ROBUST ADAPTIVE SPARSE SYSTEM IDENTIFICATION BY USING WEIGHTED ℓ_1 BALLS AND MOREAU ENVELOPES

Konstantinos Slavakis[1] *Yannis Kopsinis*[2] *Sergios Theodoridis*[2]

[1]University of Peloponnese,
Dept. of Telecommunications Science and
Technology, Tripolis 22100, Greece.
Email: slavakis@uop.gr

[2]University of Athens,
Dept. of Informatics and Telecommunications,
Athens 15784, Greece.
Emails: kopsinis@ieee.org, stheodor@di.uoa.gr

ABSTRACT

This paper presents a novel approach to the time-recursive sparse system identification task by revisiting the classical Wiener-Hopf equation. The proposed methodology is built on the concept of the Moreau envelope of a convex function. The objective of employing such a convex analytic tool is twofold: i) it penalizes the deviations from the Wiener-Hopf equation, which are often met in practice due to outliers, model inaccuracies, etc, and ii) it fortifies the method against strongly correlated input signal samples. The resulting algorithm enjoys a clear geometrical description; the a-priori information on sparsity is exploited by the introduction of a sequence of weighted ℓ_1 balls, and the recursions are obtained by simple relations based on the generic tool of projections onto closed convex sets. The method is tested against the state-of-the-art batch and time-recursive techniques, and in several scenarios, which also include signal recovery tasks. The proposed design shows a competitive performance in cases where the model is corrupted by Gaussian noise, and excels in scenarios of non-Gaussian heavy-tailed noise processes, albeit at a higher complexity.

1. INTRODUCTION

First, let us introduce some notation. The set of all integers, nonnegative integers, positive integers, and real numbers are denoted by $\mathbb{Z}, \mathbb{Z}_{\geq 0}$, $\mathbb{Z}_{>0}$, and $\mathbb{R}$. Vector and matrix quantities are denoted by boldfaced symbols, and the operator $(\cdot)^t$ stands for vector/matrix transposition. Given two integers $j_1, j_2 \in \mathbb{Z}$, such that $j_1 \leq j_2$, let $\overline{j_1, j_2} := \{j_1, j_1+1, \ldots, j_2\}$. Throughout the manuscript, $n \in \mathbb{Z}_{\geq 0}$ denotes discrete time.

The present study focuses on the following task: given a sequence of observations $(\boldsymbol{x}_n)_{n\in\mathbb{Z}_{\geq 0}} \subset \mathbb{R}^L$, $L \in \mathbb{Z}_{>0}$, and $(y_n)_{n\in\mathbb{Z}_{\geq 0}} \subset \mathbb{R}$, that are related via the linear model:

$$y_n = \boldsymbol{x}_n^t \boldsymbol{h}_* + v_n, \quad \forall n \in \mathbb{Z}_{\geq 0}, \tag{1}$$

obtain an estimate of the unknown $\boldsymbol{h}_*$, which is assumed to be *sparse*. The sequence $(v_n)_{n\in\mathbb{Z}_{\geq 0}} \subset \mathbb{R}$ stands for a zero mean noise process. The input signal $(\boldsymbol{x}_n)_{n\in\mathbb{Z}_{>0}}$ is considered to be independent of $(v_n)_{n\in\mathbb{Z}_{\geq 0}}$. The approach, which we follow in this paper, is that of the time-adaptive nature. In other words, we search for algorithms that operate via simple recursive rules, in an online fashion, as the new training data are received at each time instant, $n \in \mathbb{Z}_{>0}$. This approach is in contrast to *batch* methods, where processing is performed on data blocks and off-line.

Next we provide with a few definitions related to the concept of sparsity. The support of a vector $\boldsymbol{h}$ is defined as $\operatorname{supp}(\boldsymbol{h}) := \{i \in \overline{1, L} :\ h_i \neq 0\}$, and the ℓ_0-norm of $\boldsymbol{h}$ is simply the cardinality of its support, i.e., $\|\boldsymbol{h}\|_{\ell_0} := \#\operatorname{supp}(\boldsymbol{h})$. The vector $\boldsymbol{h}_* \in \mathbb{R}^L$ is called *sparse* if $\|\boldsymbol{h}_*\|_{\ell_0} \ll L$. Sparsity-aware methods have been gaining, recently, an interest of exponential growth, due to the revolutionary point of view that the *Compressed Sensing* or *Sampling* (CS) framework [1] has brought into estimation tasks; if sparsity, which pervades a very large number of signal/system models, is appropriately utilized, usually by convex analysis, then it takes far fewer samples than it was traditionally necessary in order to perfectly reconstruct the sparse signal/system.

For the sake of illustration, let us assume here that $(\boldsymbol{x}_n)_{n\in\mathbb{Z}_{\geq 0}}$ is weakly stationary. It is well-known [2] that if we define $\boldsymbol{X}_n := [\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{L\times(n+1)}$, $\boldsymbol{y}_n := [y_0, \ldots, y_n]^t \in \mathbb{R}^{n+1}$, $\forall n \in \mathbb{Z}_{\geq 0}$, then the Least-Squares (LS) estimation task has the following solution set: $V := \arg\min_{\boldsymbol{h}\in\mathbb{R}^L} \mathsf{E}\|\boldsymbol{X}_n^t \boldsymbol{h} - \boldsymbol{y}_n\|^2 = \{\boldsymbol{h} \in \mathbb{R}^L : \boldsymbol{R}\boldsymbol{h} = \boldsymbol{r}\} \neq \emptyset$, where E stands for expectation, $\|\cdot\|$ for the Euclidean norm in $\mathbb{R}^L$, $\boldsymbol{R} := \mathsf{E}(\boldsymbol{x}_n \boldsymbol{x}_n^t)$, and $\boldsymbol{r} := \mathsf{E}(y_n \boldsymbol{x}_n)$, $\forall n \in \mathbb{Z}_{\geq 0}$. It is an easy task to verify by (1) that $\boldsymbol{h}_*$ belongs to the set V, which is defined by the *Wiener-Hopf* equation: $\boldsymbol{R}\boldsymbol{h} = \boldsymbol{r}$. The classical Recursive Least-Squares (RLS) algorithm [2] is built around this equation; having available a *certain* sequence of estimates $(\tilde{\boldsymbol{R}}_n, \tilde{\boldsymbol{r}}_n)_{n\in\mathbb{Z}_{\geq 0}}$ of $(\boldsymbol{R}, \boldsymbol{r})$, the RLS computes the inverse $\tilde{\boldsymbol{R}}_n^{-1}$ efficiently, and uses also the knowledge of $\tilde{\boldsymbol{r}}_n$ in order to obtain an estimate of $\boldsymbol{h}_*$, $\forall n \in \mathbb{Z}_{\geq 0}$.

