looking for pointwise anomalies, the philosophy of their estimator is comparable to ours. Moreover, they note that in the presence of multiple outliers, the Soft threshold is not able to deal with masking and swamping effects, and propose to replace Soft-thresholding with Hard-thresholding obtaining a great improvement. Then they generalize the idea and introduce an entire class of methods, namely $\Theta - IPOD$ which uses a general Θ thresholding operator instead of the Soft thresholding operator, discovering important results and outstanding performance. In this direction of research, we explore the possibility to improve our estimator, substituting the first step of our procedure, i.e., the multivariate Soft thresholding operator expressed in Equation (10), by an appropriate Θ operator as done by the authors of [13]. Specifically, we substitute Equation (10), by the following Hard-thresholding operator

$$\hat{\mathbf{a}}^{(k)} = \mathbf{r}^{(k)} \mathbb{I}\left(\left\| \mathbf{r}^{(k)} \right\|_2 < \lambda_2 \right), \tag{13}$$

where $\mathbb{I}(\cdot)$ is the indicator function defined in Equation (3). We stress that substituting Equation (13) into the first step of our procedure, the second step becomes equivalent to a skipped-mean loss penalized with a group lasso penalty. We sketch below how to derive it.

$$\begin{aligned} F_2(\boldsymbol{\beta}) &= F(\boldsymbol{\beta}, \hat{\mathbf{a}}) \\ &= \frac{1}{2} \sum_{k=1}^{K} \left\| \mathbf{y}^{(k)} - \boldsymbol{W}\boldsymbol{\beta}^{(k)} - \hat{\mathbf{a}}^{(k)} \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_{2,1} + \frac{\lambda_2^2}{2} \sum_{k=1}^{K} \mathbb{I}\left(\hat{\mathbf{a}}^{(k)} \neq 0 \right) \\ &= \frac{1}{2} \sum_{k \in I} \left\| \mathbf{y}^{(k)} - \boldsymbol{W}\boldsymbol{\beta}^{(k)} \right\|_2^2 + \frac{\lambda_2^2}{2} card(I^c) + \lambda_1 \|\boldsymbol{\beta}\|_{2,1} \end{aligned} \tag{14}$$

where $I = \left\{ k : \hat{\mathbf{a}}^{(k)} = 0 \right\} = \left\{ k : \left\| \mathbf{r}^{(k)} \right\|_2 \leq \lambda_2 \right\}$ and I^c its complement.

Hence, substituting Soft threshold with Hard threshold automatically makes the second step equivalent to the following penalized skipped-mean loss

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{K \cdot d \times 1}}{argmin} \left\{ \sum_{k=1}^{K} \rho_{\lambda_2}^{SM} \left(\mathbf{y}^{(k)} - \boldsymbol{W}\boldsymbol{\beta}^{(k)} \right) + \lambda_1 ||\boldsymbol{\beta}||_{2,1} \right\}.$$

with loss defined as:

$$\rho_{\lambda_2}^{SM}(v) = \begin{cases} \|v\|_2^2/2 & \text{if} \quad \|v\|_2 \leq \lambda_2 \\ \lambda_2^2/2 & \text{if} \quad \|v\|_2 > \lambda_2, \end{cases} \tag{15}$$

for any $v \in \mathbb{R}^{n \times 1}$. Note that skipped-mean loss is not-convex and is a special case of Hampel loss, that belongs to the family of redescending M-estimators, see [24], with high breakdown point, close to 0.5. This means that is robust to outliers. For a summary, see Appendix 1 in [16].

Figure 1 shows a comparison between quadratic loss, Huber loss $\rho_{\lambda_2}^{H}(v)$, and skipped-mean loss $\rho_{\lambda_2}^{SM}(v)$.

We can conclude that, when using Hard threshold instead of Soft threshold, we solve the following problem instead of the problem in Equation (7)

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{a}}) \in \underset{\boldsymbol{\beta}\in\mathbb{R}^{Kd},\, \mathbf{a}\in\mathbb{R}^{Kn}}{argmin} \left\{ \frac{1}{2}\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta} - \mathbf{a}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_{2,1} + \frac{\lambda_2^2}{2}\|\mathbf{a}\|_{2,0} \right\}, \tag{16}$$

with $\|\mathbf{a}\|_{2,0} = \sum_{k=1}^{K} \mathbb{I}\left(\mathbf{a}^{(k)} \neq 0\right)$, the l_0 grouped norm.

Of course, the problem in Equation (16) is not convex, because the complexity penalty is not convex; therefore, we can only aspire to find one of the possible local minima of the objective function. Nonconvex functions may possess local optima that are not global optima, and our iterative method may terminate undesirably in one of these local optima. From a statistical perspective, although theoretical results for nonconvex penalties have been studied in [14,15] proving that all local optima are essentially as good as a global optimum, we cannot apply those results because our penalty does not satisfy their hypothesis. Therefore, our finding for this second procedure is purely empirical and we include the side condition $\|\hat{\boldsymbol{\beta}}\|_1 + \|\mathbf{a}\|_1 < R$ to guarantee the existence of at least one local/global optima. In addition, we will require $R \geq \|\boldsymbol{\beta}_0\|_1 + \|\mathbf{a}_0\|_1$ so that the true regression vector $\boldsymbol{\beta}_0$ and the true anomaly vector $\mathbf{a}_0$, that satisfy Equation (4), are feasible.

The proposed procedure is summarized in Algorithm 2.

Algorithm 2 Calculate $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{a}}\right)$ with the Hard Thresholding operator

Input data:

- K curves $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(K)}$;
- parameters for the calculation of the low Q-factor and high Q-factor analysis filters p, q, s, J;
- maximum number of iterations max_iter;
- $\epsilon > 0$

(1) Build matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi} \rightarrow \boldsymbol{X}$, see Equation (5)

(2) Set initial estimate for $j = 0$, $\hat{\boldsymbol{\beta}}^0$ and $\hat{\mathbf{a}}^0 = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^0$

while no convergence **do**

$j = j + 1$

i. solve $\hat{\boldsymbol{\beta}}^j = \underset{\boldsymbol{\beta}\in\mathbb{R}^{Kd}}{argmin} \left\{ \frac{1}{2}||\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta} - \hat{\mathbf{a}}^{j-1}||_2^2 + \lambda_1||\boldsymbol{\beta}||_{2,1} + \lambda_2\|\hat{\mathbf{a}}^{j-1}\|_{2,0} \right\}$

ii. evaluate residuals $\hat{\mathbf{r}}^j = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^j = \begin{bmatrix} \hat{\mathbf{r}}^{(1),j} \\ \hat{\mathbf{r}}^{(2),j} \\ \vdots \\ \hat{\mathbf{r}}^{(K),j} \end{bmatrix}$

iii. Hard threshold the residuals $\hat{\mathbf{a}}^{(k),j} = \mathbf{r}^{(k),j}\mathbb{I}\left(\left\|\mathbf{r}^{(k),j}\right\|_2 < \lambda_2\right)$,

iv. check convergence $\frac{\left\|\hat{\mathbf{a}}^j - \hat{\mathbf{a}}^{j-1}\right\|_2}{\left\|\hat{\mathbf{a}}^{j-1}\right\|_2} < \epsilon$ or $j =$max_iter

end while

Output: $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{a}}\right) = \left(\hat{\boldsymbol{\beta}}^j, \hat{\mathbf{a}}^j\right)$

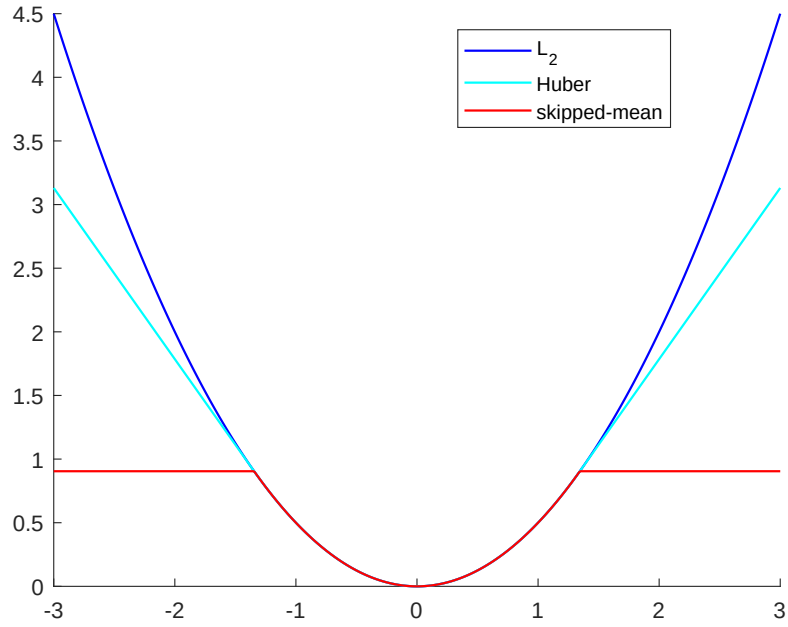

Figure 1. Loss functions: L_2, Huber and skipped - mean.

4. Implementation

In this section, we describe all the issues that make the proposed procedure effective: namely the initialization step, the numerical algorithm, and the choice of the regularization parameters λ_1 and λ_2.

