C-HiLasso: A Collaborative Hierarchical Sparse Modeling Framework

Pablo Sprechmann, *Student Member, IEEE*, Ignacio Ramírez, *Student Member, IEEE*, Guillermo Sapiro, *Senior Member, IEEE*, and Yonina C. Eldar, *Senior Member, IEEE*

***Abstract*—Sparse modeling is a powerful framework for data analysis and processing. Traditionally, encoding in this framework is performed by solving an ℓ_1-regularized linear regression problem, commonly referred to as *Lasso* or *Basis Pursuit*. In this work we combine the sparsity-inducing property of the Lasso at the individual feature level, with the block-sparsity property of the *Group Lasso*, where sparse groups of features are jointly encoded, obtaining a sparsity pattern hierarchically structured. This results in the *Hierarchical Lasso (HiLasso)*, which shows important practical advantages. We then extend this approach to the collaborative case, where a set of simultaneously coded signals share the same sparsity pattern at the higher (group) level, but not necessarily at the lower (inside the group) level, obtaining the collaborative HiLasso model *(C-HiLasso)*. Such signals then share the same active groups, or classes, but not necessarily the same active set. This model is very well suited for applications such as source identification and separation. An efficient optimization procedure, which guarantees convergence to the global optimum, is developed for these new models. The underlying presentation of the framework and optimization approach is complemented by experimental examples and theoretical results regarding recovery guarantees.**

***Index Terms*—Collaborative coding, hierarchical models, sparse models, structured sparsity.**

I. Introduction and Motivation

SPARSE signal modeling has been shown to lead to numerous state-of-the-art results in signal processing, in addition to being very attractive at the theoretical level. The standard model assumes that a signal can be efficiently represented by a sparse linear combination of atoms from a given or learned dictionary. The selected atoms form what is usually referred to as the *active set*, whose cardinality is significantly smaller than the size of the dictionary and the dimension of the signal.

In recent years, it has been shown that adding structural constraints to this active set has value both at the level of representation robustness and at the level of signal interpretation (in particular when the active set indicates some physical properties of the signal); see [1]–[4] and references therein. This leads to *group* or *structured* sparse coding, where instead of considering the atoms as singletons, the atoms are grouped, and a few groups are active at a time. An alternative way to add structure (and robustness) to the problem is to consider the simultaneous encoding of multiple signals, requesting that they all share the same active set. This is a natural collaborative filtering approach to sparse coding; see, for example, [5]–[10].

In this work, we extend these approaches in a number of directions. First, we present a hierarchical sparse model, where not only a few (sparse) groups of atoms are active at a time, but also each group enjoys internal sparsity.[1] At the conceptual level, this means that the signal is represented by a few groups (classes), and inside each group only a few members are active at a time. A simple example of this is a piece of music (numerous applications in genomics and image processing exist as well), where only a few instruments are active at a time (each instrument is a group), and the sound produced by each instrument at each instant is efficiently represented by a few atoms of the subdictionary/group corresponding to it. Thereby, this proposed hierarchical sparse coding framework permits to efficiently perform source identification and separation, where the individual sources (classes/groups) that generated the signal are identified at the same time as their representation is reconstructed (via the sparse code inside the group). An efficient optimization procedure, guaranteed to converge to the global optimum, is proposed to solve the hierarchical sparse coding problems that arise in our framework. Theoretical recovery bounds are derived, which guarantee that the output of the optimization algorithm is the true underlying signal.

Next, we go one step beyond. Continuing with the above example, if we know that the same few instruments will be playing simultaneously during different passages of the piece, then we can assume that the active groups at each instant, within the same passage, will be the same. We can exploit this information by applying the new hierarchical sparse coding approach in a collaborative way, enforcing that the same groups will be active at all instants within a passage (since they are of the same instruments and then efficiently representable by the same subdictionaries), while allowing each group for each music instant to have its own unique internal sparsity pattern (depending on how the sound of each instrument is represented at each instant). We propose a collaborative hierarchical sparse coding framework following this approach, *(C-HiLasso)*, along with an efficient optimization procedure. We then comment on results regarding the correct recovery of the underlying active groups.

[1]While we consider only two levels of sparsity, the proposed framework is easily extended to multiple levels.

Manuscript received June 01, 2010; revised November 20, 2010 and February 24, 2011; accepted May 12, 2011. Date of publication May 27, 2011; date of current version August 10, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Olgica Milenkovic. P. Sprechmann and I. Ramírez contributed equally to this work. This work was supported in part by the NSF, NSSEFF, ONR, NGA, and ARO.

P. Sprechmann, I. Ramírez, and G. Sapiro are with the Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: sprec009@umn.edu).

Y. C. Eldar is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2011.2157912

The proposed optimization techniques for both *HiLasso* and *C-HiLasso* is based on the Proximal Method [11], more specifically, on its particular implementation for sparse problems, *Sparse Reconstruction by Separable Approximation* (SpaRSA) [12]. This is an iterative method which solves a subproblem at each iteration which, in our case, has a closed form and can be solved in linear time. Furthermore, this closed form solution combines a vector thresholding and a scalar thresholding, naturally yielding to the desired hierarchical sparsity patterns.

The rest of the paper is organized as follows. Section II provides an introduction to traditional sparse modeling and presents our proposed HiLasso and C-HiLasso models. We discuss their relationship with the recent works of [2] and [13]–[17]. In Section III we describe the optimization techniques applied to solve the resulting sparse coding problems and we discuss its relationship with other optimization methods recently proposed in the literature [16], [18]. Theoretical recovery guarantees for HiLasso in the noiseless setting are developed in Section IV, demonstrating improved performance when compared with Lasso and Group Lasso. We also comment on existing results regarding correct recovery of group-sparse patterns in the collaborative case. Experimental results and simulations are given in Section V, and finally concluding remarks are presented in Section VI.

II. Collaborative Hierarchical Sparse Coding

A. Background: Lasso and Group Lasso

Assume we have a set of data samples $\mathbf{x}_j \in \mathbb{R}^m, j = 1,\dots,n$, and a dictionary of p atoms in $\mathbb{R}^m$, assembled as a matrix $\mathbf{D} \in \mathbb{R}^{m\times p}, \mathbf{D} = [\mathbf{d}_1\mathbf{d}_2\dots\mathbf{d}_p]$. Each sample $\mathbf{x}_j$ can be written as $\mathbf{x}_j = \mathbf{D}\mathbf{a}_j + \epsilon, \mathbf{a}_j \in \mathbb{R}^p, \epsilon \in \mathbb{R}^m$, that is, as a linear combination of the atoms in the dictionary $\mathbf{D}$ plus some perturbation ϵ, satisfying $\|\epsilon\|_2 \ll \|\mathbf{x}_j\|_2$. The basic underlying assumption in sparse modeling is that, for all or most j, the "optimal" $\mathbf{a}_j$ has only a few nonzero elements. Formally, if we define the ℓ_0 cost as the pseudo-norm counting the number of nonzero elements of $\mathbf{a}_j, \|\mathbf{a}_j\|_0 := |\{k : a_{kj} \neq 0\}|$, then we expect that $\|\mathbf{a}_j\|_0 \ll p$ and $\|\mathbf{a}_j\|_0 \ll m$ for all or most j. Seeking the sparsest representation $\mathbf{a}$ is known to be NP-hard. To determine $\mathbf{a}_j$ in practice, a multitude of efficient algorithms have been proposed, which achieve high correct recovery rates. The ℓ_1-minimization method is the most extensively studied recovery technique. In this approach, the nonconvex ℓ_0 norm is replaced by the convex ℓ_1 norm, leading to

$$\min_{\mathbf{a}\in\mathbb{R}^p} \|\mathbf{a}\|_1 \quad \text{s.t.} \quad \|\mathbf{x}_j - \mathbf{D}\mathbf{a}\|_2^2 \leq \epsilon. \tag{II.1}$$

The use of general purpose or specialized convex optimization techniques allows for efficient reconstruction using this strategy. The above approximation is known as the Lasso [19] or Basis Pursuit [20], [21]. A popular variant is to use the unconstrained version

$$\min_{\mathbf{a}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}_j - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1 \tag{II.2}$$

where λ is an appropriate parameter value, usually found by cross-validation, or based on statistical principles [22].

The fact that the $\|\cdot\|_1$ regularizer induces sparsity in the solution $\mathbf{a}_j$ is desirable not only from a regularization point of view, but also from a model selection perspective, where one wants to identify the relevant factors (atoms) that conform each sample $\mathbf{x}_j$. In many situations, however, the goal is to represent the relevant factors not as singletons but as groups of atoms. For a dictionary of p atoms, we define groups of atoms through their indices, $G \subseteq \{1,\dots,p\}$. Given a group G of indexes, we denote the subdictionary of the columns indexed by them as $\mathbf{D}_{[G]}$, and the corresponding set of reconstruction coefficients as $\mathbf{a}_{[G]}$. Define $\mathcal{G} = \{G_1,\dots,G_q\}$ to be a partition of $\{1,\dots,p\}$.[2] In order to perform model selection at the group level (relative to the partition $\mathcal{G}$), the Group Lasso problem was introduced in [1],

$$\min_{\mathbf{a}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}_j - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\psi_{\mathcal{G}}(\mathbf{a}), \tag{II.3}$$

where $\psi_{\mathcal{G}}$ is the Group Lasso regularizer defined in terms of $\mathcal{G}$ as $\psi_{\mathcal{G}}(\mathbf{a}) := \sum_{G\in\mathcal{G}} \left\|\mathbf{a}_{[G]}\right\|_2$. The function $\psi_{\mathcal{G}}$ can be seen as a generalization of the ℓ_1 regularizer, as the latter arises from the special case $\mathcal{G} = \{\{1\},\{2\},\dots,\{p\}\}$ (the groups are singletons), and as such, its effect on the groups of $\mathbf{a}$ is also a natural generalization of the one obtained with the Lasso: It "turns on/off" atoms in groups.

We can always consider the "noiseless" sparse coding problem $\min_{\mathbf{a}\in\mathbb{R}^p}\{\psi(\mathbf{a}) : \mathbf{x}_j = \mathbf{D}\mathbf{a}\}$, for a generic regularizer $\psi(\cdot)$, as the limit of the Lagrangian sparse coding problem $\min_{\mathbf{a}\in\mathbb{R}^p}\left\{\frac{1}{2}\|\mathbf{x}_j - \mathbf{D}\mathbf{a}\|_2^2 + \lambda\psi(\mathbf{a})\right\}$ when $\lambda \to 0$. In the remainder of this section, as well as in Section III, we only present the corresponding Lagrangian formulations.

B. The Hierarchical Lasso

The Group Lasso trades sparsity at the single-coefficient level with sparsity at a group level, while, inside each group, the solution is generally dense. Let us consider for example that each group is a subdictionary trained to efficiently represent, via sparse modeling, an instrument, a type of image, or a given class of signals in general. The entire dictionary $\mathbf{D}$ is then appropriate to represent all classes of the signal as well as mixtures of them, and Group Lasso will properly represent (dense) mixtures with one group or subdictionary per class. At the same time, since each class is properly represented in a sparse mode via its corresponding group or subdictionary, we expect sparsity inside its groups as well (which is not achieved by Group Lasso, whose solutions are dense inside each group). This will become even more critical in the collaborative case, where signals will share groups because they are of the same class, but will not necessarily share the full active sets, since they are not the same signal. To achieve the desired in-group sparsity, we simply reintroduce the ℓ_1 regularizer together with the group regularizer, leading to the proposed *Hierarchical Lasso (HiLasso)* model,[3]

$$\min_{\mathbf{a}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}_j - \mathbf{D}\|_2^2 + \lambda_2\psi_{\mathcal{G}}(\mathbf{a}) + \lambda_1\|\mathbf{a}\|_1. \tag{II.4}$$

[2]While in this paper we concentrate and develop the important nonoverlapping case, it will be clear that the concepts of collaborative hierarchical sparse modeling introduced here apply to the case of overlapping groups as well.

[3]We can similarly define a hierarchical sparsity model with ℓ_0 instead of ℓ_1.

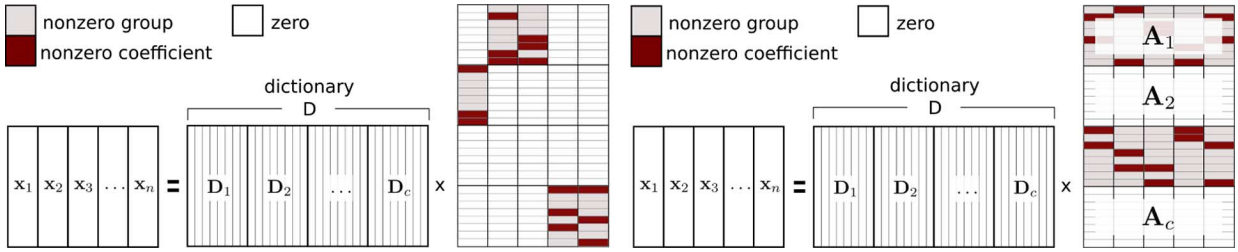

Fig. 1. Sparsity patterns induced by HiLasso (left) and C-HiLasso (right) model selection programs. Notice that the C-HiLasso imposes the same group-sparsity pattern in all the samples (same class), whereas the in-group sparsity patterns can vary between samples (samples themselves are different).