A novel view of the least-squares rationale was introduced in [3]. Non-smooth convex analytic arguments [4] were used in order to penalize deviations from the Wiener-Hopf equation. Given the estimates $(\tilde{\boldsymbol{R}}_n, \tilde{\boldsymbol{r}}_n)$, such a penalization was achieved by user-defined, convex, and not necessary differentiable, loss functions which evaluate the deviation $\tilde{\boldsymbol{R}}_n \boldsymbol{h} - \tilde{\boldsymbol{r}}_n$, for any $\boldsymbol{h} \in \mathbb{R}^L$. Put in geometrical terms, an equivalent description of the previous penalization task is the construction of a sequence of closed convex sets $(S_n)_{n\in\mathbb{Z}_{\geq 0}}$ which contain $\boldsymbol{h}_*$ with high probability. In order to solve the associated convex feasibility task, i.e., find a point in $\bigcap_{n\geq n_0} S_n$, for some $n_0 \in \mathbb{Z}_{\geq 0}$, the study in [3] introduced a sparsity-aware, time-recursive algorithm, realized by simple iterations, with a computational complexity of order $\mathcal{O}(3L^2)$. A notable advantage of the proposed method was that the computation of a sequence of inverses $(\tilde{\boldsymbol{R}}_n^{-1})_{n\in\mathbb{Z}_{\geq 0}}$, as in the classical RLS, was no longer necessary. The methodology of [3] belongs to the rich algorithmic family of [5–7]. In such a way, convergence results, based on very recent advances of time-adaptive convex feasibility tasks [7] can be obtained. More importantly, the study of [3] benefits from the rich variety of tools for exploiting the available a-priori information [7], such as the sparsity of $\boldsymbol{h}_*$. The method showed excellent performance for sparse signal recovery tasks [3].

However, it was observed that the performance of the method in [3] deteriorates when applied to system identification tasks where the input signal samples $(\boldsymbol{x}_n)_{n\in\mathbb{Z}_{\geq 0}}$ show strong correlation. The goal of the present paper is to propose a novel technique in order to remedy this sensitivity of [3]. The new approach is based on the notion of the Moreau envelope of a convex function [4,8], which has been very recently popularized in signal processing tasks [9,10]. To validate this new methodology, a series of experiments is presented for both signal recovery and system identification tasks. Indeed, the new methodology shows competitive behavior when compared to the state-of-the-art batch and time-adaptive techniques in scenarios where the noise process $(v_n)_{n\in\mathbb{Z}_{\geq 0}}$ of (1) is Gaussian, and excels in scenarios where $(v_n)_{n\in\mathbb{Z}_{\geq 0}}$ becomes heavy-tailed and non-Gaussian.

2. PENALIZING THE DEVIATIONS FROM THE WIENER-HOPF EQUATION

In this section, we will present in short the basic ingredient behind the methodology of [3]. We will also give a geometrical explanation for the reason behind the sensitivity of [3] to system identification tasks where the input signal samples show strong correlation.

Assume that available to us, at the time instant n, are the estimates $\tilde{\boldsymbol{R}}_n$ and $\tilde{\boldsymbol{r}}_n$ of $\boldsymbol{R}$ and $\boldsymbol{r}$, respectively. For example, in this study, we will use $\forall n \in \mathbb{Z}_{\geq 0}$, $\tilde{\boldsymbol{R}}_n := \frac{1}{N}\sum_{j=n-N+1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^t$, $\tilde{\boldsymbol{r}}_n := \frac{1}{N}\sum_{j=n-N+1}^{n} y_j \boldsymbol{x}_j$, where $N \in \mathbb{Z}_{>0}$ denotes the size of a window.

Having at our disposal the estimates $(\tilde{\boldsymbol{R}}_n, \tilde{\boldsymbol{r}}_n)$, and by mimicking the Wiener-Hopf equation, a natural choice for a place to look for $\boldsymbol{h}_*$ is the affine set $V_n := \{\boldsymbol{h} \in \mathbb{R}^L : \tilde{\boldsymbol{R}}_n \boldsymbol{h} = \tilde{\boldsymbol{r}}_n\}$. In general, there is no guarantee that $\boldsymbol{h}_*$ belongs to V_n. Outliers, model inaccuracies, as well as calibration errors may result into estimates $(\tilde{\boldsymbol{R}}_n, \tilde{\boldsymbol{r}}_n)$ which form a V_n that deviates from the one generated by the classical Wiener-Hopf equation, i.e., $V := \{\boldsymbol{h} \in \mathbb{R}^L : \boldsymbol{R}\boldsymbol{h} = \boldsymbol{r}\}$. It is therefore desirable to "enlarge" V_n to a set S_n that accommodates such deviations, shows robustness to outliers and inaccuracies, which are often met in practice, particularly for small values of n, and increases the probability of having $\boldsymbol{h}_*$ lying into S_n, $\forall n \in \mathbb{Z}_{\geq 0}$.