As described in Section 3, we need to initialize the two step procedure by a robust estimator of parameters $\boldsymbol{\beta}$ which gives a robust estimator of $\boldsymbol{X\beta}$ to plug into expression of Equation (9). First, we note that vector $(\boldsymbol{X\beta})^t = \left(\left(\boldsymbol{W\beta}^{(1)}\right)^t, ..., \left(\boldsymbol{W\beta}^{(K)}\right)^t\right)$, hence to initialize our procedure is enough to establish a robust estimator for each of the signal component $f^{(k)} = \boldsymbol{W\beta}^{(k)}$. In our procedure, under the prior knowledge that the number of channels with anomaly are less then the half number of channels, we initialize $f^{(k)}$ by the median of the data channels, namely $median\left(y^{(1)}, ..., y^{(K)}\right)$

As regards the numerical algorithm we note that the first step described in Equation (9) has a closed-form solution given in Equation (10) and Equation (13) in the case of Soft and Hard threshold respectively; the second step in Equation (8) requires the solution of a linear regression problem with a group Lasso penalty. Hence for the solution of the second step we employed the `grpreg` R package that implements the efficient Group Descendent Algorithm presented in [25,26]. This algorithm works groupwise by using the separability of the model (5) in terms of a group of variables, i.e., it updates each group of variables freezing the other groups to their current value until convergence. The updating of each group of variables is performed through a multivariate soft-thresholding operator, under the assumption of "orthonormal groups". We stress that the "orthonormal group" property refers to the condition $\boldsymbol{X}_{G_j}^t \boldsymbol{X}_{G_j} = I$, with $\boldsymbol{X}_{G_j}$ denoting the reduction of design matrix $\boldsymbol{X}$ to columns of the subset of index G_j. When this condition is not satisfied the `grpreg` automatically orthonormalizes the design matrix, but this practice leads to a slight modification of the l_1/l_2-norm contained in the penalty, as pointed out in [27,28]. However, this is not our case, because the design matrix defined in Equation (5) satisfies the "orthonormal groups" property by construction, and hence, we can take complete advantage of the Group Descendent Algorithm implemented in the `grpreg` package.

As in any penalized regression approach, the choice of the regularization parameter is really strategic. In particular, in the proposed procedure the regularization parameter has two components, namely λ_1 and λ_2. The latter controlling the degree of regularization in the regression model on parameter $\boldsymbol{\beta}$, the last controlling the threshold over which the

residual of the previous regression is considered *not acceptable*. Under the data model described in Equation (1), the expected size of the residual $\mathbf{r}^{(k)}$, in absence of anomaly in channel k, is given by $\sigma\sqrt{n}$, hence it is natural to fix $\lambda_2 = \hat{\sigma}\sqrt{n}$ where $\hat{\sigma}$ is an estimator of the noise variance. In particular, we adopt $\hat{\sigma} = median\left(\left\|\mathbf{r}^{(1)})\right\|_2, ..., \left\|\mathbf{r}^{(K)}\right\|_2\right)/\sqrt{n}$ where $\mathbf{r}^{(k)}$ is the residual after having initialized $f^{(k)}$ as described above. The choice of the regularization λ_1 is more demanding; we avoided CV criterion since it is quite computationally heavy and in principle when applied on the $(\mathbf{y}, \boldsymbol{X})$ data, a large prediction error can be due to both a suboptimal $\boldsymbol{\beta}$ as well as to the presence of an anomaly. Hence we suggest tuning parameter λ_1 by the BIC criterion, which is much less computationally demanding and easily modifiable to our setting. In particular, we adopted the following formula

$$BIC(\lambda_1) = nR\, log(RSS(\lambda_1)/nR) + df(\lambda_1)\, log(nR)$$

where $df(\lambda_1)$ is evaluated according to formula (22) in [26] and $RSS(\lambda_1) = \|\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\mathbf{a}}\|_2^2$, with $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{a}}$ estimates obtained for λ_1. The grid of λ_1 values has been obtained as in Section 3.5 of [26].

5. Simulations and Real Examples

To show the performance of the proposed methodology, we performed numerical experiments both in a simulation setting and in a real dataset case.

5.1. Synthetic Data

We generated data according to the model in Equation (1) with some known functions $f^k(t)$ and some known anomalies and evaluated the performance of the proposed techniques in terms of signals reconstruction and anomalies detection.

Since the proposed method deals with nominal signals with fast oscillating characteristics, we choose the HiSine and the TwoChirp signals (represented in Figures 2 and 3) among the classical signals employed in the literature. Both can be well represented by using an appropriate RADWT. In particular, a RADWT with Q-factor almost 5 ($p = 8$, $q = 9$, $s = 3$ and $j = 10$) is appropriate for their sparse representation.

As regards the anomalies, there are no assumptions on their shape, i.e., they can be of any duration and shape and they can appear in any of the signal components, so we decided to model them by the following formula

$$a(t) = M \cdot I(t_1 \leq t \leq t_2) \tag{17}$$

to represent situations where the mean of the signal undergoes to a shift during a specific interval of time (t_1, t_2); where the shift $M = h \cdot \max_{1 \leq i \leq n,k}\left(\mathbf{y}^{(k)}(t_i)\right)$ with $h = 0.5$ or 2, to mimic situation where the anomaly is weak or strong with respect to the underlying signals. We generated data according to model (1), building K signals, $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(K)}$ of length $n = 256$ and for such a choice the W matrix has dimension $n \times d = 256 \times 695$. Each signal has been generated randomizing the original signal $\mathbf{f}$ (*HiSine* or *Twochirp*) by the formula $\mathbf{f}^{(k)} = \mathbf{f} \cdot u$, where u is a vector of i.i.d. variables with distribution Unif$(0, 1)$. These generations guarantee that the true underlying signals $\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(K)}$ have similar shape and the same sparsity patterns. According to the model in Equation (1), we added noise to each channel as $\boldsymbol{\varepsilon}^{(k)} \sim N(\mathbf{0}, \sigma^2 I)$ with σ^2 related to the variance of the true underlying signal realizing three different type of $SNR = (0.5, 1, 6)$ to mimic situations of different severity of noise. For completeness, here is the SNR expression

$$SNR = \frac{\frac{1}{K}\sum_{i=1}^{K} \mathrm{Var}\left(\mathbf{f}^{(k)}\right)}{\sigma^2}.$$

Finally, we added anomalies $\mathbf{a}^{(k)}$ to some channels as in Formula (17).

To sketch, we have considered two settings:

- $\mathbf{f} = HiSine$, $K = 3$, $\mathbf{a}^{(1)}$ as in Formula (17) with $(t_1, t_2) = (78, 177)$;
- $\mathbf{f} = TwoChirp$, $K = 5$, $\mathbf{a}^{(1)}$ as in Formula (17) with $(t_1, t_2) = (78, 177)$ and $\mathbf{a}^{(4)}$ as in Formula (17) with $(t_1, t_2) = (157, 256)$.

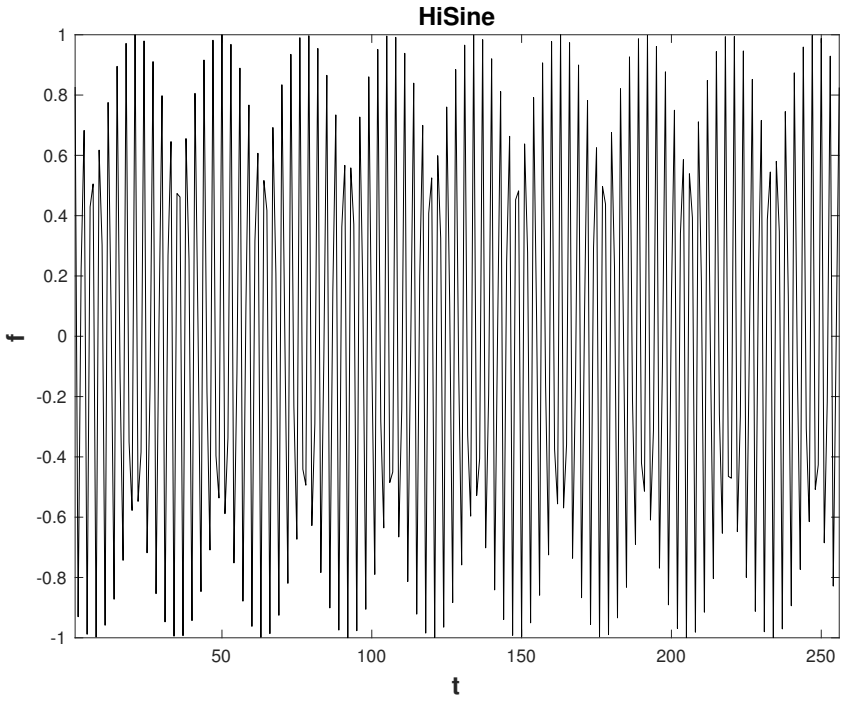

Figure 2. HiSine signal.

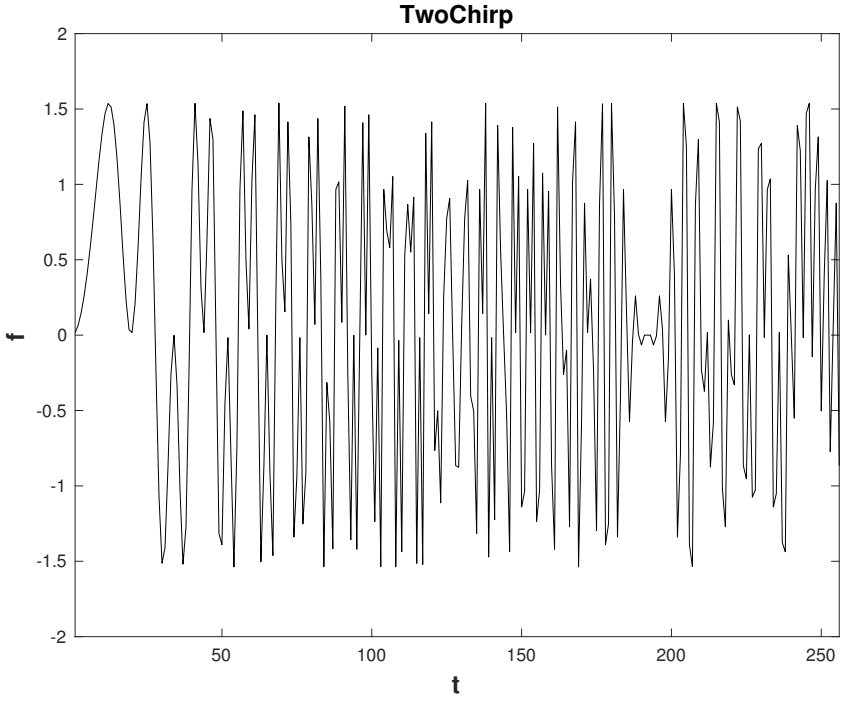

Figure 3. TwoChirp signal.