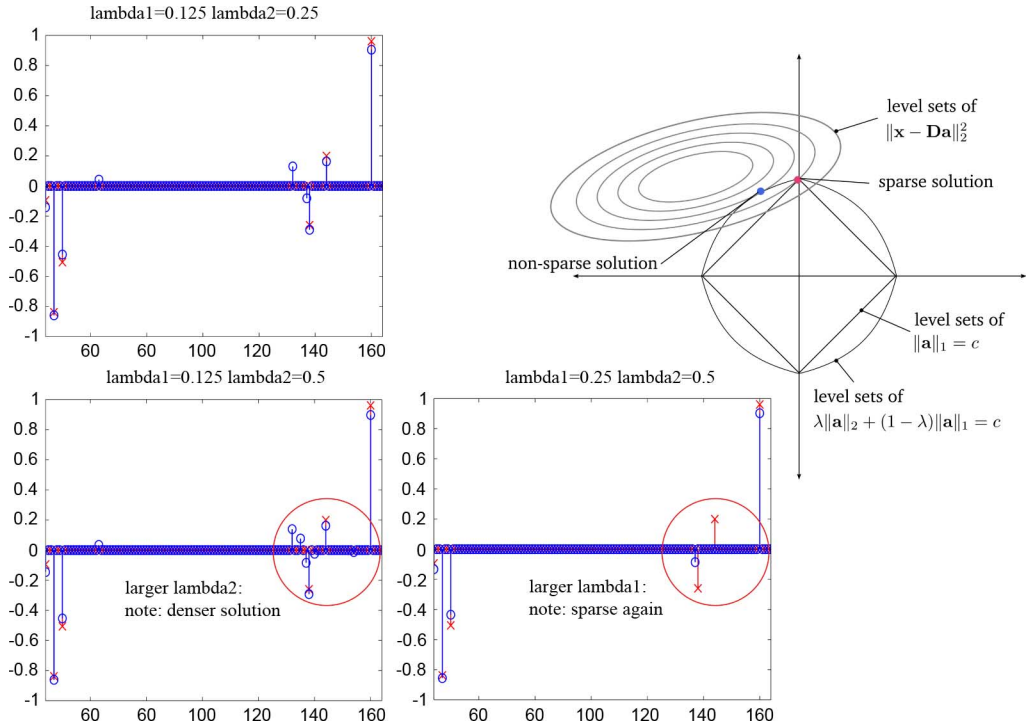

Fig. 2. Effect of different combinations of λ_1 and λ_2 on the solutions of the HiLasso coding problem. Three cases are given in which we want to recover a sparse signal (red crosses) $\mathbf{a}_0$ by means of the solution $\mathbf{a}$ of the HiLasso problem (blue dots). In this example we have two active groups out of ten possible (the sub dictionaries associated to each group have 30 atoms) and $\mathbf{a}_0 = 8$ (four nonzero coefficient per active group). The estimate that is closest to $\mathbf{a}_0$ in ℓ_1 norm is shown in the top left. As the ratio $\frac{\lambda_2}{\lambda_1}$ increases (bottom left), the level sets of the regularizer $\psi_{\mathcal{G}}(\cdot)$ become rounder, thus encouraging denser solutions. This is depicted in the rightmost figure for a simple case of $q = 1$ groups. Increasing λ_1 again (bottom right) increases sparsity, although here the final effect is too strong and some nonzero coefficients are not detected.

The hierarchical sparsity pattern produced by the solutions of (II.4) is depicted in Fig. 1 (left). For simplicity of the description, we assume that all the groups have the same number of elements. The extension to the general case is obtained by multiplying each group norm by the square root of the corresponding group size. This model then achieves the desired effect of promoting sparsity at the group/class level while at the same time leading to overall sparse feature selection. As mentioned above, additional levels of hierarchy can be considered as well, e.g., with groups inside the blocks. This is relevant for example in audio analysis.

As with models such as Lasso and Group Lasso, the optimal parameters λ_1 and λ_2 are application and data dependent. In some specific cases, closed form solutions exist for such parameters. For example, for signal restoration in the presence of noise using Lasso ($\lambda_2 = 0$), the GSURE method provides a simple way to compute the optimal λ_1 [22]. As extending such methods to HiLasso (or the C-HiLasso model presented next) is beyond the scope of this work, we rely on cross-validation for the choice of such parameters. The selection of λ_1 and λ_2 has an important influence on the sparsity of the obtained solution. Intuitively, as $\frac{\lambda_2}{\lambda_1}$ increases, the group constraint becomes dominant and the solution tends to be more sparse at a group level but less sparse within groups (see Fig. 2). This relation allows in practice to intuitively select a set of parameters that performs well. We also noticed empirically that the selection of the parameters is quite robust, since small variations in their numerical value don't change considerably the obtained results.

Some recent modeling frameworks for sparse coding do not rely on the selection of such model parameters, e.g., following the minimum description length criterion in [23], or nonparametric Bayesian techniques in [24]. Applying such techniques to the here proposed models is subject of future research.

C. Collaborative Hierarchical Lasso

In numerous applications, one expects that certain collections of samples $\mathbf{x}_j$ share the same active components from the dictionary, that is, that the indexes of the nonzero coefficients in $\mathbf{a}_j$ are the same for all the samples in the collection. Imposing such dependency in the ℓ_1 regularized regression problem gives rise to the so called collaborative (also called "multitask" or "simultaneous") sparse coding problem [5], [9], [10], [25]. Considering the coefficients matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{p \times n}$ associated with the reconstruction of the samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$, this model is given by

$$\min_{\mathbf{A} \in \mathbb{R}^{p \times n}} \frac{1}{2} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda \sum_{k=1}^{p} \left\|\mathbf{a}^k\right\|_2 \tag{II.5}$$

where $\mathbf{a}^k \in \mathbb{R}^n$ is the kth row of $\mathbf{A}$, that is, the vector of the n different values that the coefficient associated to the kth atom takes for each sample $j = 1, \ldots, n$. If we now extend this idea to the Group Lasso, we obtain a *collaborative Group Lasso (C-GLasso)* formulation

$$\min_{\mathbf{A} \in \mathbb{R}^{p \times n}} \frac{1}{2} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda \psi_{\mathcal{G}}(\mathbf{A}), \tag{II.6}$$

where $\psi_{\mathcal{G}}(\mathbf{A}) = \sum_{G \in \mathcal{G}} \left\|\mathbf{A}^G\right\|_F$, and $\mathbf{A}^G$ is the submatrix formed by all the rows belonging to group G. This regularizer is the natural collaborative extension of the regularizer in (II.3).

In this paper, we take an additional step and treat this together with the hierarchical extension presented in the previous section. The combined model that we propose, *C-HiLasso*, is given by

$$\min_{\mathbf{A} \in \mathbb{R}^{p \times n}} \frac{1}{2} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda_2 \psi_{\mathcal{G}}(\mathbf{A}) + \lambda_1 \sum_{j=1}^{n} \|\mathbf{a}_j\|_1. \tag{II.7}$$

The sparsity pattern obtained using (II.7) is shown in Fig. 1 (right). The C-GLasso is a particular case of our model when $\lambda_1 = 0$. On the other hand, one can obtain independent Lasso solutions for each $\mathbf{x}_i$ by setting $\lambda_2 = 0$. We see that (II.7) encourages all the signals to share the same groups (classes), while the active set inside each group is signal dependent. We thereby obtain a collaborative hierarchical sparse model, with collaboration at the class level (all signals collaborate to identify the classes), and freedom at the individual levels inside the class to adapt to each particular signal. This new model is particularly well suited, for example, when the data vectors have missing components. In this case combining the information from all the samples is very important in order to obtain a correct representation and model (group) selection. This can be done by slightly changing the data term in (II.7). For each data vector $\mathbf{x}_j$ one computes the reconstruction error using only the observed elements. Note that the missing components do not affect the other terms of the equation. Examples will be shown in Section V.

D. Relationship to Recent Literature

A number of recent works have addressed hierarchy, grouping and collaboration within the sparse modeling community. We now discuss the ones most closely related to the proposed Hi-Lasso and C-HiLasso models.

In [2], the authors propose a general framework in which one can define a regularization term to encourage a variety of sparsity patterns, and provide theoretical results (different to the ones developed here) for the single-signal case. The Hi-Lasso model presented here, in the single signal scenario, can be seen as a particular case of that model (where the groups in [2] should be blocks and singletons), although the particularly and important case of hierarchical structure introduced here is not mentioned in that paper. In [13] the authors simultaneously (see [26]) proposed a model that coincides with ours again in the single-signal scenario. None of these approaches develop the collaborative framework introduced here, nor the theoretical guarantees. The recovery of mixed signals with ℓ_0 optimization was addressed in [17]. This model does not include block sparsity (no hierarchy), collaboration, or the theoretical results we obtain here.

The special case of C-HiLasso when $\lambda_1 = 0$, C-GLasso, is investigated in [27], where a theoretical analysis of the signal recovery properties of the model is developed. Collaborative coding with structured sparsity has also been used recently in the context of gene expression analysis [14], [15]. In [14], the authors propose a model, that can be interpreted as a particular case of the collaborative approach presented here, in which a set of signals is simultaneously coded using a small (sparse) number of atoms of the dictionary. They modify the classical collaborative sparse coding regularization so that each signal can use any subset of the detected atoms. This is equivalent to our model when the groups have only one element and therefore there is no hierarchy in the coding. A collaborative model is presented in [15], where signals sharing the same active atoms are grouped together in a hierarchical way by means of a tree structure. The regularization term proposed is analogous to the one proposed in our work, but it is used to group signals rather than atoms (features), having once again no hierarchical coding.

Tree-based sparse coding has also been used recently to learn dictionaries [16], [18]. Under this model, if a particular learned atom is not used in the decomposition of a signal, then none of its descendants (in terms of the given tree structure) can be used. Although not explicitly considered in these works, the HiLasso model is an important particular case, among the wide spectrum of hierarchical sparse models considered in this line of work, where the hierarchy has two levels and no single atoms are in the upper level.

To conclude, while particular instances of the proposed C-HiLasso have been recently reported in the literature, none of them are as comprehensive. C-HiLasso includes both collaboration, at a block/group level, and hierarchical coding. Such collaborative hierarchical structure is novel and fundamental to address new important problems such as collaborative source identification and separation. The new theoretical results presented here extend the block sparsity results of [3], [28], complementing the modeling and algorithmic work.

III. Optimization

A. Single-Signal Problem: Hilasso

In the last decade, optimization of problems of the form of (II.2) and (II.3) have been deeply studied, and there exist very

efficient algorithms for solving them. Recently, Wright *et al.* [12] proposed a framework, spaRSA, for solving the general problem

$$\min_{\mathbf{a}\in\mathbb{R}^p} f(\mathbf{a}) + \lambda\psi(\mathbf{a}) \tag{III.8}$$

be a smooth and convex function, while $\psi : \mathbb{R}^p \rightarrow \mathbb{R}$ only needs to be finite and convex in $\mathbb{R}^p$. This formulation, which is a particular case of the Proximal Method framework developed by Nesterov [11], includes as important particular cases the Lasso, Group , Lasso and HiLasso problems by setting $f(\cdot)$ as the reconstruction error and then choosing the corresponding regularizers for $\psi(\cdot)$. When the regularizer, $\psi(\cdot)$, is group separable, the optimization can be subdivided into smaller problems, one per group. The framework becomes powerful when these subproblems can be solved efficiently. This is the case of the Lasso and Group Lasso (with non overlapping groups) settings, and also of the HiLasso, as we will show later in this Section. In all cases, the solution of the subproblems are obtained in linear time.

The spaRSA algorithm generates a sequence of iterates $\{\mathbf{a}^{(t)}\}_{t\in\mathbb{N}}$ that, under certain conditions, converges to the solution of (III.8). At each iteration, $\mathbf{a}^{(t+1)}$ is obtained by solving

$$\min_{\mathbf{z}\in\mathbb{R}^p} (\mathbf{z}-\mathbf{a}^{(t)})^\top \nabla f(\mathbf{a}^{(t)}) + \frac{\alpha^{(t)}}{2}\left\|\mathbf{z}-\mathbf{a}^{(t)}\right\|_2^2 + \lambda\psi(\mathbf{z}) \tag{III.9}$$

for a sequence of parameters $\{\alpha^{(t)}\}_{t\in\mathbb{N}}$, $\alpha^{(t)} = \alpha_0\eta^t$, where $\alpha_0 > 0$ and $\eta > 1$ need to be chosen properly for the algorithm to converge (see [12] for details). It is easy to show that (III.9) is equivalent to

$$\min_{\mathbf{z}\in\mathbb{R}^p} \frac{1}{2}\left\|\mathbf{z}-\mathbf{u}^{(t)}\right\|_2^2 + \frac{\lambda}{\alpha^{(t)}}\psi(\mathbf{z}) \tag{III.10}$$

where $\mathbf{u}^{(t)} = \mathbf{a}^{(t)} - \frac{1}{\alpha^{(t)}}\nabla f(\mathbf{a}^{(t)})$. In this new formulation, it is clear that the first term in the cost function can be separated elementwise. Thus, when the regularization function $\psi(\mathbf{z})$ is group separable, so is the overall optimization, and one can solve (III.10) independently for each group, leading to

$$\mathbf{a}_{[G]}^{(t+1)} = \arg\min_{\mathbf{z}\in\mathbb{R}^{|G|}} \frac{1}{2}\left\|\mathbf{z}-\mathbf{u}_{[G]}^{(t)}\right\|_2^2 + \frac{\lambda}{\alpha^{(t)}}\psi_{\mathcal{G}}(\mathbf{z}),$$

$\mathbf{z}_{[G]}$ being the corresponding variable for the group. In the case of HiLasso, this becomes

$$\mathbf{a}_{[G]}^{(t+1)} = \arg\min_{\mathbf{z}\in\mathbb{R}^{|G|}} \frac{1}{2}\|\mathbf{z}-\mathbf{w}\|_2^2 + \frac{\lambda_2}{\alpha^{(t)}}\|\mathbf{z}\|_2 + \frac{\lambda_1}{\alpha^{(t)}}\|\mathbf{z}\|_1 \tag{III.11}$$

where we have defined $\mathbf{w} = \mathbf{u}_{[G]}^{(t)}$. Problem (III.11) is a second order cone program (SOCP), for which one could use generic solvers. However, since it needs to be solved many times within the spaRSA iterations, it is crucial to solve it efficiently. It turns out that (III.11) admits a closed form solution with cost linear in the dimension of $\mathbf{w}$. By inspecting the subgradient of (III.11) for the case where the optimum $\mathbf{z}^* \neq 0$,

$$\mathbf{w} - \left(1 + \frac{\tilde{\lambda}_2}{\|\mathbf{z}^*\|_2}\right)\mathbf{z}^* \in \tilde{\lambda}_1\partial\|\mathbf{z}^*\|_1,$$

where we have defined $\tilde{\lambda}_2 = \frac{\lambda_2}{\alpha^{(t)}}$ and $\tilde{\lambda}_1 = \frac{\lambda_1}{\alpha^{(t)}}$. If we now define $C(\mathbf{z}^*) = 1 + \frac{\tilde{\lambda}_2}{\|\mathbf{z}^*\|_2}$, we observe that each element of $C(\mathbf{z}^*)\mathbf{z}^*$ is the solution of the well known scalar soft thresholding operator

$$\begin{aligned} z_i^* &= \frac{1}{C(\mathbf{z}^*)}\mathrm{sgn}(w_i)\max\{0, |w_i| - \tilde{\lambda}_1\} \\ &= \frac{h_i}{C(\mathbf{z}^*)}, \quad i = 1, \ldots, g \end{aligned} \tag{III.12}$$

where we have defined $h_i = \mathrm{sgn}(w_i)\max\{0, |w_i| - \tilde{\lambda}_1\}$, the result of the scalar thresholding of $\mathbf{w}$. Taking squares on both sides of (III.12) and summing over $i = 1, \ldots, g$ we obtain

$$\|\mathbf{z}^*\|_2^2 = C^2(\mathbf{z}^*)\|\mathbf{h}\|_2^2 = \frac{\|\mathbf{z}^*\|_2^2}{(\|\mathbf{z}^*\|_2 + \tilde{\lambda}_2)^2}\|\mathbf{h}\|_2^2,$$

from which the equality $\|\mathbf{z}^*\|_2 = \|\mathbf{h}^*\|_2 - \tilde{\lambda}_2$ follows. Since all terms are positive, this can only hold as long as $\|\mathbf{h}^*\|_2 > \tilde{\lambda}_2$, which gives us a vector thresholding condition on the solution $\mathbf{z}^*$ in terms of $\|\mathbf{h}\|_2$. It is easy to show that $\|\mathbf{h}^*\|_2 \leq \tilde{\lambda}_2$ is a sufficient condition for $\mathbf{z}^* = 0$. Thus, we obtain

$$\mathbf{a}_{[G]}^{(t+1)} = \begin{cases} \frac{\max\{0, \|\mathbf{h}\|_2 - \tilde{\lambda}_2\}}{\|\mathbf{h}\|_2}\mathbf{h}, & \|\mathbf{h}\|_2 > 0 \\ \mathbf{0}, & \|\mathbf{h}\|_2 = 0. \end{cases} \tag{III.13}$$