Let any convex loss function $\mathcal{L} : \mathbb{R}^L \to \mathbb{R}$. Motivated by the path of *robust statistics* [11], we will enhance $\mathcal{L}$ with robustness properties. First of all, for every n, define the composite function $\Theta_n(\boldsymbol{h}) := \mathcal{L}(\tilde{\boldsymbol{R}}_n \boldsymbol{h} - \tilde{\boldsymbol{r}}_n)$, $\forall \boldsymbol{h} \in \mathbb{R}^L$. Notice that since $\mathcal{L}$ is convex, and $\tilde{\boldsymbol{R}}_n \boldsymbol{h} - \tilde{\boldsymbol{r}}_n$ is an affine transformation of $\boldsymbol{h}$, the loss function Θ_n is also convex [4]. Now, given a tolerance $\varepsilon \geq 0$, let us introduce here the *ε-insensitive* version of Θ_n; $\Theta_n^{(\varepsilon)} : \mathbb{R}^L \to [0,\infty) : \boldsymbol{h} \mapsto \max\{0, \Theta_n(\boldsymbol{h}) - \varepsilon\}$. The 0-*th level set* of $\Theta_n^{(\varepsilon)}$ is defined as: $\forall n \in \mathbb{Z}_{\geq 0}$, $\operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)} := \{\boldsymbol{h} \in \mathbb{R}^L : \Theta_n^{(\varepsilon)}(\boldsymbol{h}) \leq \varepsilon\}$. If $\mathcal{L}$ and ε are chosen such that $\mathcal{L}(\boldsymbol{0}) \leq \varepsilon$, then clearly $V_n \subset \operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)}$. Hence, $S_n := \operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)}$ serves as a candidate for the "enlarged" V_n. In other words, we allow our solution set to be larger than V_n. The "shape" of $\operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)}$ is dictated by the choice of $\mathcal{L}$. A standard choice for $\mathcal{L}$ is the quadratic function $\mathcal{L} = \frac{1}{2}\|\cdot\|^2$. Nevertheless, the study in [3] gave the freedom to employ any convex and not necessarily differentiable $\mathcal{L}$, provided that the associated subgradients (see Section 3.2) exist and are available in closed form. In such a way, the choice of $\mathcal{L}$ is not constrained by the differentiability condition, and the designer can choose from a large variety of convex functions in order to penalize the deviations $\tilde{\boldsymbol{R}}_n \boldsymbol{h} - \tilde{\boldsymbol{r}}_n$, like any ℓ_p-norm, with $p \in \mathbb{Z}_{>0}$, the ℓ_∞-norm, the negative log function, the Huber loss, etc.

The methodology of [3] exhibited excellent performance in signal recovery tasks, where the input signal $(\boldsymbol{x}_n)_{n \in \mathbb{Z}_{\geq 0}}$ becomes an i.i.d. vector-valued random process. However, it deteriorates when applied to system identification problems, in scenarios where the input signal $(\boldsymbol{x}_n)_{n \in \mathbb{Z}_{\geq 0}}$ is a strongly correlated process. An explanation for such a sensitivity comes from the shape of the 0-th level set $\operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)}$. For example, if one chooses the quadratic function, as the basic module $\mathcal{L}$, then for a strongly correlated input signal, the condition number of the matrix $\tilde{\boldsymbol{R}}_n$ is large, and $\operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)}$ becomes highly elongated. Solving a convex feasibility problem, i.e., finding a point in $\bigcap_{n \geq n_0} \operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)}$, for some $n_0 \in \mathbb{Z}_{\geq 0}$, may result into a point that is located far from the desired $\boldsymbol{h}_*$, in spite of the fact that the latter is also contained, with high probability, in the elongated $\operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)}$, for a large number of $n \in \mathbb{Z}_{\geq 0}$.

Hence, to remedy such an unpleasant situation, a solution would be to "blow up" the elongated $(\operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)})_{n \in \mathbb{Z}_{\geq 0}}$. Naturally, the following question arises: how can we "blow up" the closed convex sets $(\operatorname{lev}_{\leq 0} \Theta_n^{(\varepsilon)})_{n \in \mathbb{Z}_{\geq 0}}$ in a way that respects their original shape, or in other words, the characteristics of the matrix $\tilde{\boldsymbol{R}}_n$, and, more importantly, escapes from the classical approach of diagonally loading $\tilde{\boldsymbol{R}}_n$, $\forall n \in \mathbb{Z}_{\geq 0}$? The Moreau envelope [4,8] of a convex function gives an answer to this question.

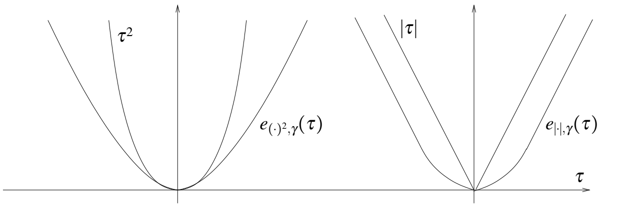


Figure 1: Illustration of the Moreau envelope for two simple examples of convex functions: i) the differentiable quadratic function, and ii) the non-differentiable absolute value function. Notice that the Moreau envelope of the absolute value function is a scaled version of the well-known Huber function [12], widely used in robust statistics [11].

3. A FEW ELEMENTS OF CONVEX ANALYSIS

3.1 Moreau envelopes.

Given a convex function $\Theta : \mathbb{R}^L \to \mathbb{R}$, its *Moreau envelope* [4,8–10], for some $\gamma > 0$, is defined as the function

$$e_{\Theta,\gamma}(\boldsymbol{h}) := \min_{\boldsymbol{v} \in \mathbb{R}^L} \left(\Theta(\boldsymbol{v}) + \frac{1}{2\gamma}\|\boldsymbol{h} - \boldsymbol{v}\|^2 \right), \quad \forall \boldsymbol{h} \in \mathbb{R}^L. \tag{2}$$

An illustration of the Moreau envelope is given in Fig. 1. It turns out [8–10] that the minimizer of (2) is unique, so that one can define the *proximity mapping of index γ of* Θ [8–10] as follows:

$$\operatorname{prox}_{\gamma\Theta}(\boldsymbol{h}) := \arg\min_{\boldsymbol{v} \in \mathbb{R}^L} \left(\Theta(\boldsymbol{v}) + \frac{1}{2\gamma}\|\boldsymbol{h} - \boldsymbol{v}\|^2 \right), \quad \forall \boldsymbol{h} \in \mathbb{R}^L. \tag{3}$$

Fact 1 (Selected properties of the Moreau envelope [4, 9, 10])**.** Given a convex function $\Theta : \mathbb{R}^L \to \mathbb{R}$, its Moreau envelope $e_{\Theta,\gamma} : \mathbb{R}^L \to \mathbb{R}$ satisfies the following properties.