To summarize, for each of these two settings ($\mathbf{f} = HiSine$ and $\mathbf{f} = TwoChirp$), we analyzed two different levels of anomalies ($h = 0.5$ and $h = 2$ in Formula (17)) and three

different levels of SNR ($SNR = 0.5, 1$ and 6). These choices are arbitrary but they were dictated by the idea of understanding a sort of *breakdown point* of our procedure. Naturally, we expect that the weaker the anomalies and the stronger the noise, the worse the problem. Finally, we evaluated performance by computing the MSE for each signal's component as

$$\mathrm{MSE}_k = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}^{(k)}(t_i) - \mathbf{f}^{(k)}(t_i)\right)^2,\ k = 1, \ldots, K;$$

with $\hat{\mathbf{f}}^{(k)}$ estimate of $\mathbf{f}^{(k)}$, and by computing the relative number of (falsely) detected anomalies (FPR) as well as the relative number of not detected anomalies (FNR), namely

$$\begin{aligned}\mathrm{FPR} &= \frac{1}{K-A}\sum_{i=1}^{K}\mathbb{I}\left(\left\|\hat{\mathbf{a}}^{(k)}\right\|_2 \neq 0 \wedge \left\|\mathbf{a}^{(k)}\right\|_2 = 0\right)\\ \mathrm{FNR} &= \frac{1}{A}\sum_{i=1}^{K}\mathbb{I}\left(\left\|\hat{\mathbf{a}}^{(k)}\right\|_2 = 0 \wedge \left\|\mathbf{a}^{(k)}\right\|_2 \neq 0\right),\end{aligned}$$

A being the true number of channel with anomalies. In addition, we evaluated the intensity of the estimated anomalies to compare with the intensity of the simulated ones and to analyze the detection capabilities of both our proposals as

$$\frac{\left\|\hat{\mathbf{a}}^{(k)}\right\|_2^2}{n}, \quad k = 1, \ldots, K.$$

To be robust to the particular realization in generating synthetic data (and corresponding noise), each experiment was run several times. In particular, we ran 10 instances for each of the 12 considered cases (2 (function settings) $\times$ 2 (levels of anomaly) $\times$ 3 (SNRs)). For the sake of exposition, in the following we show separately results obtained for each function's settings. Specifically, Figures 4–9 and Tables 1 and 2 show results of the 6 different combinations of levels of anomaly and SNRs obtained when the function is $\mathbf{f} = HiSine$ and Figures 10–15 and Tables 3 and 4 show analogous results when the function is $\mathbf{f} = TwoChirp$. Moreover, we show in Figures 5 and 6 the plots of the shape of the unknown signals and the goodness of reconstructions by the Hard thresholding procedure in the two extreme cases, weak anomaly ($h = 0.5$) with highest noise ($SNR = 0.5$) and strong anomaly ($h = 2$) with lowest noise level ($SNR = 6$) for the first setting $\mathbf{f} = HiSine$; while in Figures 7 and 8 we show the same results obtained by the Soft thresholding procedure. For the second setting $\mathbf{f} = TwoChirp$, four analogous plots are given in Figures 11 and 12 and Figures 13 and 14 for the Hard and Soft thresholding procedures respectively.

Comment on the results. Concerning anomaly detection performance, we observe that for both procedures, Hard and Soft, and both settings, $\mathbf{f} = HiSine$ and $\mathbf{f} = TwoChirp$, it is worst in the case of weak anomaly ($h = 0.5$, Tables 1 and 3) with respect to the case of strong anomaly ($h = 2$, Tables 2 and 4). This is a consequence of the masking effect: the weaker the anomaly the easier the noise, and the anomalies can mask each other. However, there is an interesting difference between the two procedures. In particular, for the Hard thresholding, we observe that FNR is always zero, while there can be some $FPR > 0$ in the more difficult setting, i.e., when the anomalies are weaker and the level of noise is higher. This means that the Hard thresholding procedure always detects the true anomalies, being prone to some false positives detection when stressing the problem setting. This observation is relevant from the practical point of view because in such extreme cases one can always adopt a post-processing rule to classify the detected anomalies, being confident enough not to miss any anomaly. On the contrary, the Soft thresholding is much more conservative, hence when stressing the setting, i.e., when the anomalies are weaker ($h = 0.5$) and the level of noise higher, the Soft thresholding procedure is more prone to false negatives detection. Indeed, for the Soft we always get $FPR = 0$, while stressing the setting we get some $FNR > 0$ (the opposite of Hard). This is not a drawback, rather a different

way of reacting to the difficulty of the problem. However, from a practical point of view, when using Soft instead of Hard it can be a disadvantage to have false-negative detection because when the result undergoes post-processing the anomalies not detected are not further analyzed. However, although our two procedures are not the definitive answer to the AD problem, it is important from a practical point of view to be aware that the Hard thresholding tends to select more anomalies (and therefore to have some false positives) while the Soft threshold tends to select fewer anomalies (and therefore to have some false negatives). In addition, Figures 9 and 15 show the boxplots of the intensity of the simulated anomalies and of the estimated anomalies by both Hard and Soft thresholding. We can note the good performance of the Hard thresholding procedure with respect to the Soft thresholding one, both in the case of weak anomalies and strong anomalies, with a better intensities estimate in the last case.

As regards the estimation performance, we observe that for both procedures, Hard and Soft, and both settings, $\mathbf{f} = HiSine$, and $\mathbf{f} = TwoChirp$, it decreases when increasing the level of noise (i.e., decreasing the SNR). In fact, the MSE significantly decreases when moving from the left to the right in all box-plot shown in Figure 4 for the first setting $\mathbf{f} = HiSine$ and in Figure 10 for the second setting $\mathbf{f} = TwoChirp$; panels (a)–(c) refer to Hard thresholding for $h = 0.5$ and $h = 2$, respectively, and panels (b) and (d) refer to Soft thresholding for $h = 0.5$ and $h = 2$, respectively. Interestingly, in line with what we observed on the detection performance results, we have significant differences in terms of MSE performance on the anomalous channels between the Hard and the Soft. See box-plots in positions 1, 4, 7, corresponding to the first channel for the $HiSine$ signal and box-plots in positions 1, 4, 6, 9, 11, 14, corresponding to the first and the fourth channels for the $TwoChirp$ signal. The anomalous channels always have a better MSE for the Hard than for the Soft. This is a direct consequence of our double-step procedure: when we correctly detect the anomaly (in the first step), we correctly adjust the residual (in the second step) and hence we obtain a better estimation. On the channels without anomaly, detection performance is comparable.

Finally, we stress that what we have observed on average (looking at box-plots) about estimation performance, can be inspected visually for a randomly chosen realization looking at the first channel in Figure 5 (Hard thresholding, $h = 0.5$, $SNR = 0.5$), Figure 6 (Hard thresholding, $h = 2$, $SNR = 6$), Figure 7 (Soft thresholding, $h = 0.5$, $SNR = 0.5$) and Figure 8 (Soft thresholding, $h = 2$, $SNR = 6$), which refer to $\mathbf{f} = HiSine$ and looking at the first and fourth channels in Figure 11 (Hard thresholding, $h = 0.5$, $SNR = 0.5$), Figure 12, (Hard thresholding, $h = 2$, $SNR = 6$), Figure 13 (Soft thresholding, $h = 0.5$, $SNR = 0.5$) and Figure 14 (Soft thresholding, $h = 2$, $SNR = 6$), which refer to $\mathbf{f} = TwoChirp$.

Table 1. False positives rates FPR (%) and false negatives rates FNR (%) for the simulations carried out on HiSine signals, $h = 0.5$. We report average values of the indicators over 10 simulations.

HiSine	$SNR = 0.5$	$SNR = 1$	$SNR = 6$
FPR_{hard}	0.05	0.05	0
FNR_{hard}	0	0	0
FPR_{soft}	0	0	0
FNR_{soft}	1	1	1

Table 2. False positives rates FPR (%) and false negatives rates FNR (%) for the simulations carried out on HiSine signals, $h = 2$. We report average values of the indicators over 10 simulations.

HiSine	$SNR = 0.5$	$SNR = 1$	$SNR = 6$
FPR_{hard}	0	0	0
FNR_{hard}	0	0	0
FPR_{soft}	0	0	0
FNR_{soft}	0.1	0	0

Table 3. False positives rates FPR (%) and false negatives rates FNR (%) for the simulations carried out on TwoChirp signals, $h = 0.5$. We report average values of the indicators over 10 simulations.

TwoChirp	$SNR = 0.5$	$SNR = 1$	$SNR = 6$
FPR_{hard}	0	0	0
FNR_{hard}	0	0	0
FPR_{soft}	0	0	0
FNR_{soft}	1	1	0.8

Table 4. False positives rates FPR (%) and false negatives rates FNR (%) for the simulations carried out on TwoChirp signals, $h = 2$. We report average values of the indicators over 10 simulations.