The above expression requires g scalar thresholding operations, and one vector thresholding, which is also linear with respect to the group size g. Therefore, for all groups, the cost of solving the subproblem (III.11) is linear in m, the same as for Lasso and Group Lasso. The complete HiLasso optimization algorithm is summarized in Algorithm 1. The parameter η has very little influence in the overall performance (see [12] for details); we used $\eta = 2$ in all our experiments. Note that, as expected, the solution to the subproblem for the cases $\lambda_2 = 0$ or $\lambda_1 = 0$, corresponds respectively to scalar soft thresholding and vector soft thresholding. In particular, when $\lambda_2 = 0$, the proposed optimization reduces to the Iterative Soft Thresholding algorithm [29].

Algorithm 1: HiLasso Optimization Algorithm

Input Data $\mathbf{X}$, dictionary $\mathbf{D}$, group set $\mathcal{G}$, constants $\alpha_0 > 0$, $\eta > 1$, $c > 0$, $0 < \alpha_{\min} < \alpha_{\max}$

Output The optimal point $\mathbf{a}^*$

Initialize $t := 0, \mathbf{a}^{(0)} := \mathbf{0}$

while *stopping criterion is not satisfied* **do**

  choose $\alpha t \in [\alpha_{\min}, \alpha_{\max}]$

  set $\mathbf{u}^{(t)} := \mathbf{a}^{(t)} - \frac{1}{\alpha^{(t)}}\nabla f(\mathbf{a}^{(t)})$;

  while *stopping criterion is not satisfied* **do**

    //Here we use the group separability of (III.10) and solve (III.11) for each group

    for $i := 1$ *to* q **do**

      Compute $\mathbf{a}_{[G]}^{(t+1)}$ as the solution to (III.13);

    end

    set $\alpha^{(t+1)} := \eta\alpha^{(t)}$;

  end

  set $t := t + 1$;

end

B. Optimization of the Collaborative Hilasso

The multisignal (collaborative) case is equivalent to the one-dimensional case where the signal is a concatenation of the columns of $\mathbf{X}$, and the dictionary is an $nm \times np$ block-diagonal matrix, where each of the n blocks is a copy of the original dictionary $\mathbf{D}$. However, in practice, it is not needed to build such (possibly very large) dictionary, and we can operate directly with the matrices $\mathbf{D}$ and $\mathbf{X}$ to find $\mathbf{A}$. If we define the matrix $\mathbf{U}^{(t)} \in \mathbb{R}^{m\times n}$ whose ith column is given by $\mathbf{u}_i^{(t)} = \mathbf{a}_i^{(t)} - \frac{1}{\alpha^{(t)}}\nabla f(\mathbf{a}_i^{(t)})$, we get the following SpaRSA iterates:

$$\mathbf{A}^{(t+1)} = \arg\min_{\mathbf{Z}\in\mathbb{R}^{m\times n}} \frac{1}{2}\left\|\mathbf{Z}-\mathbf{U}^{(t)}\right\|_F^2 + \frac{\lambda_2}{\alpha^{(t)}}\|\mathbf{Z}\|_F + \frac{\lambda_1}{\alpha^{(t)}}\sum_{j=1}^{n}\|\mathbf{z}_j\|_1$$

which again is group separable, so that it can be solved as q independent problems in the corresponding bands of $\mathbf{U}^{(t)}$,

$$(\mathbf{A}^{(t+1)})^G = \arg\min_{\mathbf{Z}\in\mathbb{R}^{g\times n}} \frac{1}{2}\left\|\mathbf{Z}-(\mathbf{U}^{(t)})^G\right\|_F^2 + \frac{\lambda_2}{\alpha^{(t)}}\|\mathbf{Z}\|_F + \frac{\lambda_1}{\alpha^{(t)}}\sum_{j=1}^{n}\|\mathbf{z}_j\|_1.$$

The correspondent closed form solutions for these subproblems, which are obtained in an analogous way to (III.12)–(III.13), are given by

$$(\mathbf{A}^{(t+1)})^G = \begin{cases} \frac{\max\{0,\|\mathbf{H}\|_F-\tilde{\lambda}_2\}}{\|\mathbf{H}\|_F}\mathbf{H}, & \|\mathbf{H}\|_F > 0 \\ \mathbf{0}, & \|\mathbf{H}\|_F = 0 \end{cases},$$
$$h_{ij} = \operatorname{sgn}(w_{ij})\max\{0, |w_{ij}| - \tilde{\lambda}_1\} \qquad \text{(III.14)}$$

and we have defined $\mathbf{W} := (\mathbf{U}^{(t)})^G$. As mentioned in Section II-D, [18] addresses a wide spectrum of hierarchical sparse models for coding and dictionary learning. They propose a proximal method optimization procedure that, when restricted to the formulation of HiLasso, is very similar to the one developed in Section III-A. The main difference with our method is that they solve the subproblem (III.10) using a dual approach (based on conic duality) that finds the exact solution in a finite number of operations. Our method, being tailored to the specific case of HiLasso, provides such solution in closed form, requiring just two thresholdings, both linear in the dimension of $\mathbf{X}$, $n\times m$.

IV. Theoretical Guarantees

In our current theoretical analysis, we study the case of a single measurement vector (signal) $\mathbf{x}$ (we comment on the collaborative case at the end of this section), and assume that there is no measurement noise or perturbation, so that $\mathbf{x} = \mathbf{D}\mathbf{a}$. Without loss of generality, we further assume that the cardinality $|G_r| = g, r = 1,\ldots,q$, that is, all groups in $\mathcal{G}$ have the same size. The goal is to recover the code $\mathbf{a}$, from the observed $\mathbf{x}$, by solving the noise-free HiLasso problem:

$$\min_{\mathbf{a}\in\mathbb{R}^p}\{\lambda\psi_{\mathcal{G}}(\mathbf{a}) + (1-\lambda)\|\mathbf{a}\|_1 \quad \text{s.t.} \quad \mathbf{x} = \mathbf{D}\mathbf{a}\}. \qquad \text{(IV.15)}$$

Note that we have replaced the two regularization parameters λ_1 and λ_2 by a single parameter λ, since scaling does not effect the optimal solution. Therefore, we can always assume that $\lambda_1 + \lambda_2 = 1$.

Our goal is to develop conditions under which the HiLasso program of (IV.15) will recover the true unknown vector $\mathbf{a}$. As we will see, the resulting set of recoverable signals is a superset of those recoverable by Lasso, that is, HiLasso is able to recover signals for which Lasso (or Group Lasso) will fail to do so.

We assume throughout this section that $\mathbf{a}$ has group sparsity k, namely, no more than k of the group vectors $\mathbf{a}_{[G_i]}, i = 1,\ldots,q$, have nonzero norm. In addition, within each group, we assume that not more than s elements are non zero, that is, $\|\mathbf{a}_{[G]}\|_0 \le s$.

For $\lambda = 1$, (IV.15) reduces to the Group Lasso problem, (II.3), whereas with $\lambda = 0$, (IV.15) becomes equivalent to the Lasso problem, (II.2). Both cases have been treated previously in the literature and sufficient conditions have been derived on the sparsity levels and on the dictionary $\mathbf{D}$ to ensure that the resulting optimization problem recovers the true unknown vector $\mathbf{a}$. For example, in [3], [30], [31], conditions are given in terms of the restricted isometry property (RIP) of $\mathbf{D}$. In an alternative line of work, recovery conditions are based on coherence measures, which are easier to compute [28], [32]. Here, we follow the same spirit and consider coherence bounds that ensure recovery using the HiLasso approach. We also draw from [10] to briefly describe conditions under which the probability of error of recovering the correct groups, using the special case of the C-HiLasso with $\lambda_1 = 0$ (C-GLasso), falls exponentially to 0 as the number of collaborating samples n grows. Finally, recent theoretical results on block sparsity were reported in [33]. In particular, bounds on the number of measurements required for block sparse recovery were developed under the assumption that the measurement matrix $\mathbf{D}$ has a basis of the null-space distributed uniformly in the Grassmanian. The model is a block-sparse model, without hierarchical or collaborative components.

In this section, we extend the groupwise indexing notation to refer both to subsets of rows and columns of arbitrary matrices as $\mathbf{W}_{[F,G]} := \{w_{ij} : i \in F, j \in G\}$. This is, $\mathbf{W}_{[F,G]} = \mathbf{I}_{[F]}^T\mathbf{W}\mathbf{J}_{[G]}$, where $\mathbf{I}$ and $\mathbf{J}$ are the identity matrices of the column and row spaces of $\mathbf{W}$, respectively. We define the sets $\Omega = \{1,2,\ldots,p\}$ and $\Gamma = \{1,2,\ldots,g\}$, and use $\bar{S}$ to denote the complement of a set of indexes S, either with respect to Ω or Γ, depending on the context. The set difference between S and T is denoted as $S \setminus T$, $\emptyset$ represents the empty set, and $|S|$ denotes the cardinality of S.

A. Block-Sparse Coherence Measures

We begin by reviewing previously proposed coherence measures. For a given dictionary $\mathbf{D}$, the (standard) coherence is defined as $\mu := \max_{i,j\neq i\in\Gamma}|\mathbf{d}_i^{\mathrm{T}}\mathbf{d}_j|$. This coherence was extended to the block-sparse setting in [28], leading to the definition of *block coherence*:

$$\mu_B := \max\left\{\frac{1}{g}\rho(\mathbf{D}_{[G]}^{\mathrm{T}}\mathbf{D}_{[F]}), G, F \in \mathcal{G}, G \neq F\right\}$$

where $\rho(\cdot)$ is the spectral norm, that is, $\rho(\mathbf{Z}) := \lambda_{\max}^{\frac{1}{2}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})$, with $\lambda_{\max}(\mathbf{W})$ denoting the largest eigenvalue of the positive semi-definite matrix $\mathbf{W}$. An alternate atomwise measure of block coherence is given by the *cross coherence*

$$\chi := \max\left\{\max\left\{|\mathbf{d}_i^T\mathbf{d}_j|, i \in G, j \in F\right\} G, F \in \mathcal{G}, G \neq F\right\}. \tag{IV.16}$$

When $g = 1$ (each block is a singleton), $\mathbf{D}_{[G_r]} = \mathbf{d}_r$, so that as expected, $\chi = \mu_B = \mu$. While μ_B and χ quantify global properties of the dictionary $\mathbf{D}$, local block properties are characterized by the *subcoherence*, defined as

$$\nu := \max\left\{\max\left\{|\mathbf{d}_i^T\mathbf{d}_j|, i, j \in G, i \neq j\right\} G \in \mathcal{G}\right\}. \tag{IV.17}$$

We define $\nu = 0$ for $g = 1$. Clearly, if the columns of $\mathbf{D}_{[G]}$ are orthonormal for each group G, then $\nu = 0$. Assuming the columns of $\mathbf{D}$ have unit norm, it can be easily shown that μ, ν, χ and μ_B all lie in the range $[0, 1]$. In addition, we can easily prove that $\nu,\ \mu_B, \chi \leq \mu$. In our setting, $\mathbf{a}$ is block sparse, but has further internal structure: each subvector of $\mathbf{a}$ is also sparse. In order to quantify our ability to recover such signals, we expect that an appropriate coherence measure will be based on the definition of block sparsity, but will further incorporate the internal sparsity as well. Let $\mathbf{M} := \mathbf{D}^{\mathrm{T}}\mathbf{D}$ denote the Gram matrix of $\mathbf{D}$. Then, the standard block coherence μ_B is defined in terms of the largest singular value of an off-diagonal $g \times g$ subblock of $\mathbf{M}$. In a similar fashion, we will define *sparse block coherence* measures in terms of *sparse singular values*. As we will see, two different definitions will play a role, depending on where exactly the sparsity within the block enters. To define these, we note that the *spectral norm* $\rho(\mathbf{Z})$ of a matrix $\mathbf{Z}$ can be defined as

$$\rho(\mathbf{Z}) := \max_{\mathbf{u},\mathbf{v}} |\mathbf{u}^{\mathrm{T}}\mathbf{Z}\mathbf{v}| \quad \text{s.t.} \quad \|\mathbf{u}\|_2 = 1, \|\mathbf{v}\|_2 = 1.$$

Alternatively, we can define $\rho(\mathbf{Z})$ as the largest singular value of $\mathbf{Z}$, $\rho(\mathbf{Z}) := \sigma_{\max}(\mathbf{Z}) = \sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}$,

$$\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}) := \max_{\mathbf{v}} \mathbf{v}^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})\mathbf{v} \quad \text{s.t.} \quad \|\mathbf{v}\|_2 = 1.$$