1. $\forall \gamma > 0$, and $\forall \boldsymbol{h} \in \mathbb{R}^L$, we have $e_{\Theta,\gamma}(\boldsymbol{h}) \leq \Theta(\boldsymbol{h})$.
2. The Moreau envelope $e_{\Theta,\gamma}$ converges pointwise to Θ as $\gamma \to 0$, i.e., $\lim_{\gamma \to 0} e_{\Theta,\gamma}(\boldsymbol{h}) = \Theta(\boldsymbol{h})$, $\forall \boldsymbol{h} \in \mathbb{R}^L$.
3. The Moreau envelope $e_{\Theta,\gamma}$ is differentiable with $\nabla e_{\Theta,\gamma}(\boldsymbol{h}) = (\boldsymbol{h} - \operatorname{prox}_{\gamma\Theta}(\boldsymbol{h}))/\gamma$, $\forall \boldsymbol{h} \in \mathbb{R}^L$. □

For example, it can be easily verified, by a simple differentiation, that the proximity mapping, of index $\gamma > 0$, which relates to the function:

$$\Theta_n(\boldsymbol{h}) := \frac{1}{2}\|\tilde{\boldsymbol{R}}_n \boldsymbol{h} - \tilde{\boldsymbol{r}}_n\|^2, \quad \forall \boldsymbol{h} \in \mathbb{R}^L, \forall n \in \mathbb{Z}_{\geq 0}, \tag{4}$$

is

$$\operatorname{prox}_{\gamma\Theta_n}(\boldsymbol{h}) = \left(\tilde{\boldsymbol{R}}_n^2 + \frac{1}{\gamma}\boldsymbol{I} \right)^{-1} \left(\frac{\boldsymbol{h}}{\gamma} + \tilde{\boldsymbol{R}}_n \tilde{\boldsymbol{r}}_n \right). \tag{5}$$

The previous expression of the proximity mapping helps us not only to calculate the value of the Moreau envelope $e_{\Theta_n,\gamma}$ at a point $\boldsymbol{h}$, but also its differential $\nabla e_{\Theta_n,\gamma}(\boldsymbol{h})$, by employing Fact 1.3.

3.2 Subgradients.

Given a convex function $\Theta : \mathbb{R}^L \to \mathbb{R}$, the subdifferential $\partial\Theta$ is defined as the set-valued mapping: $\boldsymbol{h} \mapsto \partial\Theta(\boldsymbol{h}) := \{\boldsymbol{y} \in \mathbb{R}^L : \forall \boldsymbol{v} \in \mathbb{R}^L, \boldsymbol{y}^t(\boldsymbol{v} - \boldsymbol{h}) + \Theta(\boldsymbol{h}) \leq \Theta(\boldsymbol{v})\}$. In the case where Θ is continuous at $\boldsymbol{h}$, then $\partial\Theta(\boldsymbol{h}) \neq \varnothing$ [4]. Any element in $\partial\Theta(\boldsymbol{h})$ will be called a *subgradient* of Θ at $\boldsymbol{h}$, and will be denoted by $\Theta'(\boldsymbol{h})$. If Θ is differentiable at $\boldsymbol{h}$, then $\partial\Theta(\boldsymbol{h})$ becomes a singleton, and the unique element of $\partial\Theta(\boldsymbol{h})$ is nothing but the classical differential of Θ at $\boldsymbol{h}$. Notice, also, that $\mathbf{0} \in \partial\Theta(\boldsymbol{h}) \Leftrightarrow \boldsymbol{h} \in \arg\min_{\boldsymbol{v}\in\mathbb{R}^L} \Theta(\boldsymbol{v})$.

Hence, the subdifferential mapping of the ε-insensitive version of the Moreau envelope $e_{\Theta_n,\gamma}$ becomes as follows:

$$\partial e^{(\varepsilon)}_{\Theta,\gamma}(\boldsymbol{h}) = \begin{cases} \{\mathbf{0}\}, & \text{if } e_{\Theta,\gamma}(\boldsymbol{h}) < \varepsilon, \\ \{\tau \nabla e_{\Theta,\gamma}(\boldsymbol{h}) : \tau \in [0,1]\}, & \text{if } e_{\Theta,\gamma}(\boldsymbol{h}) = \varepsilon, \\ \{\nabla e_{\Theta,\gamma}(\boldsymbol{h})\}, & \text{if } e_{\Theta,\gamma}(\boldsymbol{h}) > \varepsilon. \end{cases} \tag{6}$$

The previous results can be reproduced by using standard arguments of convex analysis, e.g., [4].

3.3 Projection mappings onto closed convex sets.

Given any nonempty *closed convex* set $C \subset \mathbb{R}^L$, the *(metric) projection onto C* is defined as the mapping $P_C : \mathbb{R}^L \to C$ which takes any $\boldsymbol{h} \in \mathbb{R}$ to the (unique) point in C that lies the closest from $\boldsymbol{h}$, i.e., $\|\boldsymbol{h} - P_C(\boldsymbol{h})\| = \inf_{\boldsymbol{v}\in C} \|\boldsymbol{h} - \boldsymbol{v}\|$. The *relaxed (metric) projection mapping* is defined as: $T_C^{(\alpha)} := I + \alpha(P_C - I)$, $\alpha \in (0,2)$, where I denotes the identity mapping in $\mathbb{R}^L$.

The *weighted ℓ_1 ball* is defined as the closed convex set: $B_{\ell_1}[\boldsymbol{w},\rho] := \{\boldsymbol{h} \in \mathbb{R}^L : \sum_{i=1}^L w_i|h_i| \le \rho\}$, where $\boldsymbol{w} := [w_1,\ldots,w_L]^t \in \mathbb{R}^L$ has positive components, and ρ is a positive number denoting the radius of the weighted ball. For example, the standard ℓ_1 ball is nothing but $B_{\ell_1}[\mathbf{1},\rho]$, where $\mathbf{1} \in \mathbb{R}^L$ is a vector of unities. The closed form expression of the metric projection $P_{B_{\ell_1}[\boldsymbol{w},\rho]}$ can be found in [13].

4. THE ALGORITHM

So far, we introduced all the necessary tools on which our new algorithm will be built on. In this section, we will introduce a simple recursion which generates a sequence of estimates $(\boldsymbol{h}_n)_{n\in\mathbb{Z}_{\ge 0}}$ for approximating the desired sparse $\boldsymbol{h}_*$.