TwoChirp	$SNR = 0.5$	$SNR = 1$	$SNR = 6$
FPR_{hard}	0	0	0
FNR_{hard}	0	0	0
FPR_{soft}	0	0	0
FNR_{soft}	0	0	0

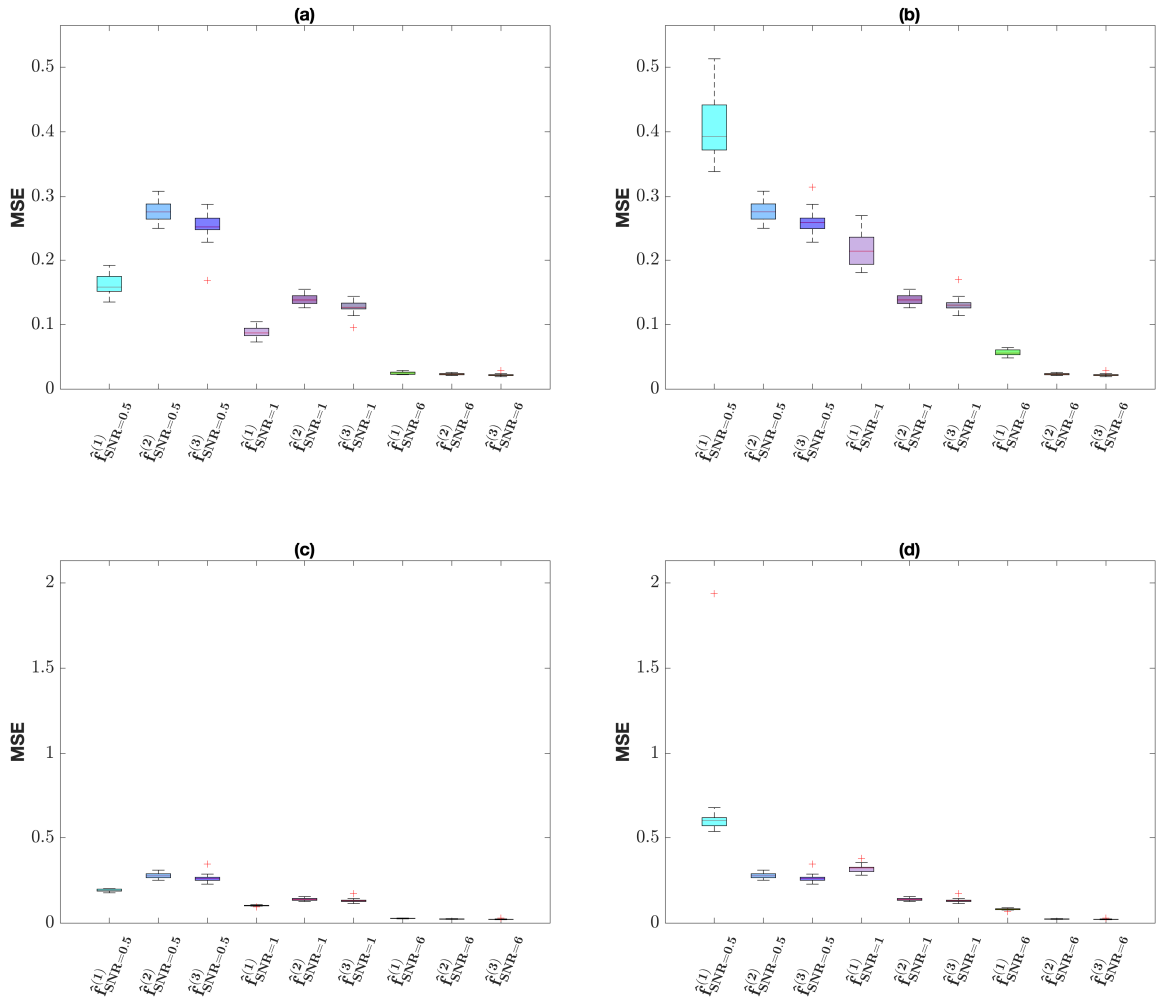

Figure 4. Mean Square Error for the estimated signals $\hat{\mathbf{f}}^{(1)}$, $\hat{\mathbf{f}}^{(2)}$, $\hat{\mathbf{f}}^{(3)}$ for the HiSine. First row: (**a**) Hard and (**b**) Soft estimators for $h = 0.5$. Second Row: (**c**) Hard and (**d**) Soft estimators for $h = 2$. Blue colors refer to $SNR = 0.5$ (first 3 boxes), violet colors refer to $SNR = 1$ (second 3 boxes), green colors refer to $SNR = 6$ (last 3 boxes).

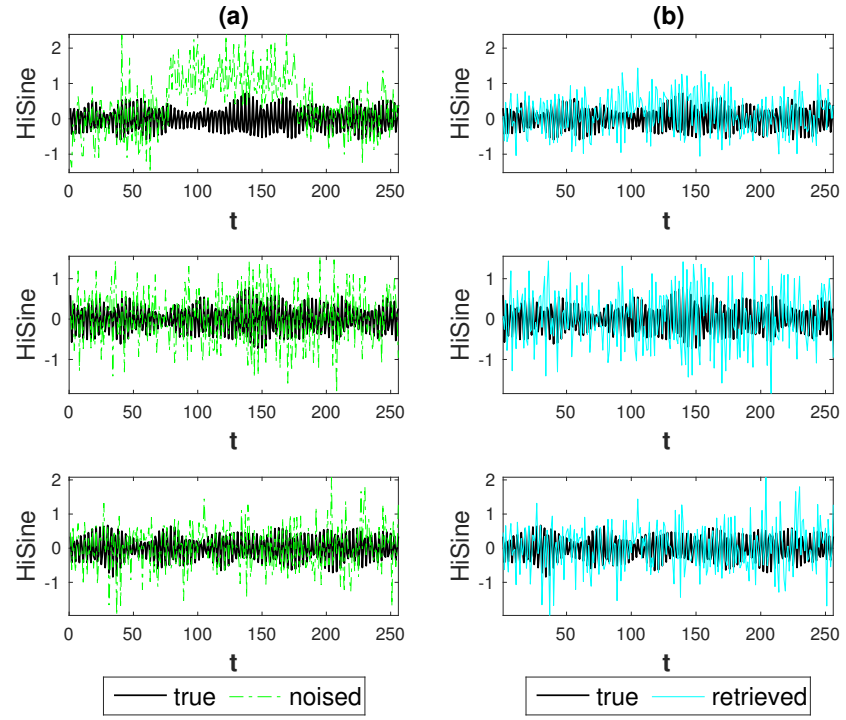

Figure 5. First setting: HiSine signals, weak anomaly ($h = 0.5$), highest noise level ($SNR = 0.5$). Result for Hard thresholding procedure: (**a**) true signals (black) and noised signals (green); (**b**) true signals (black) and retrieved signals (cyan).

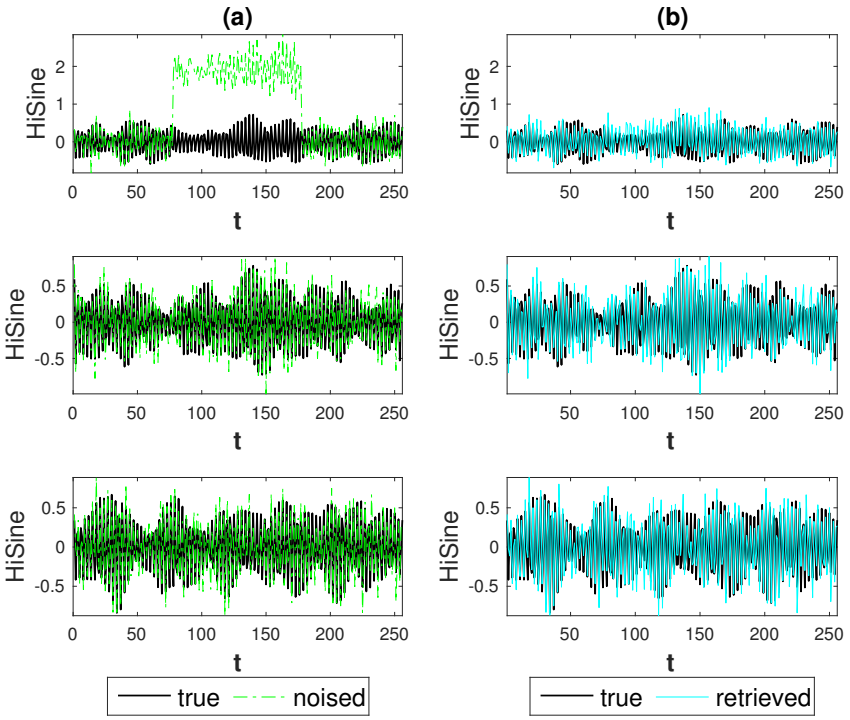

Figure 6. First setting: HiSine signals, strong anomaly ($h = 2$), lowest noise level ($SNR = 6$). Result for Hard thresholding procedure: (**a**) true signals (black) and noised signals (green); (**b**) true signals (black) and retrieved signals (cyan).