We now develop sparse analogs of $\rho(\mathbf{Z})$ and $\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})$. As we will see, the simple square-root relation no longer holds in this case. The *largest sparse singular value* is defined as [34]

$$\rho^{ss}(\mathbf{Z}) := \max_{\mathbf{u},\mathbf{v}} |\mathbf{u}^{\mathrm{T}}\mathbf{Z}\mathbf{v}| \quad \text{s.t.} \quad \|\mathbf{u}\|_2 = 1,\ \|\mathbf{v}\|_2 = 1,\ \|\mathbf{u}\|_0 \leq s,\ \|\mathbf{v}\|_0 \leq s. \tag{IV.18}$$

Similarly, the *largest sparse eigenvalue* of $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ is defined as [34]–[36]

$$\lambda_{\max}^{s}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}) := \max_{\mathbf{v}} \mathbf{v}^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})\mathbf{v} \quad \text{s.t.} \quad \|\mathbf{v}\|_2 = 1,\ \|\mathbf{v}\|_0 \leq s. \tag{IV.19}$$

The *sparse matrix norm* is then given by

$$\rho^{s}(\mathbf{Z}) := \sqrt{\lambda_{\max}^{s}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}. \tag{IV.20}$$

Note that, in general, $\rho^s(\mathbf{Z})$ is not equal to $\rho^{ss}(\mathbf{Z})$. It is easy to see that $\rho^{ss}(\mathbf{Z}) \leq \rho^s(\mathbf{Z})$. For any matrix $\mathbf{Z}$, $\rho^{ss}(\mathbf{Z}) = \rho(\mathbf{Z}_{[F,G]})$ and $\rho^s(\mathbf{Z}) = \rho(\mathbf{Z}_{[T]})$, where F, G, T are subsets of $\Gamma = \{1, 2, \ldots, g\}$ of size s, chosen to maximize the corresponding singular value. Using (IV.18) and (IV.20), we define two sparse block coherence measures:

$$\mu_B{}^{ss} := \max\left\{\frac{1}{g}\rho^{ss}(\mathbf{D}_{[G]}^{\mathrm{T}}\mathbf{D}_{[F]}), G, F \in \mathcal{G}, G \neq F\right\} \tag{IV.21}$$

$$\mu_B{}^{s} := \max\left\{\frac{1}{g}\rho^{s}(\mathbf{D}_{[G]}^{\mathrm{T}}\mathbf{D}_{[F]}),\ G, F \in \mathcal{G}, G \neq F\right\}. \tag{IV.22}$$

The choice of scaling is to ensure that $\mu_B^s, \mu_B^{ss} \leq \mu_B$.

Note that, while $\rho^s(\mathbf{Z})$ (also referred to in the literature as *sparse principal component analysis* (SPCA)) and $\rho^{ss}(\mathbf{Z})$ are in general NP-hard to compute, in many cases they can be computed exactly, or approximated, using convex programming techniques [34]–[36].

The following proposition establishes some relations between these new definitions and the standard coherence measures.

Proposition 1: The sparse block-coherence measures μ_B^{ss}, μ_B^s satisfy

$$0 \leq \mu_B^{ss} \leq \frac{s}{g}\mu, \quad 0 \leq \mu_B^s \leq \sqrt{\frac{s}{g}}\mu. \tag{IV.23}$$

Proof: The inequalities $\mu_B^{ss}, \mu_B^s \geq 0$ follow immediately from the definition. We obtain the upper bounds by rewriting $\rho^{ss}(\mathbf{Z})$ and $\rho^s(\mathbf{Z})$ and then using the Geršgorin theorem,

$$\begin{aligned}\rho^{ss}(\mathbf{Z}) &= \lambda_{\max}^{\frac{1}{2}}(\mathbf{Z}_{[F,G]}^{\mathrm{T}}\mathbf{Z}_{[F,G]}) \overset{(a)}{\leq} \sqrt{\max_{l}\sum_{r=1}^{s}|e_{lr}|} \\ &\leq \sqrt{s\max_{l,r}|e_{lr}|} \end{aligned} \tag{IV.24}$$

$$\begin{aligned}\rho^{s}(\mathbf{Z}) &= \lambda_{\max}^{\frac{1}{2}}(\mathbf{Z}_{[T]}^{\mathrm{T}}\mathbf{Z}_{[T]}) \overset{(b)}{\leq} \sqrt{\max_{l}\sum_{r=1}^{s}|e'_{lr}|} \\ &\leq \sqrt{s\max_{l,r}|e'_{lr}|} \end{aligned} \tag{IV.25}$$

where e_{lr} and e'_{lr} are the elements of $\mathbf{E} = \mathbf{Z}_{[F,\Gamma]}^{\mathrm{T}}\mathbf{Z}_{[F,\Gamma]}$ and $\mathbf{E}' = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$, and (a), (b) are a consequence of Geršgorin's disc theorem.

The entries of $\mathbf{Z} = \mathbf{D}_{[G_i]}^{\mathrm{T}}\mathbf{D}_{[G_j]}$ for $i \neq j$ have absolute value smaller than or equal to μ, and the size of $\mathbf{Z}$ is $g \times g$. Therefore, $|e_{k\ell}| \leq s\mu^2$ and $|e'_{k\ell}| \leq g\mu^2$. Substituting these values into (IV.24) and (IV.25) concludes the proof of the upper bounds on μ_B^{ss} and μ_B^s. ∎

B. Recovery Proof

Our main recovery result is stated as follows. Suppose that $\mathbf{a}$ is a block k-sparse vector with blocks of length g, where each block has sparsity exactly s,[4] and let $\mathbf{x} = \mathbf{D}\mathbf{a}$. We rearrange the columns in $\mathbf{D}$ and the coefficients in $\mathbf{a}$ so that the first k groups, $\{G_1, G_2, \ldots, G_k\}$ correspond to the nonzero (active) blocks. Within each block $G_i, i \leq k$, the first s indices, represented by the set S_i, correspond to the s nonzero coefficients in that block, and the index set $T_i = G_i \setminus S_i$ represents its $(g - s)$ inactive elements, so that $G_i = [S_i\ T_i]$. The set $G_0 = \bigcup_{i=1}^{k} G_i$ contains the indexes of all the active blocks of $\mathbf{a}$, whereas $\bar{G}_0 = \Omega \setminus G$

[4]These conditions are nonlimiting, since we can always complete the vector with zeros.

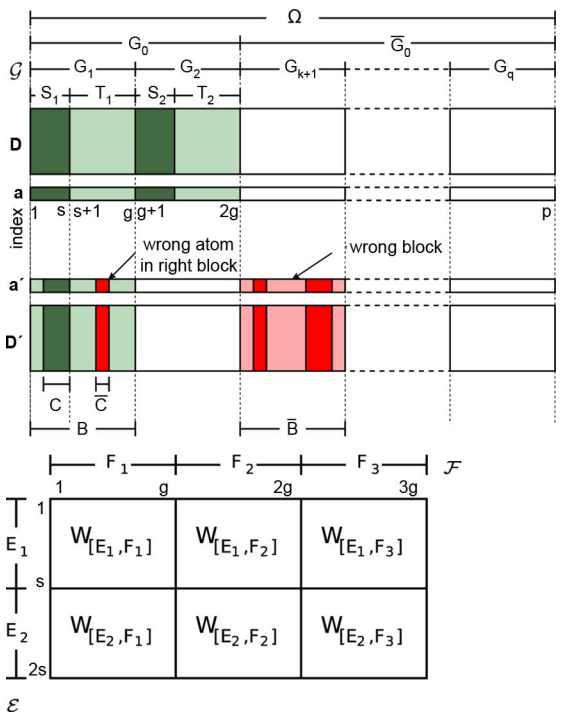

Fig. 3. Top: Indexing conventions, here shown for $g = 8$, $k = 2$, and $s = 3$. Shaded regions correspond to active elements/atoms. Active blocks are light-colored, and active elements/coefficients are dark colored. Here, $\mathbf{a}'$ represents an alternate representation of $\mathbf{x}$, $\mathbf{x} = \mathbf{D}\mathbf{a}'$. Blocks and atoms that are not part of the true solution $\mathbf{a}$ are marked in red. Bottom: Partitioning of a matrix $\mathbf{W}$ performed by the measure $\rho_{[\mathcal{E},\mathcal{F}]}(\mathbf{W})$ with $\mathcal{E} = \{E_1, E_2\}$ and $\mathcal{F} = \{F_1, F_2, F_3\}$, where $|E_i| = s$ and $|F_j| = g$.

contains the inactive ones. Similarly, $S_0 = \bigcup_{i=1}^{k} S_i$ contains the indexes of all the active coefficients/atoms in $\mathbf{a}$ and $\mathbf{D}$ respectively, $\bar{S}_0 = \Omega \setminus S_0$ indexes the inactive coefficients/atoms in $\frac{\mathbf{a}}{\mathbf{D}}$, and $T_0 = \bigcup_{i=1}^{k} T_i$ indexes the inactive coefficients/atoms within the active blocks. These indexing conventions are exemplified in Fig. 3 (top). With these conventions we can write $\mathbf{x} = \mathbf{D}_{[G_0]}\mathbf{a}_{[G_0]} = \mathbf{D}_{[S_0]}\mathbf{a}_{[S_0]}$.

An important assumption that we will rely on throughout, is that the columns of $\mathbf{D}_{[G_0]}$ must be linearly independent for any G_0 as defined above. Under this assumption, $\mathbf{D}_{[S_0]}^{\mathsf{T}}\mathbf{D}_{[S_0]}$ is invertible and we can define the pseudo-inverse $\mathbf{H} := (\mathbf{D}_{[S_0]}^{\mathsf{T}}\mathbf{D}_{[S_0]})^{-1}\mathbf{D}_{[S_0]}^{\mathsf{T}}$. For reasons that will become clear later, we will also need a second, oblique pseudo-inverse, $\mathbf{Q} := (\mathbf{D}_{[S_0]}^{\mathsf{T}}(\mathbf{I}-\mathbf{P})\mathbf{D}_{[S_0]})^{-1}\mathbf{D}_{[S_0]}^{\mathsf{T}}(\mathbf{I}-\mathbf{P})$, where $\mathbf{P}$ is an orthogonal projection onto the range of $\mathbf{D}_{[T_0]}$, that is, $\mathbf{P}\mathbf{D}_{[T_0]} = \mathbf{D}_{[T_0]}$. It is easy to check that

$$\mathbf{Q}\mathbf{D}_{[T_0]} = \mathbf{0} \quad \text{and} \quad \mathbf{Q}\mathbf{D}_{[S_0]} = \mathbf{I}. \tag{IV.26}$$

Equipped with these definitions we can now state our main result.

Theorem 1: Let $\mathbf{a}$ be a block k-sparse vector with blocks of length g, where each block has sparsity s. Let $\mathbf{x} = \mathbf{D}\mathbf{a}$ for a given matrix $\mathbf{D}$. A sufficient condition for the HiLasso algorithm (IV.15) to recover $\mathbf{a}$ from $\mathbf{x}$ is that, for some $\alpha \leq 1$,

$$\rho_{[S_0,\bar{G}_0]}(\mathbf{Q}\mathbf{D}_{[\bar{G}_0]}) < \alpha \tag{IV.27}$$

$$\|\mathbf{H}\mathbf{D}_{[\bar{G}_0]}\|_{1,1} < \gamma, \quad \gamma \leq 1 + \frac{\lambda(1-\alpha)}{\sqrt{g}(1-\lambda)} \tag{IV.28}$$

$$\|\mathbf{H}\mathbf{D}_{[T_0]}\|_{1,1} < 1. \tag{IV.29}$$

Here $\rho_{[\mathcal{E},\mathcal{F}]}(\mathbf{Z}) := \max_{F\in\mathcal{F}} \sum_{E\in\mathcal{E}} \rho(\mathbf{Z}_{[E,F]})$, is the block spectral norm defined in [28], the blocks defined by the sets of index sets $\mathcal{E}$ and $\mathcal{F}$ [see Fig. 3 (bottom)]. We also define $\mathcal{S}_0 = \{S_i : i = 1,\ldots,k\}$, $\bar{\mathcal{G}}_0 = \{G_i : i = k+1,\ldots,q\}$ and $\mathcal{T}_0 = \{T_i : i = 1,\ldots,k\}$. Finally, $\|\mathbf{Z}\|_{1,1} := \max_r \|\mathbf{z}_r\|_1$, where $\mathbf{z}_r$ is the rth column of $\mathbf{Z}$.

The above theorem can be interpreted as follows. With $\gamma = 1$, the conditions (IV.28)–(IV.29) are sufficient both for Lasso ($\lambda = 0$) and HiLasso to recover $\mathbf{a}$. However, if there exists a $\gamma > 1$ for which condition (IV.28) holds, then HiLasso will be able to recover $\mathbf{a}$ in a situation where Lasso is not guaranteed to do so. The idea is that, for $0 < \lambda < 1$, HiLasso trades off between the minimization of its ℓ_1 and ℓ_2 terms, by tightening the ℓ_2 term ($\alpha \leq 1$) to improve group recovery, while loosening the ℓ_1 term ($\gamma > 1$). Also, although not yet clear from conditions (IV.27)–(IV.29), we will see in Theorem 2 that the final data independent bounds are also a relaxation of the ones corresponding to Group Lasso when the solutions are block-dense. Therefore, the proposed model outperforms both standard Lasso and Group Lasso with regard to recovery guarantees. This is also reflected in the experimental results presented in the next section.

The sufficient conditions (IV.27)–(IV.29) depend on $\mathbf{D}_{[S_0]}$ and therefore on the nonzero blocks in $\mathbf{a}$, G_0, and the nonzero locations within the blocks, S_0, which, of course, are not known in advance. Nonetheless, Theorem 2 provides sufficient conditions ensuring that (IV.27)–(IV.29) hold, which are independent of the unknown signals, and depend only on the dictionary $\mathbf{D}$.

We now prove Theorem 1.

Proof: To prove that (IV.15) recovers the correct vector $\mathbf{a}$, let $\mathbf{a}'$ be an alternative solution satisfying $\mathbf{x} = \mathbf{D}\mathbf{a}'$. We will show that $\lambda\psi_{\mathcal{G}}(\mathbf{a}) + (1-\lambda)\|\mathbf{a}\|_1 < \lambda\psi_{\mathcal{G}}(\mathbf{a}') + (1-\lambda)\|\mathbf{a}'\|_1$. Let the set G_0 contain the indexes of all elements in the active blocks of $\mathbf{a}$. Let G_0' contain the indexes of the active blocks in $\mathbf{a}'$. Then $\mathbf{x} = \mathbf{D}_{[G_0]}\mathbf{a}_{[G_0]} = \mathbf{D}_{[G_0']}\mathbf{a}'_{[G_0']}$.