4.1 Learning from the training data

The sequentially arriving training data $(\boldsymbol{x}_n, y_n)_{n\in\mathbb{Z}_{\ge 0}}$ are exploited in order to form an associated sequence of estimates $(\tilde{\boldsymbol{R}}_n, \tilde{\boldsymbol{r}}_n)_{n\in\mathbb{Z}_{\ge 0}}$ of $(\boldsymbol{R},\boldsymbol{r})$. According to the discussion in Sections 2 and 3, we adopt the following steps: a) we choose as the seed convex function the LS one, $\mathcal{L} := \|\cdot\|^2/2$, b) we form the following sequence of composite functions $\Theta_n(\boldsymbol{h}) := \mathcal{L}(\tilde{\boldsymbol{R}}_n\boldsymbol{h} - \tilde{\boldsymbol{r}}_n)$, $\forall \boldsymbol{h} \in \mathbb{R}^L$, $\forall n \in \mathbb{Z}_{\ge 0}$, given in (4), c) we form their Moreau envelopes, $(e_{\Theta_n,\gamma})_{n\in\mathbb{Z}_{\ge 0}}$, of some index $\gamma > 0$, and finally d) we construct their ε-insensitive versions $(e^{(\varepsilon)}_{\Theta_n,\gamma})_{n\in\mathbb{Z}_{\ge 0}}$, for some user-defined $\varepsilon > 0$. The values of $(e^{(\varepsilon)}_{\Theta_n,\gamma})_{n\in\mathbb{Z}_{\ge 0}}$ as well as the choices for a subgradient ${e^{(\varepsilon)}_{\Theta_n,\gamma}}'(\boldsymbol{h})$, at some point $\boldsymbol{h}$, can be obtained by the discussion in Section 3, and more specifically, by (5), Fact 1.3, and (6).

4.2 Exploiting sparsity

As it was observed in [1,14,15], and as we also verified in [13] in the case of an unknown system that we know it is sparse, the recursive re-weighting of an ℓ_1-norm term improves not only the convergence speed of the algorithm, but also decreases its mis-adjustment level. We will follow the same approach also in this study and consider a sequence $(B_{\ell_1}[\boldsymbol{w}_n,\rho_n])_{n\in\mathbb{Z}_{\ge 0}}$, where ρ_n stands for the radius of the weighted ℓ_1 balls, $\forall n \in \mathbb{Z}_{\ge 0}$. To this end, for a time instant $n \in \mathbb{Z}_{\ge 0}$, and assuming that we have at our disposal the current estimate $\boldsymbol{h}_n$, we inductively define $\boldsymbol{w}_n$, and thus $B_{\ell_1}[\boldsymbol{w}_n,\rho_n]$, as follows: $w_{n,i} := 1/(\max\{|h_{n,i}|,\check{\varepsilon}\})$, $\forall i \in \overline{1,L}$, $\forall n \in \mathbb{Z}_{\ge 0}$, where $\check{\varepsilon}$ is a user-defined sufficiently small positive parameter, introduced in order to avoid divisions by zeros. By following similar arguments to the ones in [13], it can be shown that the previous choice of $(\boldsymbol{w}_n)_{n\in\mathbb{Z}_{\ge 0}}$ leads to an interpretation of the parameter ρ_n, $n \in \mathbb{Z}_{\ge 0}$, as an estimate of $\|\boldsymbol{h}_*\|_{\ell_0}$. Such a perspective suggests that any information about $\|\boldsymbol{h}_*\|_{\ell_0}$, obtained either a-priori or in an online fashion, can be used in the definition of the sequence of parameters $(\rho_n)_{n\in\mathbb{Z}_{\ge 0}}$. More accurately, any over-estimation of $\|\boldsymbol{h}_*\|_{\ell_0}$ can serve as a candidate for ρ_n, $n \in \mathbb{Z}_{\ge 0}$, since if $\rho \le \rho'$, then $B_{\ell_1}[\boldsymbol{w},\rho] \subset B_{\ell_1}[\boldsymbol{w},\rho']$. To save space, the present study will assume that the information about $\|\boldsymbol{h}_*\|_{\ell_0}$, which enters the design via ρ_n, $n \in \mathbb{Z}_{\ge 0}$, is available a-priori. A scheme for employing information about $\|\boldsymbol{h}_*\|_{\ell_0}$, that is obtained in an online fashion, is deferred to a future work.

4.3 The recursion

Any element of $\mathbb{R}^L$ is suitable for the starting point $\boldsymbol{h}_0 \in \mathbb{R}^L$. Given the current estimate $\boldsymbol{h}_n$, $\boldsymbol{h}_{n+1}$ is computed as follows:

$$\boldsymbol{h}_{n+1} := \begin{cases} T^{(\alpha_n)}_{B_{\ell_1}[\boldsymbol{w}_n,\rho_n]}\left(\boldsymbol{h}_{n+1} - \lambda_n \dfrac{e^{(\varepsilon)}_{\Theta_n,\gamma}(\boldsymbol{h}_n)}{\|{e^{(\varepsilon)}_{\Theta_n,\gamma}}'(\boldsymbol{h}_n)\|^2} {e^{(\varepsilon)}_{\Theta_n,\gamma}}'(\boldsymbol{h}_n)\right), & \\ & \text{if } {e^{(\varepsilon)}_{\Theta_n,\gamma}}'(\boldsymbol{h}_n) \neq \mathbf{0}, \\ T^{(\alpha_n)}_{B_{\ell_1}[\boldsymbol{w}_n,\rho_n]}(\boldsymbol{h}_n), & \text{if } {e^{(\varepsilon)}_{\Theta_n,\gamma}}'(\boldsymbol{h}_n) = \mathbf{0}, \end{cases} \tag{7}$$

where $T^{(\alpha_n)}_{B_{\ell_1}[\boldsymbol{w}_n,\rho_n]}$ is the relaxed projection mapping, defined in Section 3.3, and ${e^{(\varepsilon)}_{\Theta_n,\gamma}}'(\boldsymbol{h}_n)$ stands for any subgradient of $e^{(\varepsilon)}_{\Theta_n,\gamma}$ at $\boldsymbol{h}_n$. Both of the user-defined parameters $\alpha_n, \lambda_n \in (0,2)$.