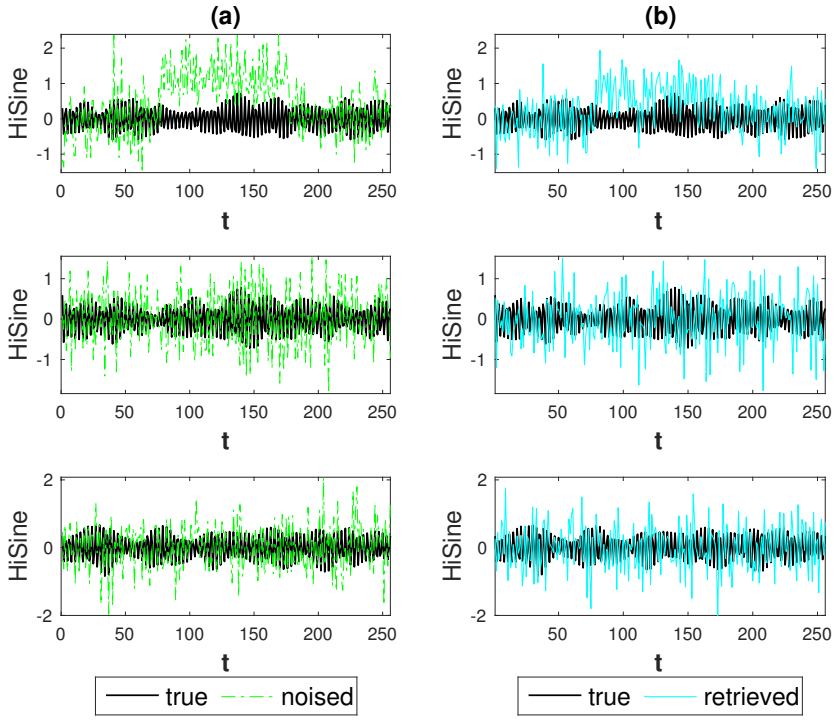

Figure 7. First setting: HiSine signals, weak anomaly ($h = 0.5$), highest noise level ($SNR = 0.5$). Result for Soft thresholding procedure: (**a**) true signals (black) and noised signals (green); (**b**) true signals (black) and retrieved signals (cyan).

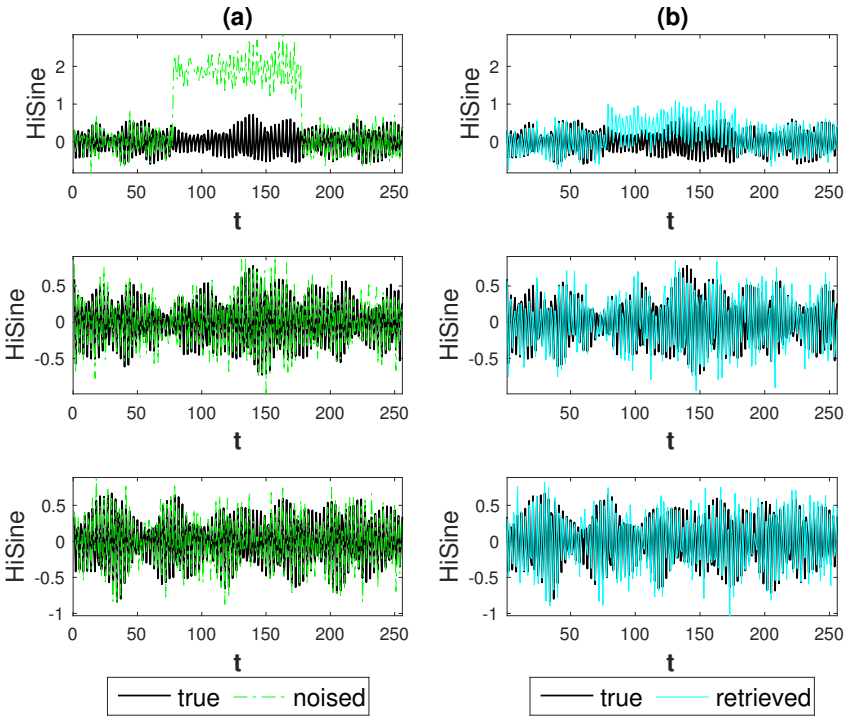

Figure 8. First setting: HiSine signals, strong anomaly ($h = 2$), lowest noise level ($SNR = 6$). Result for Soft thresholding procedure:(**a**) true signals (black) and noised signals (green); (**b**) true signals (black) and retrieved signals (cyan).

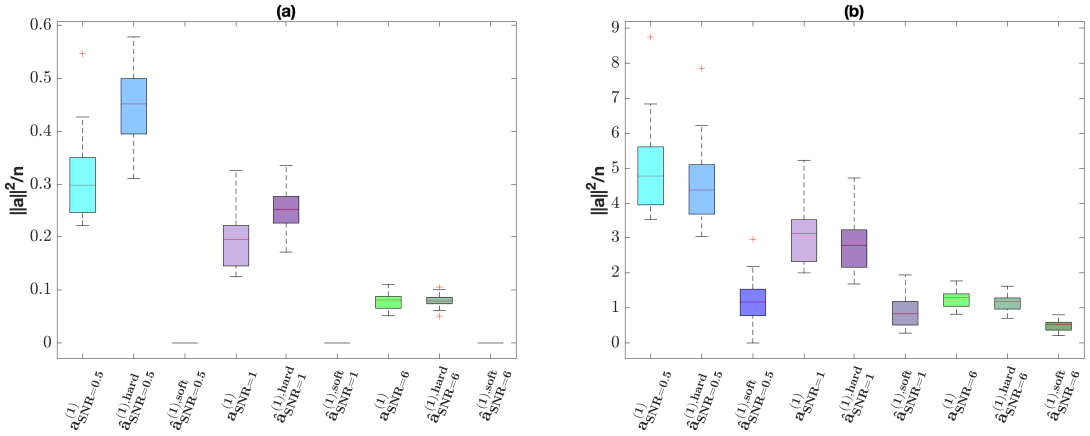

Figure 9. Intensity of the simulated anomaly $\mathbf{a^{(1)}}$, of the estimated anomaly by Hard thresholding $\hat{\mathbf{a}}^{(\mathbf{1}),\mathbf{hard}}$ and, of the estimated anomaly by Soft thresholding $\hat{\mathbf{a}}^{(\mathbf{1}),\mathbf{soft}}$ for the HiSine. Panel (**a**) refers to $h = 0.5$ and anomaly in channel 1, panel (**b**) refers to $h = 2$ and anomaly in channel 1. Blue colors refer to $SNR = 0.5$ (first 3 boxes), violet colors refer to $SNR = 1$ (second 3 boxes), green colors refer to $SNR = 6$ (last 3 boxes).

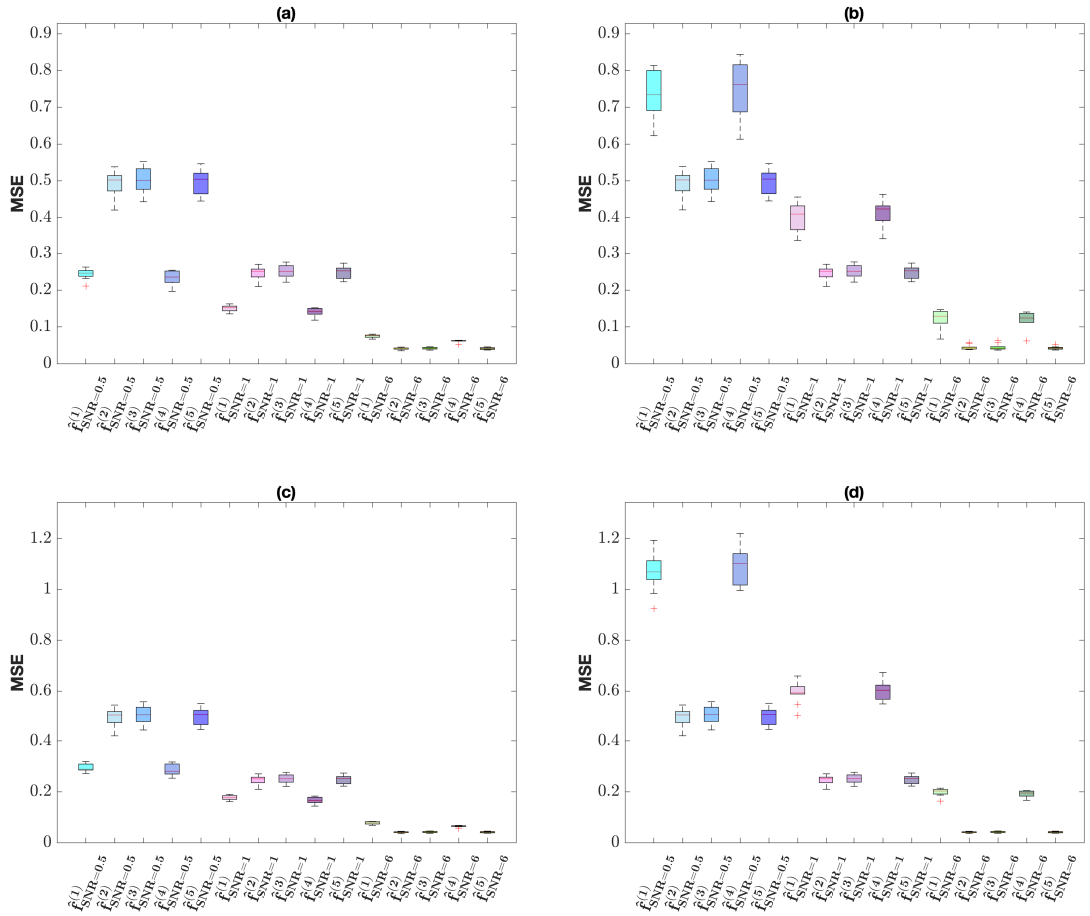

Figure 10. Mean Square Error for the estimated signals $\hat{\mathbf{f}}^{(\mathbf{1})}$, $\hat{\mathbf{f}}^{(\mathbf{2})}$, $\hat{\mathbf{f}}^{(\mathbf{3})}$, $\hat{\mathbf{f}}^{(\mathbf{4})}$, $\hat{\mathbf{f}}^{(\mathbf{5})}$ for the TwoChirp. First row: (**a**) Hard and (**b**) Soft estimators for $h = 0.5$. Second Row: (**c**) Hard and (**d**) Soft estimators for $h = 2$. Blue colors refer to $SNR = 0.5$ (first 5 boxes), violet colors refer to $SNR = 1$ (second 5 boxes), green colors refer to $SNR = 6$ (last 5 boxes).