By our assumptions, in each block of $\mathbf{a}_{[G_0]}$ there are exactly s nonzero values. Let the set $S_0 \subset G_0$ contain the indexes of all nonzero elements in $\mathbf{a}$. We thus have $|S_0| = ks$. Using (IV.26) we can write

$$\mathbf{a}_{[S_0]} = \mathbf{Q}\mathbf{D}_{[S_0]}\mathbf{a}_{[S_0]} = \mathbf{Q}\mathbf{D}_{[G_0]}\mathbf{a}_{[G_0]} = \mathbf{Q}\mathbf{D}_{[G_0']}\mathbf{a}'_{[G_0']}. \tag{IV.30}$$

To proceed, we separate G_0' into two parts: $B = G_0' \cap G_0$, and $\bar{B} = G_0' \setminus G_0$, so that $G_0' = [B\bar{B}]$ and $\mathbf{D}_{[G_0']}\mathbf{a}'_{[G_0']} = \mathbf{D}_{[B]}\mathbf{a}'_{[B]} + \mathbf{D}_{[\bar{B}]}\mathbf{a}'_{[\bar{B}]}$. We can now rewrite (IV.30) as

$$\mathbf{a}_{[S_0]} = \mathbf{Q}\mathbf{D}_{[B]}\mathbf{a}'_{[B]} + \mathbf{Q}\mathbf{D}_{[\bar{B}]}\mathbf{a}'_{[\bar{B}]} \tag{IV.31}$$

and use the triangle inequality to obtain

$$\psi_{\mathcal{G}}(\mathbf{a}_{[S_0]}) \leq \psi_{\mathcal{G}}(\mathbf{Q}\mathbf{D}_{[B]}\mathbf{a}'_{[B]}) + \psi_{\mathcal{G}}(\mathbf{Q}\mathbf{D}_{[\bar{B}]}\mathbf{a}'_{[\bar{B}]}). \tag{IV.32}$$

We now analyze the two terms in the right-hand side of (IV.32) using [28, Lemma3]:

Lemma 1: Let $\mathbf{v} \in \mathbb{R}^p$ be a vector, $\mathbf{Z} \in \mathbb{R}^{m\times p}$ be a matrix, $\mathcal{F}$ be a partition of $\Omega = \{1,2,\ldots,p\}$, and $\mathcal{E}$ a partition of $\{1,\ldots,m\}$. We then have that, $\psi_{\mathcal{G}}(\mathbf{Zv}) \le \rho_{[\mathcal{E},\mathcal{F}]}(\mathbf{Z})\psi_{\mathcal{G}}(\mathbf{v})$.[5]

Since $\bar{B} \subset \bar{G}_0$, it follows from (IV.27) that $\rho_{[\mathcal{S}_0,\bar{\mathcal{B}}]}(\mathbf{QD}_{[\bar{B}]}) < \alpha$ (here $\bar{\mathcal{B}}$ is the set of the blocks that comprise $\bar{B}$). To analyze $\rho_{[\mathcal{S}_0,\mathcal{B}]}(\mathbf{QD}_{[B]})$, we use its definition

$$\begin{aligned}\rho_{[\mathcal{S}_0,\mathcal{B}]}(\mathbf{QD}_{[B]}) &= \max_{F\in\mathcal{B}} \sum_{E\in\mathcal{S}_0} \rho((\mathbf{QD})_{[E,F]}) \\ &= \max_{F\in\mathcal{B}} \sum_{S_j:j=1,\ldots,k} \rho((\mathbf{QD})_{[S_j,F]}) \quad \text{(IV.33)}\end{aligned}$$

and analyze each of its terms. By definition of $\mathcal{B}$, each $F \in \mathcal{B}$ corresponds to some $G_i = [S_i\ T_i]$ for some $i \le k$. We can thus write $(\mathbf{QD})_{[S_j,F]} = [(\mathbf{QD})_{[S_j,S_i]}(\mathbf{QD})_{[S_j,T_i]}]$. Then, by recalling that $\mathbf{QD}_{[T_0]} = \mathbf{0}$ we see that $(\mathbf{QD})_{[S_j,T_i]} = \mathbf{0}$ for all i,j. Now, when $i = j$ we have $(\mathbf{QD})_{[S_j,S_i]} = \mathbf{I}$, thus $\rho((\mathbf{QD})_{[S_j,F]}) = \rho([\mathbf{I0}]) = 1$. When $i \ne j$, $(\mathbf{QD})_{[S_j,S_i]} = \mathbf{0}$, and $\rho((\mathbf{QD})_{[S_j,F]}) = \rho([\mathbf{00}]) = 0$ in that case. From (IV.33), we conclude that $\rho_{[\mathcal{S}_0,\mathcal{B}]}(\mathbf{QD}_{[B]}) = 1$. Plugging into (IV.32) leads to

$$\psi_{\mathcal{G}}(\mathbf{a}) < \psi_{\mathcal{G}}(\mathbf{a}'_{[B]}) + \alpha\psi_{\mathcal{G}}(\mathbf{a}'_{[\bar{B}]}). \quad \text{(IV.34)}$$

For the ℓ_1 term, we follow the same path as (IV.30) and (IV.31), now using the Moore-Penrose pseudo-inverse $\mathbf{H}$ instead, yielding $\mathbf{a}_{[S_0]} = \mathbf{HD}_{[B]}\mathbf{a}'_{[B]} + \mathbf{HD}_{[\bar{B}]}\mathbf{a}'_{[\bar{B}]}$, from which $\|\mathbf{a}\|_1 \le \left\|\mathbf{HD}_{[B]}\mathbf{a}'_{[B]}\right\|_1 + \left\|\mathbf{HD}_{[\bar{B}]}\mathbf{a}'_{[\bar{B}]}\right\|_1$ follows. Using the fact that $\|\mathbf{Wv}\|_{1,1} \le \|\mathbf{W}\|_{1,1}\|\mathbf{v}\|_1$[32], we get $\|\mathbf{a}\|_1 \le \left\|\mathbf{HD}_{[B]}\right\|_{1,1}\left\|\mathbf{a}'_{[B]}\right\|_1 + \left\|\mathbf{HD}_{[\bar{B}]}\right\|_{1,1}\left\|\mathbf{a}'_{[\bar{B}]}\right\|_1$. Now, since $B \subset G_0$, and $\left\|\mathbf{HD}_{[G_0]}\right\|_{1,1} = 1$, we have that $\left\|\mathbf{HD}_{[B]}\right\|_{1,1} \le 1$. Together with condition (IV.29) this yields

$$\|\mathbf{a}\|_1 < \left\|\mathbf{a}'_{[B]}\right\|_1 + \gamma\left\|\mathbf{a}'_{[\bar{B}]}\right\|_1. \quad \text{(IV.35)}$$

Combining (IV.34) and (IV.35) into the HiLasso cost function we get

$$\begin{aligned}\lambda\psi_{\mathcal{G}}(\mathbf{a}) + (1-\lambda)\|\mathbf{a}\|_1 < \lambda\left[\psi_{\mathcal{G}}(\mathbf{a}'_{[B]}) + \alpha\psi_{\mathcal{G}}(\mathbf{a}'_{[\bar{B}]})\right] \\ +(1-\lambda)\left[\left\|\mathbf{a}'_{[B]}\right\|_1 + \gamma\left\|\mathbf{a}'_{[\bar{B}]}\right\|_1\right]. \quad \text{(IV.36)}\end{aligned}$$

[5]Note that the statement of Lemma 1 as shown here is actually a slight generalization of [28, Lemma3], where the groups in the partitions need not have the same size.

Now, to finish the proof, we need to bound the right-hand side of (IV.36) by $\lambda\psi_{\mathcal{G}}(\mathbf{a}') + (1-\lambda)\|\mathbf{a}'\|_1$, in order to show that the alternate $\mathbf{a}'$ is not a minimum of the HiLasso problem. For any γ satisfying

$$\gamma \le 1 + \frac{\lambda(1-\alpha)\psi_{\mathcal{G}}(\mathbf{a}'_{[\bar{B}]})}{(1-\lambda)\left\|\mathbf{a}'_{[\bar{B}]}\right\|_1},$$

we have

$$\begin{aligned}\lambda[\psi_{\mathcal{G}}(\mathbf{a}'_{[B]}) + \alpha\psi_{\mathcal{G}}(\mathbf{a}'_{[\bar{B}]})] + (1-\lambda)[\left\|\mathbf{a}'_{[B]}\right\|_1 + \gamma\left\|\mathbf{a}'_{[\bar{B}]}\right\|_1] \\ \le \lambda\psi_{\mathcal{G}}(\mathbf{a}') + (1-\lambda)\|\mathbf{a}'\|_1 \quad \text{(IV.37)}\end{aligned}$$

where we have used the fact that $\|\mathbf{a}'\|_1 = \left\|\mathbf{a}'_{[B]}\right\|_1 + \left\|\mathbf{a}'_{[\bar{B}]}\right\|_1$ and $\psi_{\mathcal{G}}(\mathbf{a}') = \psi_{\mathcal{G}}(\mathbf{a}'_{[B]}) + \psi_{\mathcal{G}}(\mathbf{a}'_{[\bar{B}]})$. To obtain a signal independent relationship between γ and α, we bound $\psi_{\mathcal{G}}(\mathbf{a}'_{[\bar{B}]})$ in terms of $\left\|\mathbf{a}'_{[\bar{B}]}\right\|_1$,

$$\left\|\mathbf{a}'_{[\bar{B}]}\right\|_1 = \sum_i \left\|\mathbf{a}'_{[\bar{R}_i]}\right\|_1 \le \sum_i \sqrt{g}\left\|\mathbf{a}'_{[\bar{B}_i]}\right\|_2 = \sqrt{g}\psi_g(\mathbf{a}'_{[\bar{B}]}),$$

resulting in the condition

$$\gamma \le 1 + \frac{\lambda(1-\alpha)}{(1-\lambda)\sqrt{g}} \le 1 + \frac{\lambda(1-\alpha)\psi_{\mathcal{G}}(\mathbf{a}'_{[\bar{B}]})}{(1-\lambda)\left\|\mathbf{a}'_{[\bar{B}]}\right\|_1}$$

which completes the proof.

We conclude that we can guarantee recovery for every choice of λ as long as (IV.27)–(IV.29) are satisfied. Note that when $\lambda = 0$ (Lasso mode) we get $\gamma \le 1$, and, as expected, (IV.28)–(IV.29) reduce to the Lasso recovery condition. Also, if $\alpha = 1$ we have $\gamma \le 1$, meaning that we must tighten the constrains related to the ℓ_2 part of the cost function in order to relax the ℓ_1 part. For $\gamma > 1$, the HiLasso conditions are a relaxation of the Lasso conditions, thus allowing for more signals to be correctly recovered.

Theorem 2 below provides signal independent replacements of the conditions (IV.27)–(IV.29). The signal independent bound for (IV.27) derived here, depends on coherence measures between the dictionary $\mathbf{D}$ and its image under the projection $\mathbf{I}-\mathbf{P}$, $\mathbf{C} = (\mathbf{I}-\mathbf{P})\mathbf{D}$. Since $\mathbf{P}$ depends on S_0, $\mathbf{C}$ itself is signal dependent. Thus, we need to maximize also over all possible sets S_0. These are defined as (IV.38)–(IV.41), shown at the bottom of the page. We are now in position to state the theorem.

$$\nu_P := \max\left\{\max\left\{\max\left\{\frac{\mathbf{d}_i^{\mathrm{T}}\mathbf{c}_j}{(\mathbf{d}_i^{\mathrm{T}}\mathbf{c}_i)^{\frac{1}{2}}(\mathbf{d}_j^{\mathrm{T}}\mathbf{c}_j)^{\frac{1}{2}}}, i,j \in G, i \ne j\right\}, G \in \mathcal{G}\right\}, S_0\right\} \quad \text{(IV.38)}$$

$$\mu_P^s := \max\left\{\max\left\{\frac{1}{g}\rho^s(\mathbf{D}_{[G]}^{\mathrm{T}}\mathbf{C}_{[F]}), G, F \in \mathcal{G}, G \ne F\right\}, S_0\right\} \quad \text{(IV.39)}$$

$$\mu_P^{ss} := \max\left\{\max\left\{\frac{1}{g}\rho^{ss}(\mathbf{D}_{[G]}^{\mathrm{T}}\mathbf{C}_{[F]}), G, F \in \mathcal{G}, G \ne F\right\}, S_0\right\} \quad \text{(IV.40)}$$

$$\zeta := \max\left\{\max\{(\mathbf{d}_i^{\mathrm{T}}\mathbf{c}_i)^{-\frac{1}{2}} : i = 1,\ldots,p\}, S_0\right\}. \quad \text{(IV.41)}$$

Theorem 2: Let χ, ν_P, μ_P^s, μ_P^{ss} and ζ be the coherence measures defined respectively in (IV.16) and (IV.38)–(IV.41). Then the conditions (IV.27)–(IV.29) are satisfied if

$$\frac{\zeta^2 kg\mu_P^s}{1-(s-1)\nu_P+g\mu_P^{ss}(k-1)\zeta^2}\leq\alpha \qquad \text{(IV.42)}$$

$$\frac{ks\chi}{1-(s-1)\nu-(k-1)s\chi}<\gamma \qquad \text{(IV.43)}$$

$$\frac{ks\nu}{1-(s-1)\nu-(k-1)s\chi}<1. \qquad \text{(IV.44)}$$

We also require the denominators in (IV.42)–(IV.44) to be positive. Note that, although the interpretation of (IV.42) is rather counter-intuitive, it is easy to check that $\mu_P^s, \mu_P^{ss}\leq\mu_B$. This can be seen when $s=g$ (a case included in our theorems), in which case $\mathbf{P}=\mathbf{0}$, $\mathbf{C}=\mathbf{D}$, and $\mu_P^s=\mu_P^{ss}=\mu_B$. Therefore, the condition (IV.42) is a relaxation of the standard (dense) block-sparse recovery one [28, Theorem 2].