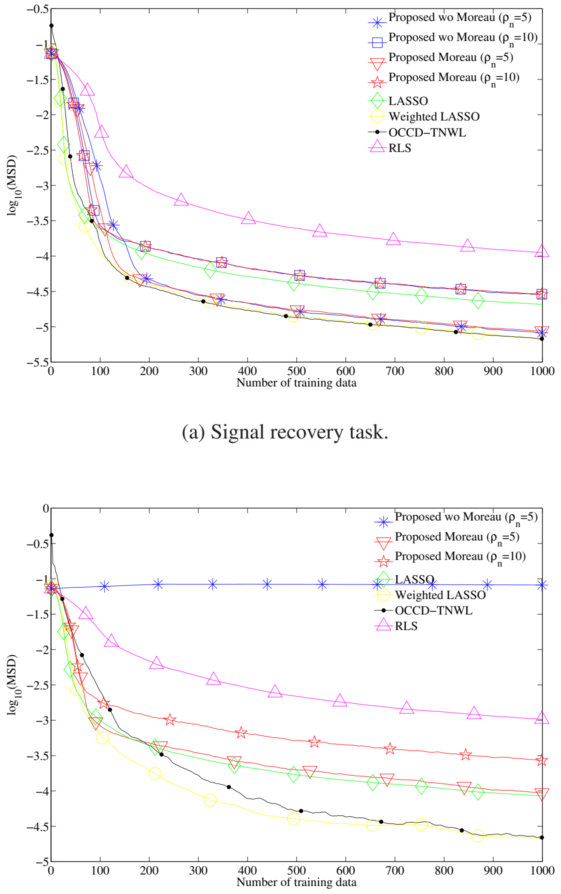
It turns out that the basic recursion (7) obtains a simple geometrical interpretation. The term in the large parenthesis of (7) is nothing but the relaxed subgradient projection mapping [7] with respect to the function $e^{(\varepsilon)}_{\Theta_n,\gamma}$. This mapping has the following remarkable property; given the current estimate $\boldsymbol{h}_n$, the subgradient projection mapping takes $\boldsymbol{h}_n$ closer to the 0-th level set $\operatorname{lev}_{\le 0} e^{(\varepsilon)}_{\Theta_n,\gamma}$, if $\boldsymbol{h}_n \notin \operatorname{lev}_{\le 0} e^{(\varepsilon)}_{\Theta_n,\gamma}$, and leaves $\boldsymbol{h}_n$ unaffected if $\boldsymbol{h}_n \in \operatorname{lev}_{\le 0} e^{(\varepsilon)}_{\Theta_n,\gamma}$. Afterwords, the relaxed projection mapping onto the weighted ℓ_1 ball $B_{\ell_1}[\boldsymbol{w}_n,\rho_n]$ is applied.

The main contribution to the complexity of the proposed algorithm comes from the matrix inversion in (5). Currently, efficient schemes to reduce the computational complexity are under investigation. We note here that the algorithm introduced in [3], where the Moreau envelope is not employed, and a matrix inversion is not necessary, is of order $\mathcal{O}(3L^2)$.

The recursion (7) belongs to the rich algorithmic family of [7]. In this way, (7) benefits from the general convergence analysis results, and the large variety of a-priori information usage found in the very recent study of [7]. A detailed description of this family, by using simple geometrical arguments, for linear and non-linear estimation tasks is given in [16]. A different philosophy, than the one presented in this study, which also exploits sparsity and the Moreau envelope in time-recursive algorithms, can be found in [17].

5. NUMERICAL EXAMPLES

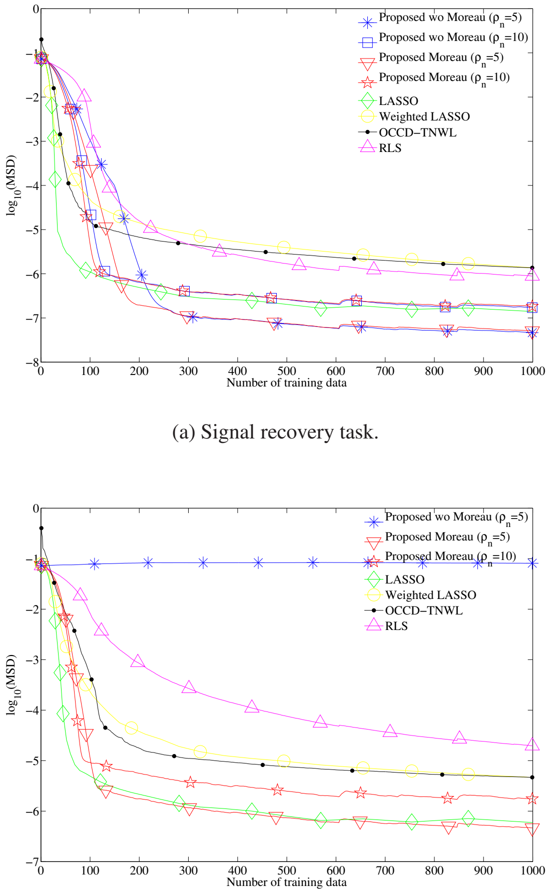
This section validates the proposed methodology for the two fundamental tasks of signal recovery and system identification. Moreover, two scenarios will be followed for the additive noise process $(v_n)_{n\in\mathbb{Z}_{\ge 0}}$ in (1); one where the noise is Gaussian, and another where it obeys the heavy-tailed student's-t distribution [18]. The proposed design will be validated against both batch and time-recursive techniques. Although the batch techniques do not fit in the time-adaptive rationale of this manuscript, since they perform computation in an off-line fashion, they were chosen as benchmarks against which the performance of the novel algorithm will be tested in the subsequent figures. Hence, the classical LASSO [19], and its variant, the weighted LASSO [14] were considered. As for the time-adaptive schemes, the classical RLS [2], and the RLS-based OCCD-TNWL [15] were employed. In the following figures, the tag "Proposed wo Moreau" corresponds to the introduced method without using the Moreau envelope, i.e., the case where $\Theta_n^{(\varepsilon)}$ takes the place of $e^{(\varepsilon)}_{\Theta_n,\gamma}$ in (7), and which was introduced in [3]. Clearly, the



(a) Signal recovery task.

(b) System identification task.