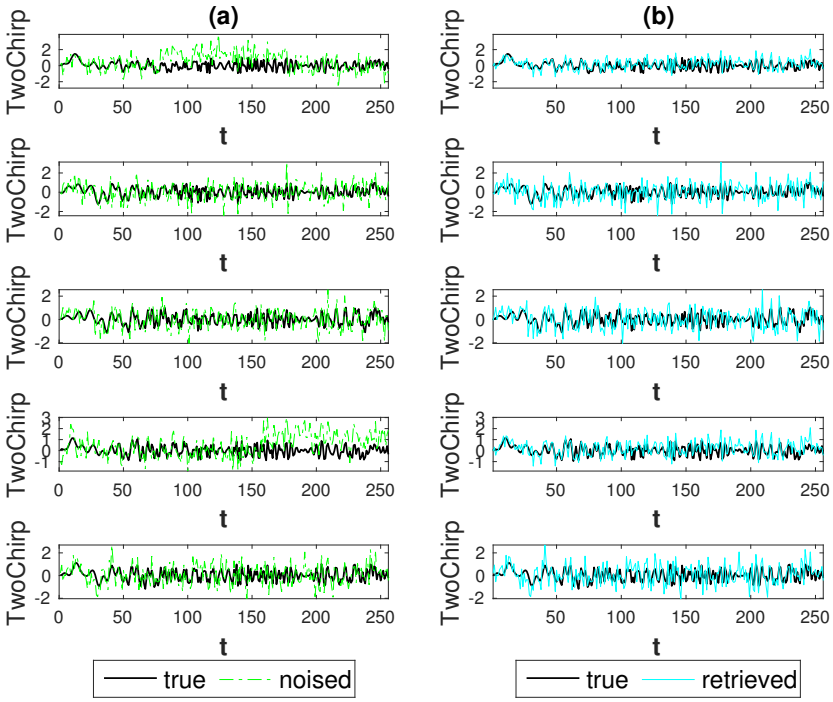

Figure 11. Second setting: TwoChirp signals, weak anomaly ($h = 0.5$), highest noise level ($SNR = 0.5$). Result for Hard thresholding procedure: (**a**) true signals (black) and noised signals (green); (**b**) true signals (black) and retrieved signals (cyan).

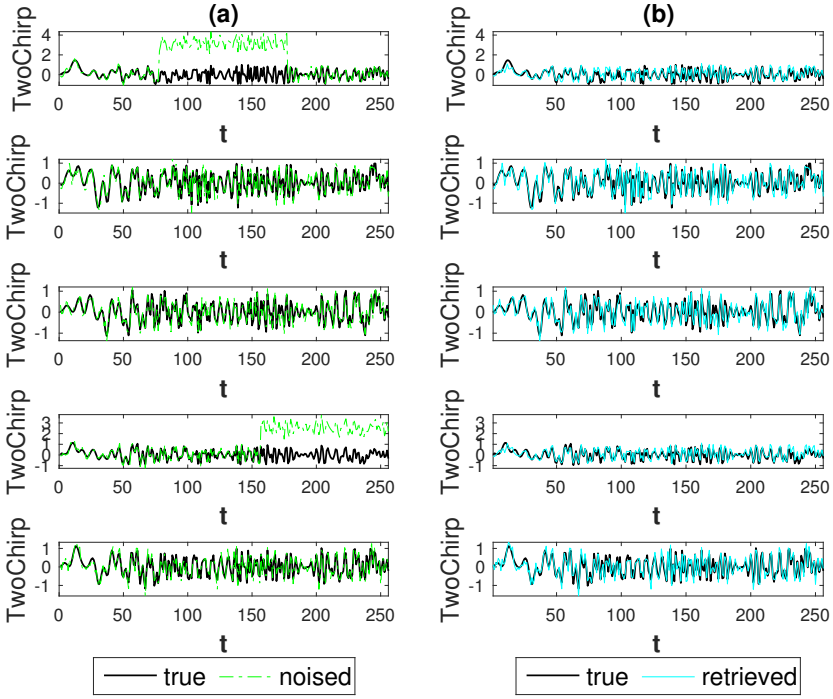

Figure 12. Second setting: TwoChirp signals, strong anomaly ($h = 2$), lowest noise level ($SNR = 6$). Result for Hard thresholding procedure: (**a**) true signals (black) and noised signals (green); (**b**) true signals (black) and retrieved signals (cyan).

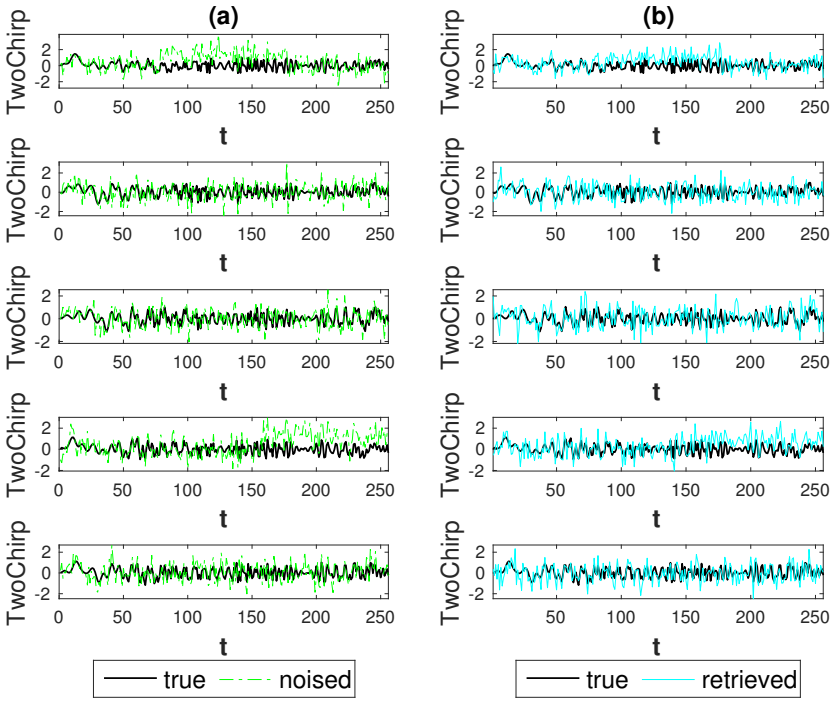

Figure 13. Second setting: TwoChirp signals, weak anomaly ($h = 0.5$), highest noise level ($SNR = 0.5$). Result for Soft thresholding procedure: (**a**) true signals (black) and noised signals (green); (**b**) true signals (black) and retrieved signals (cyan).

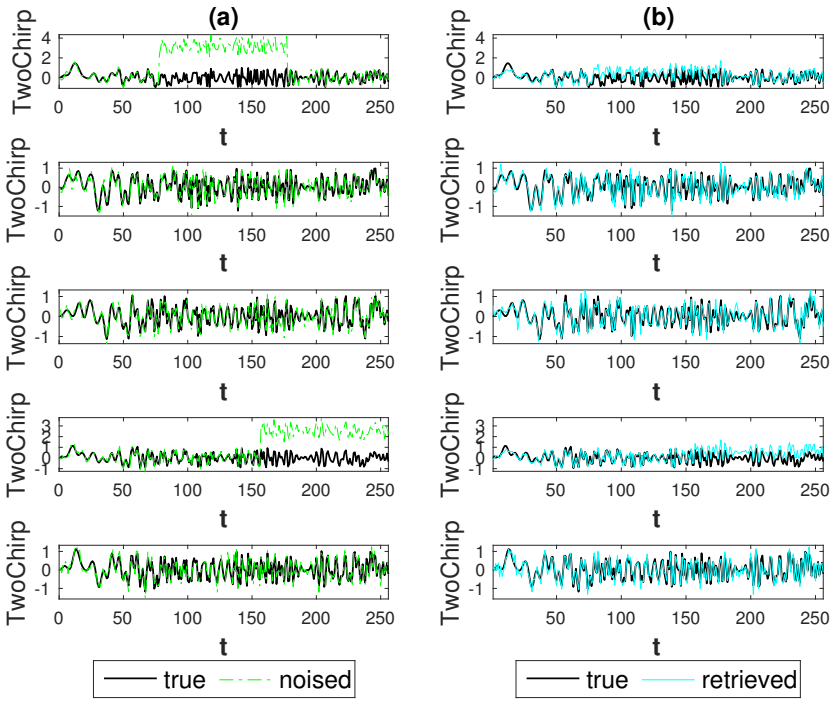

Figure 14. Second setting: TwoChirp signals, strong anomaly ($h = 2$), lowest noise level ($SNR = 6$). Result for Soft thresholding procedure: (**a**) true signals (black) and noised signals (green); (**b**) true signals (black) and retrieved signals (cyan).