Proof: Recall that

$$\mathbf{Q}\mathbf{D}_{[\bar{G}_0]}=\left(\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{C}_{[S_0]}\right)^{-1}\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{C}_{[\bar{G}_0]}.$$

Since $\rho_{[\cdot,\cdot]}(\cdot)$ is submultiplicative, [28],[6]

$$\rho_{[\mathcal{S}_0,\bar{\mathcal{G}}_0]}(\mathbf{Q}\mathbf{D}_{[\bar{G}_0]}) \leq \rho_{[\mathcal{S}_0,\mathcal{S}_0]}((\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{C}_{[S_0]})^{-1})\,\rho_{[\mathcal{S}_0,\bar{\mathcal{G}}_0]}(\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{C}_{[\bar{G}_0]}). \qquad \text{(IV.45)}$$

Applying the definitions of $\rho_{[\mathcal{S}_0,\bar{\mathcal{G}}_0]}$ and μ_P^s, we have

$$\begin{aligned}\rho_{[\mathcal{S}_0,\bar{\mathcal{G}}_0]}(\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{C}_{[\bar{G}_0]}) &= \max_{F\in\bar{\mathcal{G}}_0}\sum_{E\in\mathcal{S}_0}\rho(\mathbf{D}_{[E]}^{\mathrm{T}}\mathbf{C}_{[F]})\\ &\leq k\max_{F\in\bar{\mathcal{G}}_0}\max_{E\in\mathcal{S}_0}\{\rho(\mathbf{D}_{[E]}^{\mathrm{T}}\mathbf{C}_{[F]})\}\leq kg\mu_P^s\end{aligned} \qquad \text{(IV.46)}$$

where the last inequality in (IV.46) derives from (IV.39) and the fact that each $E\in\mathcal{S}_0$ belongs to some G_i, and $|E|=s$, thus playing the role of the set T in the definition of $\rho^s(\cdot)$. Our goal is now to obtain a bound for $\rho_{[\mathcal{S}_0,\mathcal{S}_0]}((\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{C}_{[S_0]})^{-1})$. To this end, we define $\mathbf{Z}=\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{C}_{[S_0]}$, and rewrite it as $\mathbf{Z}=\Lambda^{-1}(\mathbf{I}-(\mathbf{I}-\Lambda\mathbf{Z}\Lambda))\Lambda^{-1}$. Here Λ is a $ks\times ks$ block-diagonal scaling matrix to be defined later. Assume for now that $\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)<1$. This allows us to apply the following result from [28]:

Lemma 2: Suppose that $\rho_{[\mathcal{E},\mathcal{F}]}(\mathbf{W})<1$. Then $(\mathbf{I}+\mathbf{W})^{-1}=\sum_{k=0}^{\infty}(-\mathbf{W})^k$.

By applying Lemma 2 to $\mathbf{W}=-\mathbf{I}+\Lambda\mathbf{Z}\Lambda$ we can write $\mathbf{Z}^{-1}=\Lambda\left[\sum_{i=0}^{\infty}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)^i\right]\Lambda$. With this,

$$\begin{aligned}\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{Z}^{-1}) &\overset{(a)}{\leq} \left[\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\Lambda)\right]^2\rho_{[\mathcal{S}_0,\mathcal{S}_0]}\left(\sum_{i=0}^{\infty}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)^i\right)\\ &\overset{(b)}{\leq} \left[\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\Lambda)\right]^2\sum_{i=0}^{\infty}\rho_{[\mathcal{S}_0,\mathcal{S}_0]}\left((\mathbf{I}-\Lambda\mathbf{Z}\Lambda)^i\right)\\ &\overset{(c)}{\leq} \left[\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\Lambda)\right]^2\sum_{i=0}^{\infty}\left(\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)\right)^i\\ &\overset{(d)}{\leq} \frac{[\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\Lambda)]^2}{1-\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)}\end{aligned} \qquad \text{(IV.47)}$$

where in (a) and (c), we applied the submultiplicativity of $\rho_{[\cdot,\cdot]}(\cdot)$, (b) is a consequence if the triangle inequality, and (d) is the limit of the geometric series, which is finite when $\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)<1$.

We now bound $\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)$. First, note that, since Λ is block-diagonal, we have that $(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)_{[S_i,S_j]}=\mathbf{I}_{[S_i,S_j]}-\Lambda_{[S_i,S_i]}\mathbf{Z}_{[S_i,S_j]}\Lambda_{[S_j,S_j]}$. We then choose Λ to be a diagonal matrix with $\Lambda_{ii}=(\mathbf{d}_i^{\mathrm{T}}\mathbf{c}_i)^{-\frac{1}{2}}$, $i\in S_0$. With this choice, we have that the diagonal elements of $\mathbf{I}_{[S_j,S_j]}-\Lambda_{[S_j,S_j]}\mathbf{Z}_{[S_j,S_j]}\Lambda_{[S_j,S_j]}$ are equal to 1 for all j, and the off-diagonal elements are bounded by ν_P. Using Geršgorin Theorem we then have that

$$\rho(\mathbf{I}_{[S_j,S_j]}-\Lambda_{[S_j,S_j]}\mathbf{Z}_{[S_j,S_j]}\Lambda_{[S_j,S_j]})\leq(s-1)\nu_P, \quad j=1,\ldots,k. \qquad \text{(IV.48)}$$

As for the off-diagonal $s\times s$ blocks of $\mathbf{I}-\Lambda\mathbf{Z}\Lambda$, we have $(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)_{[S_i,S_j]}=-\Lambda_{[S_i,S_i]}\mathbf{Z}_{[S_i,S_j]}\Lambda_{[S_j,S_j]}$. We then have

$$\begin{aligned}\rho((\mathbf{I}-\Lambda\mathbf{Z}\Lambda)_{[S_i,S_j]}) &\overset{(a)}{\leq} \rho(\Lambda_{[S_i,S_i]})\rho(\mathbf{Z}_{[S_i,S_j]})\rho(\Lambda_{[S_j,S_j]})\\ &\overset{(b)}{\leq} \zeta(g\mu_P^{ss})\zeta\end{aligned} \qquad \text{(IV.49)}$$

where in (a) we used the submultiplicativity of $\rho(\cdot)$, and (b) derives from the definition of μ_P^{ss}, and the fact that, with our choice of Λ we have $\rho(\Lambda_{[S_i,S_i]})\leq\zeta$ for all i. Now we can write the definition of $\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)$ and bound its summation using (IV.48)–(IV.49):

$$\begin{aligned}&\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)\\ &\quad\leq\max_{S_j:j\leq k}\Big\{\rho(\mathbf{I}_{[S_j,S_j]}-\Lambda_{[S_j,S_j]}\mathbf{Z}_{[S_j,S_j]}\Lambda_{[S_j,S_j]})+\cdots\\ &\qquad\cdots\sum_{S_i:i\leq k,i\neq j}\rho\left(\Lambda_{[S_i,S_i]}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)_{[S_i,S_j]}\Lambda_{[S_j,S_j]}\right)\Big\}\\ &\quad\leq(s-1)\nu_P+g\mu_P^{ss}\zeta^2.\end{aligned} \qquad \text{(IV.50)}$$

By our choice of Λ, $\rho(\Lambda_{[S_i,S_i]})\leq\zeta$ and $\rho(\Lambda_{[S_i,S_j]})=0$ for $i\neq j$. Therefore, $\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\Lambda)\leq\zeta$ as well. Using this together with (IV.50) in (IV.47), we obtain

$$\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{Z}^{-1})\leq\frac{\zeta^2}{1-(s-1)\nu_P+g\mu_P^{ss}(k-1)\zeta^2}. \qquad \text{(IV.51)}$$

To ensure that $\rho_{[\mathcal{S}_0,\mathcal{S}_0]}(\mathbf{I}-\Lambda\mathbf{Z}\Lambda)<1$, we need the denominator in the above equation to be positive. Now (IV.42) follows by plugging (IV.46) and (IV.51) into (IV.45),

$$\rho_{[\mathcal{S}_0,\bar{\mathcal{G}}_0]}(\mathbf{Q}\mathbf{D}_{[\bar{G}_0]})\leq\frac{\zeta^2 kg\mu_P^s}{1-(s-1)\nu_P+g\mu_P^{ss}(k-1)\zeta^2}.$$

[6]There is a slight abuse of notation here, in that, in our case of nonsquare blocks, each norm $\rho_{[\cdot,\cdot]}(\cdot)$ in the right-hand side of the submultiplicativity inequality (IV.45) is actually a different norm. However, it is easy to see that the referred inequality holds in this case as well.

TABLE I

SIMULATED SIGNAL RESULTS. IN EVERY 2×2 CELL CONTAINS THE MSE ($\times 10^4$) AND HAMMING DISTANCE (MSE/HAMMING) FOR LASSO (TOP, LEFT), GLASSO (TOP, RIGHT), HILASSO (BOTTOM, LEFT), AND C-HILASSO (BOTTOM, RIGHT). IN THE FIRST CASE (LEFT) WE VARY THE NOISE σ WHILE KEEPING $q = 8$ AND $s = 8$ FIXED. IN THE SECOND AND THIRD CASES, WE HAVE $\sigma = 0$. FOR THE SECOND EXPERIMENT (CENTER), WE FIXED $q = 8$ WHILE CHANGING s. IN THE THIRD CASE, WE FIX $s = 12$ AND VARY THE NUMBER OF GROUPS q. BOLD BLUE INDICATES THE BEST RESULTS, ALWAYS OBTAINED FOR THE PROPOSED MODELS. IN ALL CASES, THE NUMBER OF ACTIVE GROUPS IS $k = 2$

$\sigma = 0.1$	417 / 22.0	1173 / 361.6	$s = 8$	388 / 22.0	1184 / 318.2	$q = 4$	1080 / 27.8	1916 / 221.7
	330 / 19.8	**163** / **13.3**		272 / 19.5	**96** / **16.2**		1009 / **29.8**	**742** / 30.2
$\sigma = 0.2$	564 / 21.6	1182 / 378.3	$s = 12$	1200 / 36.2	1166 / 350.4	$q = 8$	1200 / 36.2	1166 / 350.4
	399 / 22.7	**249** / **17.1**		704 / **26.5**	**413** / 29.1		704 / **26.5**	**413** / 29.1
$\sigma = 0.4$	965 / 22.7	1378 / 340.3	$s = 16$	1641 / 43.9	1093 / 338.6	$q = 12$	1030 / 41.8	840 / 447.7
	656 / **19.5**	**595** / 27.4		1100 / **32.2**	**551** / 35.0		662 / **26.4**	**4** / 29.8

Finally, we use the same ideas to bound $\|\mathbf{H}\mathbf{D}_{[\vec{G}_0]}\|_{1,1}$ and derive (IV.43). Specifically

$$\|\mathbf{H}\mathbf{D}_{[\vec{G}_0]}\|_{1,1} \leq \|(\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{D}_{[S_0]})^{-1}\|_{1,1}\|\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{D}_{[\vec{G}_0]}\|_{1,1}. \tag{IV.52}$$

Now

$$\|(\mathbf{D}_{[S_0]})^{\mathrm{T}}\mathbf{D}_{[\vec{G}_0]}\|_{1,1} = \max_{j\in\vec{G}_0}\sum_{i\in S_0}|\mathbf{d}_i^{\mathrm{T}}\mathbf{d}_j| \overset{(a)}{\leq} ks\chi \tag{IV.53}$$

where (a) follows from the definition of χ and the fact that $|S_0| = ks$. It remains to develop a bound on $\|(\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{D}_{[S_0]})^{-1}\|_{1,1}$. To this end we express $\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{D}_{[S_0]} = \mathbf{I} + \mathbf{W}$, and bound

$$\begin{aligned}\|\mathbf{W}\|_{1,1} &= \max_{r\leq k}\left\{\max_{i\in S_r}\left\{\sum_{j\in S_r, j\neq i}|\mathbf{d}_i^{\mathrm{T}}\mathbf{d}_j| + \sum_{j\in S_0\setminus S_r}|\mathbf{d}_i^{\mathrm{T}}\mathbf{d}_j|\right\}\right\}\\ &\leq (s-1)\nu + s(k-1)\chi\end{aligned} \tag{IV.54}$$

since for all $S_r, r \leq k$, and all $i \in S_r$, the first sum has $(s-1)$ nonzero elements bounded by ν, and the second sum has $s(k-1)$ elements bounded by χ. Now, by requiring $(s-1)\nu + s(k-1)\chi < 1$ we can apply Lemma 2 to $\mathbf{W}$ and follow the same path as the one that leads to (IV.50), now using the matrix norm properties of $\|\cdot\|_{1,1}$, to obtain

$$\|(\mathbf{D}_{[S_0]}^{\mathrm{T}}\mathbf{D}_{[S_0]})^{-1}\|_{1,1} \leq \frac{1}{1-(s-1)\nu+s(k-1)\chi}. \tag{IV.55}$$

Again, $(s-1)\nu + s(k-1)\chi < 1$ is implicit in the requirement that the above denominator be positive. Plugging (IV.55) and (IV.53) into (IV.52) yields (IV.43).

The proof for (IV.44) is analogous to that of (IV.43), only that now the upper bound on $|\mathbf{d}_i^{\mathrm{T}}\mathbf{d}_j|, i \in S_0, j \in T_0$, is $\nu \leq \mu$. Continuing as before leads to (IV.44).

Theorems 1 and 2 are for the noncollaborative case. For the collaborative case there exist results that show that both the C-Lasso [10] and C-GLasso [27] will recover the true shared active set with a probability of error that vanishes exponentially with n. Since the in-group active sets are not necessarily equal for all samples in $\mathbf{X}$, C-HiLasso could only help in recovering the group sparsity pattern. Since the C-GLasso is a special case of C-HiLasso when $\lambda_1 = 0$, we can conjecture that when $\lambda_1 > 0$, the accuracy of the C-HiLasso in recovering the correct groups will improve with larger n. Furthermore, since our results for HiLasso improve on those of the Group Lasso, it is to be expected that the accuracy of C-HiLasso, for an appropriate $\lambda_1 > 0$, will be better than that of C-GLasso.

As an intuitive explanation to why this may happen, the proofs in [10] and [27] assume a continuous probability distribution on the nonzero coefficients of the signals, and give recovery results for the average case. On the other hand, the in-group sparsity assumption of C-HiLasso implies that only s out of g samples will be nonzero within each group. This implies that, for the same group sparsity pattern, there will be much less (exactly a fraction $\frac{s}{g}$) nonzero elements in the possible signals compared to the ones that can occur under the hypothesis of C-GLasso. Since any assumed distribution of the signals under the in-group sparsity hypothesis has to be concentrated on this much smaller set of possible signals, they should be easier to recover correctly from solutions to the C-HiLasso program, compared to the dense group case of C-GLasso.