Figure 2: The i.i.d. noise process $(v_n)_{n\in\mathbb{Z}_{\geq 0}}$ follows the Gaussian distribution with zero mean and variance equal to 0.1. The unknown $\boldsymbol{h}_* \in \mathbb{R}^{100}$, with $\|\boldsymbol{h}_*\|_{\ell_0} := 5$.



(a) Signal recovery task.

(b) System identification task.

Figure 3: The i.i.d. noise process $(v_n)_{n\in\mathbb{Z}_{\geq 0}}$ follows the heavy-tailed student's-t distribution [18] with zero mean, variance equal to 0.1, and degree of freedom $\nu := 2.001$.

tag "Proposed Moreau" associates to the proposed method with the Moreau envelope approach. Moreover, the Mean Square Deviation (MSD) is defined as $\mathrm{MSD}(n) := \frac{1}{LQ}\sum_{q=1}^{Q}\|\boldsymbol{h}_* - \boldsymbol{h}_n^{(q)}\|^2$, $\forall n \in \mathbb{Z}_{\geq 0}$, where Q is the total number of independent runs of the experiment. Here, $Q := 100$. Here, the case of $L := 100$ is considered, i.e., the unknown sparse signal/system $\boldsymbol{h}_* \in \mathbb{R}^{100}$. Moreover, we let $\|\boldsymbol{h}_*\|_{\ell_0} := 5$. For each realization of the experiment, the non-zero components of $\boldsymbol{h}_*$ are drawn from a Gaussian distribution of zero mean and variance equal to 1. Their locations, within $\boldsymbol{h}_*$, are defined randomly for each realization. All of the employed methods were carefully tuned to produce their best performance for each adopted scenario.

For all the subsequent numerical examples, the following values were assigned to the parameters of the proposed method; ϵ, met in Section 2, takes the value of 0, the α_n, which defines the relaxed projection mapping of Section 3.3, is set equal to 0.75, $\forall n \in \mathbb{Z}_{\geq 0}$, the relaxation parameter λ_n of (7) takes the value of 1, $\forall n \in \mathbb{Z}_{\geq 0}$, and $\check{\epsilon}$, of Section 4.2, becomes 10^{-6}. For this paper, the size N of the sliding window, met in Section 2, is set equal to the total number of the training data. Similarly, the employed RLS-based techniques assume infinite memory with respect to the training data, and set the value of their forgetting factor equal to 1. The value of 0.75 was chosen for α_n in order to realize an *under-relaxed*, i.e., $\alpha_n < 1$, projection mapping $T_{B_{\ell_1}[\boldsymbol{w}_n,\rho_n]}^{(\alpha_n)}$. We adopted such a conservative approach, towards the weighted ℓ_1 ball, for obtaining smooth curves in all of the figures in this study. We noticed that values of $\alpha_n \geq 1$, i.e., the exact or an *over-relaxation* of the projection onto the weighted ℓ_1 ball, offer sequence of estimates with a fast speed of convergence, at the expense of non-smooth curves in the figures. Regarding ϵ, although the choice of $\epsilon := 0$ seems rather restrictive, we did so since such a choice draws some interesting implications in the numerical examples of Fig. 3. Moreover, we noticed that the proposed method resulted into similar behavior for values of $\check{\epsilon}$ within the interval $[10^{-3}, 10^{-10}]$; hence the choice of $\check{\epsilon} := 10^{-6}$.

Fig. 2 refers to the case where the noise $(v_n)_{n\in\mathbb{Z}_{\geq 0}}$ is an i.i.d. Gaussian process, with zero mean and variance equal to 0.1. In Fig. 2a, the signal recovery problem is considered. The input signal $(\boldsymbol{x}_n)_{n\in\mathbb{Z}}$ is defined as a discrete-time vector-valued Gaussian process of zero mean, such that the components of each $\boldsymbol{x}_n$ are mutually independent, with variance equal to 1. In addition, the index for the Moreau envelope is set equal to $\gamma := 10$. It is worth noticing that both the Moreau envelope approach as well as the proposed technique, without the Moreau regularization, perform equally well for the signal recovery task.

Next, in Fig. 2b, is the system identification task. The input signal $(\boldsymbol{x}_n)_{n\in\mathbb{Z}_{\geq 0}}$ becomes, now, $\boldsymbol{x}_n := [x_n, \ldots, x_{n-L+1}]^T$, where $(x_n)_{n\in\mathbb{Z}_{\geq 0}}$ is a strongly correlated Auto-Regressive (AR) process, given by $x_n := -0.9x_{n-1} + \sqrt{1-0.9^2}\xi_n$, $\forall n \in \mathbb{Z}_{\geq 0}$, and $(\xi_n)_{n\in\mathbb{Z}_{\geq 0}}$ is an i.i.d. Gaussian process, with zero mean and variance equal to

1. Due to the high correlation of the input signal, in this signal identification task, the index of the Moreau envelope is set equal to $\gamma := 10^4$. We stress here that the larger the index γ, the more dominant the Moreau regularization in (2) is. It is natural to ask for a stronger regularization, than in the signal recovery case of Fig. 2a, in order to overcome the difficulties imposed by the highly correlated AR input signal. Indeed, without the Moreau regularization, the proposed method fails to produce an acceptable performance for this sparse system identification problem. The introduced Moreau approach produces a performance which is inferior to the OCCD-TNWL. We noticed, however, that the proposed method, both with and without the Moreau regularization, resulted into a similar performance to the OCCD-TNWL for weakly correlated input signal samples.

Fig. 3 examines the case where the additive noise process $(v_n)_{n\in\mathbb{Z}_{\geq 0}}$ becomes non-Gaussian, and, in particular, heavy-tailed. For such a reason, the model (1) becomes prone to outliers. To realize such a heavy-tailed distribution, the zero mean student's-t distribution [18] was considered, with $\nu := 2.001$ degrees of freedom. The closer ν goes to 2, the heavier the tails of this pdf are. Its variance was set equal to 0.1. In Fig. 3a, the signal recovery task of the sparse signal $\boldsymbol{h}_* \in \mathbb{R}^{100}$, with $\|\boldsymbol{h}_*\|_{\ell_0} := 5$ is examined. As in the Gaussian noise case, the index of the Moreau envelope takes the value of $\gamma := 10$. It is easy to notice, by Fig. 3a, that the proposed method performs remarkably well against outliers caused by the heavy-tailed noise. The performance of the proposed method remains very close to the batch LASSO, even when the over-estimation of the support of $\boldsymbol{h}_*$ is up to 100%, i.e., 10 instead of the actual 5.