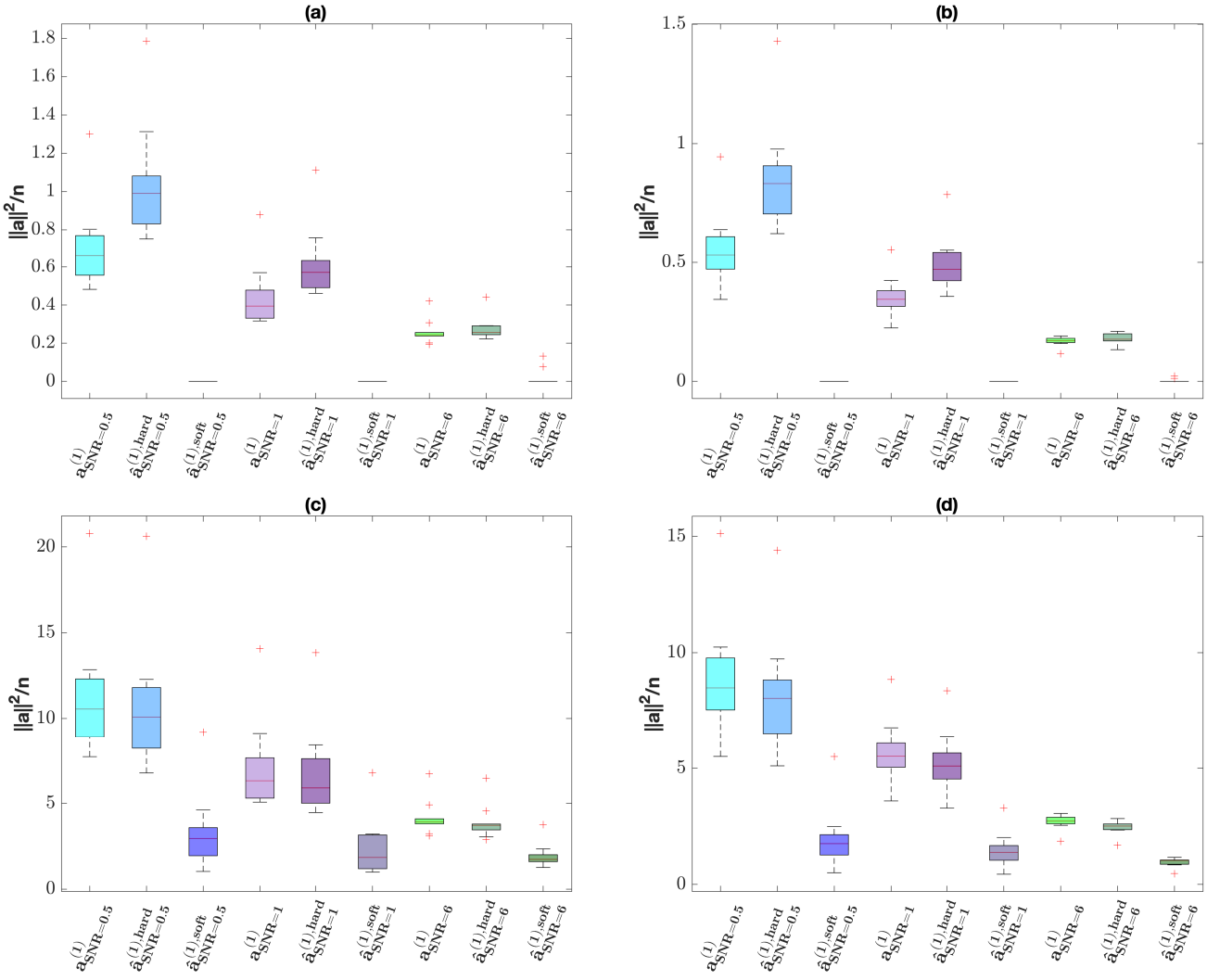

Figure 15. Intensity of the simulated anomaly $\mathbf{a}^{(1)}$, of the estimated anomaly by Hard thresholding $\hat{\mathbf{a}}^{(1),\mathbf{hard}}$ and, of the estimated anomaly by Soft thresholding $\hat{\mathbf{a}}^{(1),\mathbf{soft}}$ for the TwoChirp. Panel (**a**) refers to $h = 0.5$ and anomaly in channel 1, panel (**b**) refers to $h = 0.5$ and anomaly in channel 4, panel (**c**) refers to $h = 2$ and anomaly in channel 1, panel (**d**) refers to $h = 2$ and anomaly in channel 4. Blue colors refer to $SNR = 0.5$ (first 3 boxes), violet colors refer to $SNR = 1$ (second 3 boxes), green colors refer to $SNR = 6$ (last 3 boxes).

5.2. Real Data

To illustrate the performance of our procedures on a real example, we considered the problem of detecting anomalies for signals recording measurements on a water pump. The signals represent the behavior of the pump in both normal and abnormal conditions and a complete description of it is given in [29]. The dataset is part of a repository managed by the Delft University of Technology and is freely available at http://homepage.tudelft.nl/n9d04/occ/index.html (accessed on the 18 February 2021) (dataset 541: Delft pump AR app). Each signal has a length $n = 160$ and has been obtained by first fitting an AutoRegressive model of order 32 (AR(32)) to each of 5 vibration measurements and then by combining the coefficients in a single vector.

This dataset has been previously analyzed in [30]. In that paper, the authors proposed a data domain description, named the Support Vector Data Description (SVDD) to find the finest representation of the data such that the normal target signals are optimally clusterized and can be distinguished as best as possible from the abnormal ones. Their approach has a good performance and in some sense, it is another perspective for looking at the detection of anomalies. However, the data representation proposed in [30] does not take into account the longitudinal shape of the data, which are ordered coefficients; hence the method proposed in [30] is expected to be invariant under permutation of the data. On the other hand, our analysis is truly functional, since we treat the data as functions/signals; moreover, the dataset satisfies the hypothesis of our model. Since the underlying signals have a fast oscillating characteristic, they share a similar shape which presupposes the same sparsity pattern into the RADWT dictionary and the signals measured under abnormal conditions can be modeled as the signals measured under normal conditions plus some shift in the first part of the interval.

We considered 20 signals, 12 nominal and 8 with the anomaly, randomly chosen from the whole dataset, see Figure 16. We applied both procedures, Hard and Soft thresholding, obtaining reconstruction shown in Figure 17, panels (a) and (b). In the same figure, panels (c) and (d), we also display the retrieved residuals which represent anomalies. While the Soft thresholding has no false positives neither false negatives, the Hard thresholding has 2 false positives (channel 10 and 11), with the norm of the order of the estimated variance 0.0035, see Table 5. However, the Hard thresholding estimates better the intensity of the anomalies (which represent residuals for the regression on $\boldsymbol{\beta}$), permitting a better signals reconstruction. This does not happen for the Soft thresholding, which results in over-smoothed signal reconstruction. Finally, we can conclude that the Hard thresholding procedure has a better performance although it has two false positives, which could be easily detected by standard post-processing.

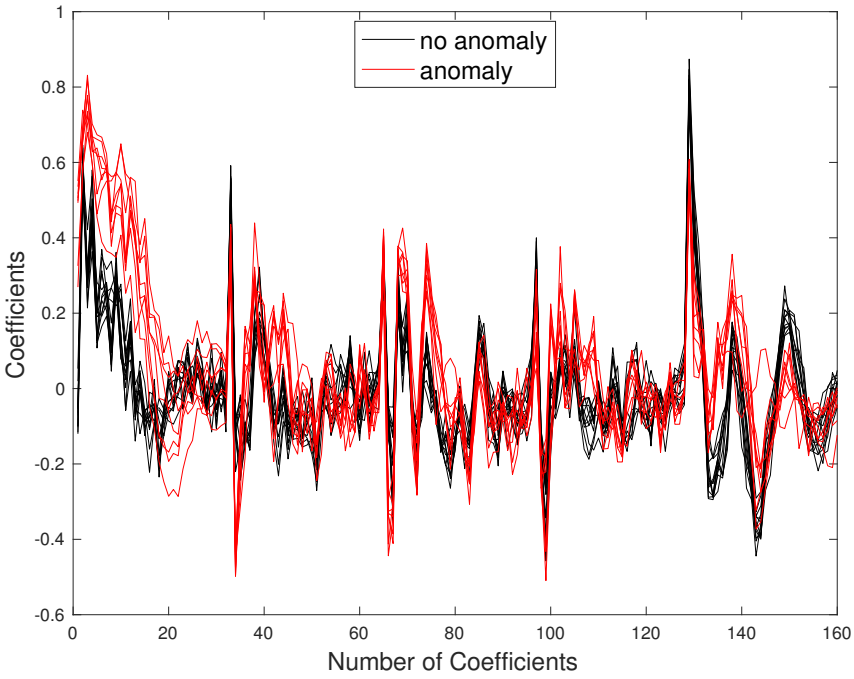

Figure 16. Pump data: red color indicates anomaly signals, black color indicates normal signals.

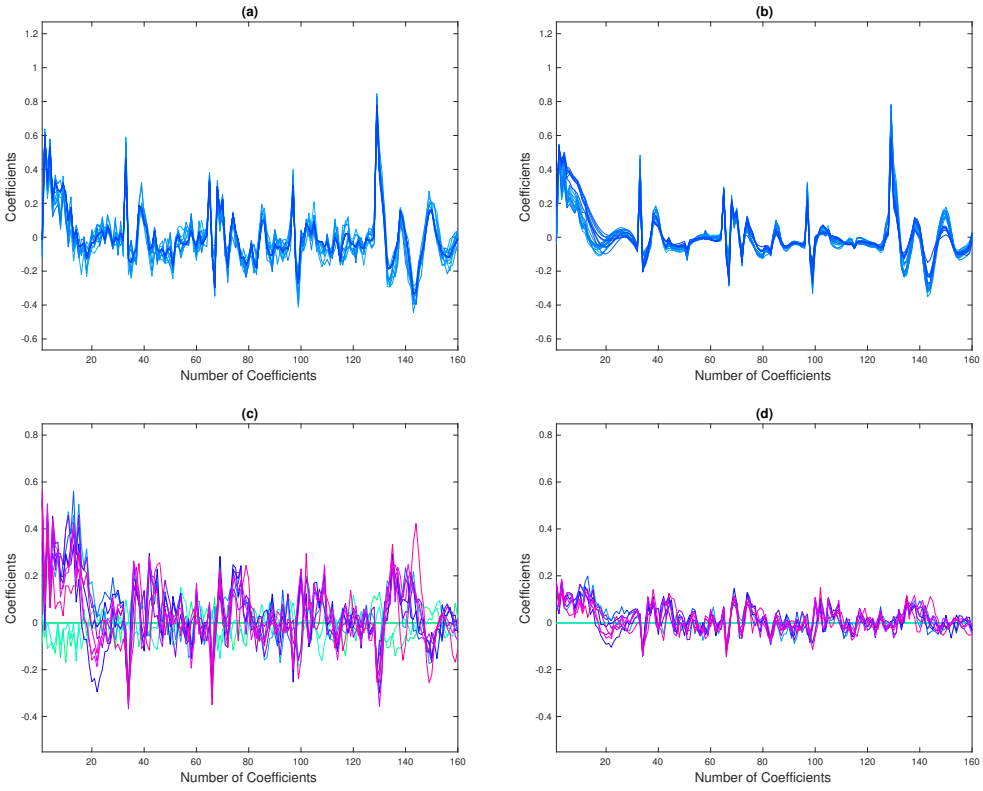

Figure 17. First row: (**a**) Pump data estimated by the Hard thresholding procedure (**b**) Pump data estimated by the Soft thresholding procedure. Second Row: (**c**) Anomalies estimated by the Hard thresholding procedure (**d**) Anomalies estimated by the Soft thresholding procedure.