V. EXPERIMENTAL RESULTS

In this section we show the strength of the proposed HiLasso and C-HiLasso models. We start by comparing our model with the standard Lasso and Group Lasso using synthetic data. We created q dictionaries, $\mathbf{D}_r, r = 1,\ldots,q$, with $g = 64$ atoms of dimension $m = 64$, and i.i.d. Gaussian entries. The columns were normalized to have unit ℓ_2 norm. We then randomly chose $k = 2$ groups to be active at each time (on all the signals). Sets of $n = 200$ normalized testing signals were generated, one per active group, as linear combinations of $s \ll 64$ elements of the active dictionaries, $\mathbf{x}_j^r = \mathbf{D}_r\mathbf{a}_j^r$. The mixtures were created by summing these signals and (eventually) adding Gaussian noise of standard deviation σ. The generated testing signals have a hierarchical sparsity structure and while they share groups, they do not necessarily share the sparsity pattern inside the groups. We then built a single dictionary by concatenating the subdictionaries, $\mathbf{D} = [\mathbf{D}_1,\ldots,\mathbf{D}_q]$, and used it to solve the Lasso, Group Lasso, HiLasso, and C-HiLasso problems. Table I summarizes the mean-square error (MSE) and Hamming distance of the recovered coefficient vectors $\mathbf{a}_j, j = 1,\ldots,n$. We observe that our model is able to exploit the hierarchical structure of the data as well as the collaborative structure. Group Lasso selects in general the correct blocks but it does not give a sparse solution within them. On the other hand, Lasso gives a solution that has nonzero elements belonging to groups that were not active in the original signal, leading to a wrong model/class selection. HiLasso gives a sparse solution that picks atoms from the correct

TABLE II
NOISY DIGIT MIXTURES RESULTS. FOUR DIFFERENT CASES ARE SHOWN: WHEN EACH SIGNAL IS A SINGLE DIGIT AND WHEN IT IS THE MIXTURE OF TWO DIFFERENT (RANDOMLY SELECTED) DIGITS, WITH AND WITHOUT ADDITIVE GAUSSIAN NOISE WITH STANDARD DEVIATION 10% OF THE PEAK VALUE. FOR THE 2-DIGITS CASE, RESULTS ARE THE AVERAGE OF EIGHT RUNS (IN EACH ROUND, A NEW PAIR OF DIGITS WAS RANDOMLY SELECTED). IN THE SINGLE-DIGIT CASE, THE RESULT IS THE AVERAGE OF THE TEN POSSIBLE SITUATIONS. BOTH AMSE AND HAMMING DISTANCE ARE SHOWN, WITH BOLD BLUE INDICATING BEST. WITHOUT NOISE, BOTH C-GLASSO AND C-HILASSO YIELD VERY GOOD RESULTS. HOWEVER, IN THE NOISY CASE, C-HILASSO IS CLEARLY SUPERIOR, SHOWING THE ADVANTAGE OF ADDING REGULARIZATION INSIDE THE GROUPS FROM A ROBUSTNESS PERSPECTIVE. SEE ALSO FIG. 4

experiment	Lasso		GLasso		HiLasso		C-GLasso		C-HiLasso	
	AMSE	Hamm	AMSE	Hamm	AMSE	Hamm	AMSE	Hamm	AMSE	Hamm
1 digit	0.06	0.43	0.07	0.78	0.02	0.19	**0.01**	**0.02**	0.02	0.06
1 digit+n	0.08	1.31	0.08	0.87	0.04	0.48	0.05	0.25	**0.02**	**0.01**
2 digit	0.09	1.46	0.08	1.86	0.02	1.18	**0.01**	**0.74**	0.02	0.90
2 digit+n	0.11	2.21	0.08	1.99	0.04	1.46	0.09	1.60	**0.03**	**0.70**

groups but still presents some minor mistakes. For the collaborative case, in all the tested configurations, no coefficients were selected outside the correct active groups, and the recovered coefficients are consistently the best ones.

In all the examples, and for each method, the regularization parameters were the ones for which the best results were obtained. One can scale the parameter λ_2 to account for different number of signals. This situation is analogous to a change in the size of the dictionary, thus, λ_2 should be proportional to the square root of the number of signals to code.

We then experimented with the USPS digits dataset, which has been shown to be well represented in the sparse modeling framework [37]. Here the signals are vectors containing the unwrapped gray intensities of 16×16 images ($m = 256$). We obtained each of the $n = 200$ samples in the testing data set as the mixture of two randomly chosen digits, one from each of the two drawn sets of digits. In this case we only have ground truth at the group level. We measure the recovery performance in terms of the average MSE of the recovered signals, $\text{AMSE} = \frac{1}{nq}\sum_{r=1}^{q}\sum_{j=1}^{n}\left\|\mathbf{x}_j^r - \hat{\mathbf{x}}_j^r\right\|_2^2$, where $\mathbf{x}_j^r$ is the component corresponding to source r in the signal j, and $\hat{\mathbf{x}}_j^r$ is the recovered one.

Using the usual training-testing split for USPS, we first learned a dictionary for each digit. We then created a single dictionary by concatenating them. In Table II we show the AMSE obtained while summing $k = 2$ different digits. We also consider the situation were only one digit is present. C-HiLasso automatically detects the number of sources while achieving the best recovery performance. As in the synthetic case, only the collaborative method was able to successfully detect the true active classes. In Fig. 4 we relax the assumption that all the signals have to contain exactly the same type and amount of classes in the mixture, further demonstrating the flexibility of the proposed C-HiLasso model.

We also used the digits dataset to experiment with missing data. We randomly discarded an average of 60% of the pixels per mixed image and then applied C-Hilasso. The algorithm is capable of correctly detecting which digits are present in the images. Some example results for this case are shown in Fig. 5. Note that this is a quite different problem than the one commonly addressed in the matrix completion literature. Here we do not aim to recover signals that all belong to a unique unknown subspace, but signals that are the combination of two nonunique spaces to be automatically identified from the available dictionary. Such unknown spaces have common models/groups for all

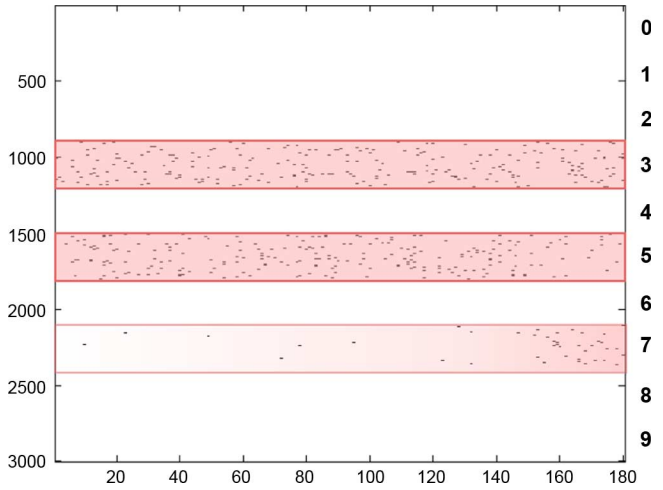

Fig. 4. In this example, we used C-HiLasso to analyze mixtures where the data set contains different number and types of sources/classes. We used a set containing 180 mixtures of digit images. The first 150 images are obtained as the sum/mixture of a number "3" and an number "5" (randomly selected). Each of the last 30 images in the set are the mixture of three numbers: "3" ,"5", and "7" (the 180 images are of course presented at random, the algorithm is not *a priori* aware which images contain two sources and which contain three). The figure shows the active sets of the recovered coefficients matrix $\mathbf{A}$ as a binary matrix the same size as $\mathbf{A}$ (atom indexes in the vertical and sample indexes in the horizontal), where black dots indicate nonzero coefficients. C-HiLasso managed to identify the active blocks while the subdictionary corresponding to "7" is mostly active for the last 30 images. The accuracy of this result depends on the relationship between the subdictionaries corresponding to each digit.

the signals in question (the coarse level of the hierarchy), but not necessarily the exact same atoms inside the groups and therefore do not necessarily belong to the same subspaces. Both levels of the hierarchy are automatically detected, e.g., the groups corresponding to "3" and "5," and the corresponding reconstructing atoms (subspaces) in each group, these last ones possibly different for each signal in the set. While we consider that the possible subspaces are to be selected from the provided dictionary (learned off-line from training data), in Section VI we discuss learning such dictionaries as part of the optimization as well (see also [38], [39]). In such cases, the standard matrix completion problem becomes a particular case of the C-HiLasso framework (with a single group and all the signals having the same active set, subspace, in the group), naturally opening numerous theoretical questions for this new more general model.[7]

[7]Prof. Carin and collaborators have new results on the case of a single group and signals in possible different subspaces of the group, an intermediate model between standard matrix completion and C-HiLasso (personal communication).

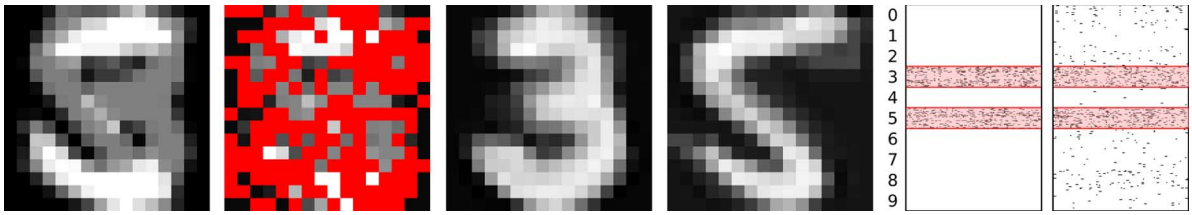

Fig. 5. Example of recovered digits (3 and 5) from a mixture with 60% of missing components. From left to right: noiseless mixture, observed mixture with missing pixels highlighted in red, recovered digits 3 and 5, and active set recovered for all samples using the C-HiLasso and Lasso, respectively. In the last two figures, the active sets are represented as in Fig. 4. The coefficients blocks for digits 3 and 5 are marked as pink bands. Notice that the C-HiLasso exploits efficiently the hypothesis of collaborative group-sparsity, succeeding in recovering the correct active groups in all the samples. The Lasso, which lacks this prior knowledge, is clearly not capable of doing so, and active sets are spread over all the groups.

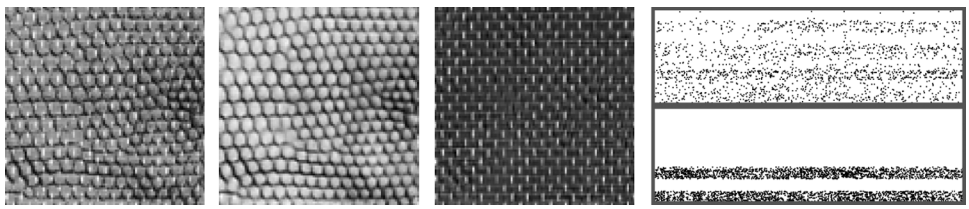

Fig. 6. Texture separation results. Left to right: Sample mixture, corresponding C-HiLasso separated textures, and comparison of the active set diagrams obtained by the Lasso (as in Fig. 5). The one for Lasso is shown on top, where all groups are wrongly active, and the one for C-HiLasso on bottom, showing that only the two correct groups are selected.

TABLE III

TEXTURE SEPARATION RESULTS. THE ROWS AND COLUMNS INDICATE THE ACTIVE TEXTURES IN EACH CELL. THE UPPER TRIANGLE CONTAINS THE AMSE ($\times 10^4$) RESULTS, WHILE THE LOWER TRIANGLE SHOWS THE HAMMING ERROR IN THE GROUP-WISE ACTIVE SET RECOVERY. WITHIN EACH CELL, RESULTS ARE SHOWN FOR THE LASSO (TOP LEFT), GROUP LASSO (BOTTOM LEFT), COLLABORATIVE GROUP LASSO (TOP RIGHT) AND COLLABORATIVE HIERARCHICAL LASSO (BOTTOM RIGHT). THE BEST RESULTS ARE IN BLUE BOLD. NOTE THAT, BOTH FOR THE AMSE AND HAMMING DISTANCE, IN 26 OUT OF 28 CASES, OUR MODEL OUTPERFORMS PREVIOUS ONES

			110	214	**18**	074	63	78	19	47	85	174	107	447	7	43
			117	**69**	069	**18**	126	**38**	47	**18**	132	**51**	102	**42**	27	**3**
	2.80	0.42			107	76	141	129	91	83	191	234	240	219	68	105
	1.36	**0.00**			182	**68**	209	**102**	100	**78**	257	**141**	245	**178**	95	**19**
	0.33	0.25	3.65	**0.00**			52	**42**	35	62	105	112	162	141	21	93
	2.06	**0.00**	2.67	0.02			158	43	83	**29**	214	**62**	200	**107**	102	**10**
	0.96	0.01	3.69	0.07	1.74	**0.00**			**49**	72	123	145	182	148	26	89
	1.97	**0.00**	2.30	**0.00**	2.42	**0.00**			81	55	224	**98**	214	**107**	85	**10**
	1.02	1.00	3.55	1.00	1.42	1.00	2.25	1.00			85	76	120	87	15	63
	2.25	**0.09**	2.52	**0.94**	3.39	**0.16**	2.85	**0.35**			120	**59**	107	**71**	41	**9**
	2.26	0.32	4.12	**0.53**	3.48	0.44	3.49	0.32	3.16	1.00			229	240	56	95
	2.50	**0.00**	3.23	0.82	3.54	**0.20**	3.11	**0.01**	4.07	**0.40**			245	**162**	117	**27**
	4.37	1.39	4.47	**0.08**	4.09	0.13	4.23	0.12	4.20	1.00	4.42	0.42			100	112
	2.51	**0.02**	2.39	0.22	2.42	**0.02**	2.76	**0.02**	2.24	**0.20**	2.96	**0.11**			102	**51**
	0.09	0.98	3.77	1.00	0.31	1.00	1.83	1.00	1.13	1.00	3.14	0.97	4.30	1.00		
	0.53	**0.00**	1.75	**0.01**	2.04	**0.00**	1.82	**0.00**	2.18	**0.00**	3.04	**0.24**	1.90	**0.18**		

We also compared the performance of C-HiLasso, Lasso, GLasso and C-GLasso (without hierarchy) in the task of separating mixed textures in an image. In this case, the set of signals $\mathbf{X}$ corresponds to all 12×12 patches in the (single) image to be analyzed. We chose eight textures from the Brodatz dataset and trained one dictionary for each one of them using one half of the respective images (these form the $g = 8$ groups of the dictionary). Then, we created an image as the sum of the other halves of the $k = 2$ textures. One can think of this experiment as a generalization to the texture separation problem proposed in [40] (without additive noise), where only two textures are present. The experiment was repeated for all possible combinations of two textures from the eight possible ones. The results are summarized in Table III. A detailed example is shown in Fig. 6. For each algorithm, the best parameters were chosen using grid search, ensuring that those were not in the edges of the grid. For Lasso and C-HiLasso, the best λ_1 is 0.0625. For GLasso and C-GLasso, the best λ_2 was, respectively, 0.05 and 75 (for the collaborative setting, we heuristically scale λ_2 with the number of signals as $\sqrt{n}$. In this experiment, $n \approx 512^2$, leading to such large value of λ_2). From Table III, we can conclude that the C-HiLasso is significantly better than the competing algorithms, both in the MSE of the recovered signals (we show the AMSE of recovering both active signals), and in the average Hamming distance between the recovered groupwise active sets and the true ones. In the latter case, we observe that, in many cases, the C-HiLasso active set recovery performance is perfect (Hamming distance 0) or near perfect,

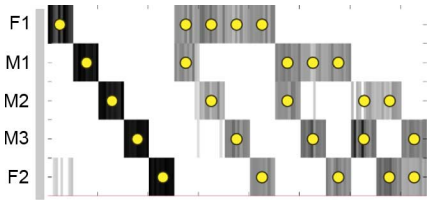

Fig. 7. Speaker identification results. Each column corresponds to the sources identified for a specific time frame, the true ones marked by yellow dots. The vertical axis indicates the estimated activity of the different sources, where darker colors indicate higher energy. For each possible combination of speakers, ten frames (15 seconds of audio) were evaluated.

whereas the other methods seldom approach a Hamming distance lower than 1.