For the case of the sparse system identification problem, depicted in Fig. 3b, the input signal follows the strongly correlated AR signal of the Gaussian noise case, met in Fig. 2. The index of the Moreau envelope takes the value of $\gamma := 10^4$. As Fig. 3b demonstrates, although the lack of the Moreau regularization results into a failure of the proposed method, the introduction of the Moreau envelope shows the best performance among all the employed techniques. The performance remains remarkably robust even when there is a large ambiguity on the over-estimation of the support of $\boldsymbol{h}_*$, i.e., the case where $\rho_n := 10$, $\forall n \in \mathbb{Z}_{\geq 0}$. As in Fig. 2, we noticed that also for the scenario of heavy-tailed noise, the proposed method, both with and without the Moreau regularization, resulted into a similar performance to the OCCD-TNWL for weakly correlated input signal samples.

Regarding the case of Fig. 3, there are some interesting implications drawn from the choice of $\epsilon := 0$. It can be verified, by the properties of the Moreau envelope, that for $\epsilon := 0$, $\mathrm{lev}_{\leq 0}\, e^{(0)}_{\Theta_n,\gamma} = V_n = \{\boldsymbol{h} \in \mathbb{R}^L : \tilde{\boldsymbol{R}}_n \boldsymbol{h} = \tilde{\boldsymbol{r}}_n\}$. We have also seen in Section 2 that $\forall n \in \mathbb{Z}_{\geq 0}$, the set V_n is an approximation of $V := \{\boldsymbol{h} \in \mathbb{R}^L : \boldsymbol{R}\boldsymbol{h} = \boldsymbol{r}\}$, defined by the Wiener-Hopf equation, and around which all the RLS-based algorithms revolve. Fig. 3 clearly implies that even if the proposed method aims to the same solution set as all the RLS-based methods do, since we set $\epsilon := 0$, the way that V_n is handled shows a remarkable robustness against non-Gaussian, heavy-tailed noise processes, as opposed to the sensitivity that the RLS-based techniques demonstrate.

6. CONCLUSIONS

Based on a new perspective of the classical Wiener-Hopf equation, the present study introduced a convex analytic framework in order to fortify projection-based time-recursive algorithms against system identification tasks with strongly correlated input signal samples. The basic tool is the Moreau envelope of a convex function. The proposed methodology enjoys a clear geometrical description, and belongs to the rich algorithmic frame of [5–7]. In this way, it benefits from the general convergence analysis results, and the large variety of a-priori information usage found in the very recent study of [7]. Such a convergence analysis, efficient methods to reduce the computational complexity of the proposed method, and numerical examples for a larger collection of signal/system scenarios are deferred to a future work.

REFERENCES

[1] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877–905, 2008.

[2] A. H. Sayed, *Fundamentals of Adaptive Filtering*, John Wiley & Sons, New Jersey, 2003.

[3] K. Slavakis, Y. Kopsinis, and S. Theodoridis, "Revisiting adaptive least-squares estimation and application to online sparse signal recovery," in *Proceedings of the IEEE ICASSP*, Prague: Czech Republic, May 2011, pp. 4292–4295.

[4] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 2004.

[5] I. Yamada and N. Ogura, "Adaptive Projected Subgradient Method for asymptotic minimization of sequence of nonnegative convex functions," *Numerical Functional Analysis and Optimization*, vol. 25, no. 7&8, pp. 593–617, 2004.

[6] K. Slavakis, I. Yamada, and N. Ogura, "The Adaptive Projected Subgradient Method over the fixed point set of strongly attracting nonexpansive mappings," *Numerical Functional Analysis and Optimization*, vol. 27, no. 7&8, pp. 905–930, 2006.

[7] K. Slavakis and I. Yamada, "Asymptotic minimization of sequences of loss functions constrained by families of quasi-nonexpansive mappings and its application to online learning," submitted for publication (preprint: `http://arxiv.org/abs/1008.5231`), Sept. 2010.

[8] J.-J. Moreau, "Fonctions convexes duales et points proximaux dans un espace Hilbertien," *Acad. Sci. Paris Sér. A Math.*, vol. 255, pp. 2897–2899, 1962.

[9] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer-Verlag, 2011.

[10] I. Yamada, M. Yukawa, and M. Yamagishi, "Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer-Verlag, 2011.

[11] P. J. Huber, *Robust Statistics*, Wiley, 1981.

[12] C. Michelot and M. L. Bougeard, "Duality results and proximal solutions of the Huber M-estimator problem," *Appl. Math. Optim.*, vol. 30, pp. 203–221, 1994.

[13] Y. Kopsinis, K. Slavakis, and S. Theodoridis, "Online sparse system identification and signal reconstruction using projections onto weighted ℓ_1 balls," *IEEE Trans. Signal Proc.*, vol. 59, no. 3, pp. 936–952, March 2011.

[14] H. Zou, "The adaptive LASSO and its oracle properties," *J. American Statistical Association*, vol. 101, pp. 1418–1429, December 2006.

[15] D. Angelosante, J. A. Bazerque, and G. B. Giannakis, "Online adaptive estimation of sparse signals: where RLS meets the ℓ_1-norm," *IEEE Trans. Signal Proc.*, vol. 58, no. 7, pp. 3436–3447, July 2010.

[16] S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections: a unifying framework for linear and nonlinear classification and regression tasks," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 97–123, Jan. 2011.

[17] Y. Murakami, M. Yamagishi, M. Yukawa, and I. Yamada, "A sparse adaptive filtering using time-varying soft-thresholding techniques," in *Proceedings of the IEEE ICASSP*, Dallas: USA, March 2010, pp. 3734–3737.

[18] D. Peel and G. J. McLachlan, "Robust mixture modelling using the t distribution," *Statistics and Computing*, vol. 10, pp. 339–348, 2000.

[19] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal. Statist. Soc. B.*, vol. 58, no. 1, pp. 267–288, 1996.