Table 5. $\|\hat{\mathbf{a}}\|^2/n$ for the Hard estimator.

Pump Data	
$\left\|\hat{\mathbf{a}}^{(1)}\right\|_2^2/n, \ldots, \left\|\hat{\mathbf{a}}^{(9)}\right\|_2^2/n$	0
$\left\|\hat{\mathbf{a}}^{(10)}\right\|_2^2/n$	**0.0037**
$\left\|\hat{\mathbf{a}}^{(11)}\right\|_2^2/n$	**0.0036**
$\left\|\hat{\mathbf{a}}^{(12)}\right\|_2^2/n$	0
$\left\|\hat{\mathbf{a}}^{(13)}\right\|_2^2/n$	0.0205
$\left\|\hat{\mathbf{a}}^{(14)}\right\|_2^2/n$	0.0225
$\left\|\hat{\mathbf{a}}^{(15)}\right\|_2^2/n$	0.0207
$\left\|\hat{\mathbf{a}}^{(16)}\right\|_2^2/n$	0.0202
$\left\|\hat{\mathbf{a}}^{(17)}\right\|_2^2/n$	0.0192
$\left\|\hat{\mathbf{a}}^{(18)}\right\|_2^2/n$	0.0190
$\left\|\hat{\mathbf{a}}^{(19)}\right\|_2^2/n$	0.0203
$\left\|\hat{\mathbf{a}}^{(20)}\right\|_2^2/n$	0.0201

6. Conclusions

In this paper, we presented two procedures for anomaly detection in multichannel signals under a structural hypothesis on the underlying signals covering some specific real-life situations. In particular, we dealt with fast oscillating signals under the assumption that they share a common specific sparsity pattern in a given dictionary. The construction of the dictionary leverages on a complete filter bank (RADWT) of $L_2(R)$ which guarantees a perfect reconstruction property and a tunable Q-factor. We based the two procedures on group penalized regression methods and we implemented them by a two-step iterative algorithm. The two methods differ in the second step iteration, the first applying Soft thresholding to the residual vectors, the second one applying Hard thresholding. Moreover, we made connections between them and the robust regression literature in a high dimensional setting obtaining interesting interpretations. From the computational point of view, we observed that the Hard thresholding procedure reveals some advantages to the Soft thresholding procedure, confirming some previous results on nonconvex robust regression.

Author Contributions: Conceptualization, D.D.C. and I.D.F.; methodology, D.D.C. and I.D.F.; validation, D.D.C. and I.D.F.; formal analysis, D.D.C. and I.D.F.; visualization, D.D.C. and I.D.F.; funding acquisition, I.D.F. All authors have read and agreed to the published version of the manuscript.

Funding: Project Piattaforma Tecnologica ADVISE-Regione Campania; project "TAILOR" (H2020-ICT-48 GA: 952215).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The pump dataset is part of a repository managed by the Delft University of Technology and is freely available at http://homepage.tudelft.nl/n9d04/occ/index.html accessed on 20 April 2021 (dataset 541: Delft pump AR app).

Acknowledgments: The authors are grateful to anonymous reviewers for improving the overall manuscript. We thank Anestis Antoniadis for the fruitful discussions that helped to improve the manuscript. This work was partially supported by the grant INdAM-GNCS Project 2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Agrawal, S.; Agrawal, J. Survey on Anomaly Detection using Data Mining Techniques. *Procedia Comput. Sci.* **2015**, *60*, 708–713. [CrossRef]
2. Fahim, M.; Sillitti, A. Anomaly Detection, Analysis and Prediction Techniques in IoT Environment: A Systematic Literature Review. *IEEE Access* **2019**, *7*, 81664–81681. [CrossRef]
3. Basora, L.; Olive, X.; Dubot, T. Recent Advances in Anomaly Detection Methods Applied to Aviation. *Aerospace* **2019**, *6*, 117. [CrossRef]
4. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly Detection: A Survey. *ACM Comput. Surv.* **2009**, *41*. [CrossRef]
5. Pimentel, M.A.F.; Clifton, D.A.; Clifton, L.; Tarassenko, L. A review of novelty detection. *Signal Process.* **2014**, *99*, 215–249. [CrossRef]
6. Ahmed, M.; Mahmood, A.N.; Hu, J. A survey of network anomaly detection techniques. *J. Netw. Comput. Appl.* **2016**, *60*, 19–31. [CrossRef]
7. Thudumu, S.; Branch, P.; Jin, J.; Singh, J.J. A comprehensive survey of anomaly detection techniques for high dimensional big data. *J. Big Data* **2020**, *7*, 42. [CrossRef]
8. Rousseeuw, P.; Hubert, M. Anomaly detection by robust statistics. *WIREs Data Mining Knowl. Discov.* **2018**, *8*, e1236. [CrossRef]
9. Hubert, M.; Rousseeuw, P.; Segaert, P. Multivariate functional outlier detection. *Stat. Methods Appl.* **2015**, *24*, 177–202. [CrossRef]
10. De Canditiis, D.; De Feis, I. Simultaneous nonparametric regression in RADWT dictionaries. *Comput. Stat. Data Anal.* **2019**, *134*, 36–57. [CrossRef]
11. Adler, A.; Elad, M.; Hel-Or, Y. Sparse Coding with Anomaly Detection. *J. Signal Process. Syst.* **2015**, *79*, 179–188. [CrossRef]
12. Pilastre, B.; Boussouf, L.; D'Escriv, S.; Tourneretca, J.Y. Anomaly detection in mixed telemetry data using a sparse representation and dictionary learning. *Signal Process.* **2020**, *168*, 1–10. [CrossRef]
13. She, Y.; Owen, A. Outlier Detection Using Nonconvex Penalized Regression. *J. Am. Stat. Assoc.* **2011**, *106*, 626–639. [CrossRef]
14. Loh, P.; Wainwright, M.J. Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *J. Mach. Learn. Res.* **2015**, *16*, 559–616.

15. Loh, P. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Stat.* **2017**, *45*, 866–896. [CrossRef]
16. Amato, U.; Antoniadis, A.; De Feis, I.; Gijbels, I. Penalised robust estimators for sparse and high-dimensional linear models. *Stat. Methods Appl.* **2021**, *30*, 1–48. [CrossRef]
17. Selesnick, I.W. Resonance-Based Signal Decomposition: A New Sparsity-Enabled Signal Analysis Method. *Signal Process.* **2011**, *91*, 2793–2809. [CrossRef]
18. Bayram, I.; Selesnick, I.W. Frequency-Domain design of overcomplete rational-dilation wavelet transform. *IEEE Trans. Signal Process.* **2009**, *57*, 2957–2972. [CrossRef]
19. Liu, H.; Zhang, J. *On the L1-Lq Regularized Regression*; Technical Report; Department of Statistics, Carnegie Mellon University: Pittsburgh, PA, USA, 2008.
20. Bach, F.R. Consistency of the Group Lasso and Multiple Kernel Learning. *J. Mach. Learn. Res.* **2008**, *9*, 1179–1225.
21. Liu, H.; Zhang, J. Estimation Consistency of the Group Lasso and its Applications. In *Artificial Intelligence and Statistics*; van Dyk, D., Welling, M., Eds.; PMLR: Clearwater Beach, FL, USA, 2009; Volume 5, pp. 376–383.
22. Yuan, M.; Lin, Y. Model selection and estimation in regression withgrouped variables. *J. R. Statist. Soc. B* **2006**, *68*, 49–67. [CrossRef]
23. Rousseeuw, P.J.; Leroy, A.M. *Robust Regression and Outlier Detection*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1987.
24. Huber, P. *Robust Statistics*; Wiley: New York, NY, USA, 1981.
25. Breheny, P.; Huang, J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat. Comput.* **2015**, *25*, 173–187. [CrossRef]
26. Breheny, P.; Huang, J. Penalized methods for bi-level variable selection. *Stat. Its Interface* **2009**, *2*, 369–380. [CrossRef]
27. Huang, J.; Breheny, P.; Ma, S. A selective review of group selection in high-dimensional models. *Stat. Sci.* **2012**, *27*, 481–499. [CrossRef] [PubMed]
28. Simon, N.; Tibshirani, R. Standardization and the group lasso penalty. *Stat. Sin.* **2012**, *22*, 983–1001. [CrossRef]
29. Ypma, A. Learning Methods for Machine Vibration Analysis and Health Monitoring. Ph.D. Thesis, Delft University of Technology, Delft, The Netherlands, 2001.
30. Tax, D.M.J.; Ypma, A.; Duin, R.P. Pump Failure Detection Using Support Vector Data Descriptions. In *International Symposium on Intelligent Data Analysis*; Advances in Intelligent Data Analysis, IDA 1999, Lecture Notes in Computer Science Volume 1642; Hand, D.J., Kok, J.N., Berthold, M.R., Eds.; Springer: Berlin/Heidelberg, Germany, 1999.