Finally, we use C-HiLasso to automatically identify the sources present in a mixture of audio signals [41]. The goal is to identify the speakers talking simultaneously on a single recording. Here, the task is not to fully reconstruct each of the unmixed sources from the observed signal but to identify which speakers are active. In this case, since the original sources do not need to be recovered, the modeling can be done in terms of features extracted from the original signals in a linear but nonbijective way.

Audio signals have in general very rich structures and their properties rapidly change over time. A natural approach is to decompose them into a set of overlapping local time-windows, where the properties of the signal remain stable. There is a straightforward analogy with the approach explained above for the texture segmentation case, where images were decomposed into collections of overlapping patches. These time-windows will collaborate in the identification.

A challenging aspect when identifying audio sources is to obtain features that are specific to each source and at the same time invariant to changes in the fundamental frequency (pitch) of the sources. In the case of speech, a common choice is to use the short-term power spectrum envelopes as feature vectors [42] (refer to [41] for details on the feature extraction process and implementation). The spectral envelope in human speech varies along time, producing different patterns for each phoneme. Thus, a speaker does not produce an unique spectral envelope, but a set of spectral envelopes that live in a union of manifolds. Since such manifolds are well represented by sparse models, the problem of speaker identification is well suited for the proposed C-HiLasso framework, where each block in the dictionary is trained for the features corresponding to a given speaker, and the overlapping time-windows collaborate in detecting the active blocks.

For this experiment, we use a dataset consisting of recordings of five different German radio speakers, two female and three male. Each recording is six minutes long. One quarter of the samples were used for dictionary training, and the rest for testing. For each speaker, we learned a subdictionary from the training dataset. For testing, we extracted ten nonoverlapping frames of 15 seconds each (including silences made by the speakers while talking), and encoded them using C-HiLasso. The experiment was repeated for all possible combinations of two speakers, and all the speakers talking alone. The results are presented in Fig. 7. C-HiLasso manages to detect automatically the number of sources very accurately, as well as the actual active speakers. Again, refer to [41] for comparisons with other sparse modeling methods (showing the clear advantage of C-HiLasso) and results obtained for the identification of wind instruments in musical recordings.

VI. Discussion

We introduced a new framework of collaborative hierarchical sparse coding, where multiple signals collaborate in their encoding, sharing code groups (models) and having (possible disjoint) sparse representations inside the corresponding groups. An efficient optimization approach was developed, which guarantees convergence to the global minimum, and examples illustrating the power of this framework were presented. At the practical level, we are currently continuing our work on the applications of this proposed framework in a number of directions, including collaborative instruments separation in music, signal classification, and speaker recognition, following the here demonstrated capability to collectively select the correct groups/models.

At the theoretical level, a whole family of new problems is opened by this proposed framework, some of which we already addressed in this work. A critical one is the overall capability of selecting the correct groups in the collaborative scenario, with missing information, and thereby of performing correct model selection and source identification and separation. Results in this direction will be reported in the future.

Finally, we have also developed an initial framework for learning the dictionary for collaborative hierarchical sparse coding, meaning the optimization is simultaneously on the dictionary and the code. As it is the case with standard dictionary learning, this is expected to lead to significant performance improvements (see [37] for the particular case of this with a single group active at a time).

Acknowledgment

The authors would like to thank Dr. T. Nguyen, who, when presented with this model, motivated them to think in a hierarchical fashion and to look at this as just the particular case of a fully hierarchical sparse coding framework. They also thank Prof. T. Luo and G. Mateos for invaluable help on optimization methods. The authors also thank Prof. L. Carin, Dr. G. Yu, and A. Castrodad for very stimulating conversations, and for the fact that their own work also motivated, in part, the example with missing information. The anonymous reviewers prompted an early mistake in the proof of Theorem 1, and that, together with their additional comments, led to improving the bounds in the theorem, as well as the overall presentation of the paper. Finally, they also want to thank the reviewer for the closed-form inner loop of the proposed optimization method, which simplified it and resulted in significant practical improvements.

References

[1] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc., Series B*, vol. 68, pp. 49–67, 2006.

[2] R. Jenatton, J. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," 2009 [Online]. Available: http://arxiv.org/pdf/0904.3523

[3] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.
[4] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.
[5] J. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.
[6] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.
[7] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.
[8] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.
[9] M. Mishali and Y. C. Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4692–4702, Oct. 2008.
[10] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, 2010.
[11] Y. Nesterov, "Gradient methods for minimizing composite objective function," in *Center for Operations Research and Econometrics (CORE)*. Louvain-la-Neuve, Belgium: Catholic Univ. of Louvain, 2007, CORE Discussion Paper 2007/76.
[12] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, 2009.
[13] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the Group Lasso and a Sparse Group Lasso Stanford Univ., Stanford, CA [Online]. Available: http://www-stat.stanford.edu/~tibs
[14] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D. Noh, J. Pollack, and P. Wang, "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," *Ann. Appl. Statist.*, vol. 4, no. 1, pp. 53–77, 2010.
[15] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," *Proc. 27th Int. Conf. Mach. Learning (ICML)*, pp. 543–550, Jun. 2010.
[16] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for sparse hierarchical dictionary learning," *ICML*, Jun. 2010.
[17] J. Starck, M. Elad, and D. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1570–1582, 2004.
[18] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for hierarchical sparse coding," Sep. 2010 [Online]. Available: http://arxiv.org/pdf/1009.2139
[19] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Stat. Soc., Series B*, vol. 58, no. 1, pp. 267–288, 1996.
[20] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.
[21] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[22] R. Giryes, M. Elad, and Y. C. Eldar, "The projected GSURE for automatic parameter tuning in iterative shrinkage methods," *Appl. Comput. Harmon. Anal.*, vol. 30, pp. 407–422, 2010.
[23] I. Ramírez and G. Sapiro, "Sparse coding and dictionary learning based on the MDL principle," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (IEEE ICASSP)*, 2011, pp. 2160–2163.
[24] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin, "Nonparametric Bayesian dictionary learning for sparse image representations," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009.
[25] B. Turlach, W. Venables, and S. Wright, "Simultaneous variable selection," *Technometrics*, vol. 27, pp. 349–363, 2004.
[26] P. Sprechmann, I. Ramirez, and G. Sapiro, "Collaborative hierarchical sparse modeling," *Proc. CISS*, Mar. 2010.
[27] P. Boufounos, G. Kutyniok, and H. Rauhut, "Sparse recovery from combined fusion frame measurements," 2010 [Online]. Available: http://arxiv.org/pdf/0912.4988v2
[28] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.
[29] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.
[30] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
[31] E. Candès, "The restricted isometry property and its implications for compressed sensing," *C. R. Acad. Sci. Paris, I Math.*, vol. 346, pp. 589–592, 2008.
[32] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[33] M. Stojnic, "Block-length dependent thresholds in block-sparse compressed sensing," Jul. 2009 [Online]. Available: http://arxiv.org/pdf/0907.3679
[34] A. d'Aspremont, L. E. Ghaoui, M. Jordan, and G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *Neural Inf. Process. Syst.*, vol. 17, pp. 434–448, 2004.
[35] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.
[36] B. Moghaddam, Y. Weiss, and S. Avidan, "Spectral bounds for sparse PCA: Exact & greedy algorithms," *Neural Inf. Process. Syst.*, vol. 18, 2006.
[37] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence," presented at the 23rd IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), San Francisco, CA, Jun. 13–18, 2010.
[38] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin, "Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images," Apr. 2010 [Online]. Available: http://www.ima.umn.edu/preprints/apr2010/2307.pdf, IMA Preprint
[39] K. Rosenblum, L. Zelnik-Manor, and Y. C. Eldar, "Sensing matrix optimization for block-sparse decoding," Sep. 2010 [Online]. Available: http://arxiv.org/abs/1009.1533
[40] N. Shoham and M. Elad, "Alternating KSVD-denoising for texture separation," *Proc. IEEE 25th Convention of Electr. Electron. Eng. in Israel*, 2008.
[41] P. Sprechmann, I. Ramirez, P. Cancela, and G. Sapiro, "Collaborative sources identification in mixed signals via hierarchical sparse modeling," presented at the Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Prague, Czech Republic, May 22–27, 2011.
[42] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall, 1993.

Pablo Sprechmann (S'09) received the E.E. and the M.Sc. degrees in electrical and computer engineering from the Universidad de la República, Montevideo, Uruguay (UdelaR), in 2007 and 2009, respectively. Since January 2009, he has been working towards the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Minnesota (UofM).

From 2001 to 2006, he was a Teaching Assistant with the Department of Mathematics and Statistics at UdelaR, and from 2004 to 2009, he was a Teaching and Research Assistant with the Department of Electrical Engineering at UdelaR. Since January 2009, he has been a Research Assistant at UofM. His main research interests lie in the areas of signal processing and computer vision. His current research focuses on image and audio processing and sparse linear regression.

Ignacion Ramírez (S'06) received the E.E. and the M.Sc. degrees in electrical engineering from the Universidad de la República, Uruguay (UdelaR) in 2002 and 2007, respectively. He is currently working towards the Ph.D. degree in the Scientific Computation program of the University of Minnesota (UofM).

From 1999 to 2006, he was a Teaching Assistant with the Department of Electrical Engineering at UdelaR. Since July 2008, he has been a Research Assistant with the Electrical Engineering at UofM. His main research interests are applied information theory and statistical signal processing, with focus in image processing applications. His current research focuses in automatic model selection for sparse linear models.

Guillermo Sapiro (M'94–SM'03) was born in Montevideo, Uruguay, on April 3, 1966. He received his B.Sc. (*summa cum laude*), M.Sc., and Ph.D. degrees from the Department of Electrical Engineering at the Technion—Israel Institute of Technology, Haifa, in 1989, 1991, and 1993 respectively.

After postdoctoral research at the Massachusetts Institute of Technology (MIT), Cambridge, he became Member of Technical Staff at the research facilities of HP Labs, Palo Alto, CA. He is currently with the Department of Electrical and Computer Engineering at the University of Minnesota, where he holds the position of Distinguished McKnight University Professor and Vincentine Hermes-Luh Chair in Electrical and Computer Engineering. He works on differential geometry and geometric partial differential equations, both in theory and applications in computer vision, computer graphics, medical imaging, and image analysis. He has authored and coauthored numerous papers in this area and has written the book *Geometric Partial Difference Equations and Image Analysis* (Cambridge Univ. Press, 2001).

Dr. Sapiro was awarded the Gutwirth Scholarship for Special Excellence in Graduate Studies in 1991, the Ollendorff Fellowship for Excellence in Vision and Image Understanding Work in 1992, the Rothschild Fellowship for Post-Doctoral Studies in 1993, the Office of Naval Research Young Investigator Award in 1998, the Presidential Early Career Awards for Scientist and Engineers (PECASE) in 1998, the National Science Foundation Career Award in 1999, and the National Security Science and Engineering Faculty Fellowship in 2010. He is a member of SIAM. He is the funding Editor-in-Chief of the *SIAM Journal on Imaging Sciences*. He recently co-edited a special issue of IEEE IMAGE PROCESSING in this topic and a second one in the *Journal of Visual Communication and Image Representation*.

Yonina C. Eldar (S'98–M'02–SM'07) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering both from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2002.

From January 2002 to July 2002, she was a Postdoctoral Fellow at the Digital Signal Processing Group at MIT. She is currently a Professor in the Department of Electrical Engineering at the Technion—Israel Institute of Technology, Haifa. She is also a Research Affiliate with the Research Laboratory of Electronics at MIT and a Visiting Professor at Stanford University, Stanford, CA. Her research interests are in the broad areas of statistical signal processing, sampling theory and compressed sensing, optimization methods, and their applications to biology and optics.

Dr. Eldar was in the program for outstanding students at TAU from 1992 to 1996. In 1998, she held the Rosenblith Fellowship for study in electrical engineering at MIT, and in 2000, she held an IBM Research Fellowship. From 2002 to 2005, she was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. In 2004, she was awarded the Wolf Foundation Krill Prize for Excellence in Scientific Research, in 2005 the Andre and Bella Meyer Lectureship, in 2007 the Henry Taub Prize for Excellence in Research, in 2008 the Hershel Rich Innovation Award, the Award for Women with Distinguished Contributions, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technion Outstanding Lecture Award, in 2009 the Technion's Award for Excellence in Teaching, and in 2010 the Michael Bruno Memorial Award from the Rothschild Foundation. She is a member of the IEEE Signal Processing Theory and Methods technical committee and the Bio Imaging Signal Processing technical committee, an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *EURASIP Journal of Signal Processing*, the *SIAM Journal on Matrix Analysis and Applications*, and the *SIAM Journal on Imaging Sciences*, and on the Editorial Board of *Foundations and Trends in Signal Processing*.