Compressive Source Separation: Theory and Methods for Hyperspectral Imaging

Mohammad Golbabaee* *Student Member IEEE*, Simon Arberet*, and Pierre Vandergheynst

Abstract

With the development of numbers of high resolution data acquisition systems and the global requirement to lower the energy consumption, the development of efficient sensing techniques becomes critical. Recently, Compressed Sampling (CS) techniques, which exploit the sparsity of signals, have allowed to reconstruct signal and images with less measurements than the traditional Nyquist sensing approach. However, multichannel signals like Hyperspectral images (HSI) have additional structures, like inter-channel correlations, that are not taken into account in the classical CS scheme.

In this paper we exploit the linear mixture of sources model, that is the assumption that the multichannel signal is composed of a linear combination of sources, each of them having its own spectral signature, and propose new sampling schemes exploiting this model to considerably decrease the number of measurements needed for the acquisition and source separation. Moreover, we give theoretical lower bounds on the number of measurements required to perform reconstruction of both the multichannel signal and its sources. We also proposed optimization algorithms and extensive experimentation on our target application which is HSI, and show that our approach recovers HSI with far less measurements and computational effort than traditional CS approaches.

Index Terms

Compressed sensing, source separation, hyperspectral image, linear mixture model, sparsity, proximal splitting method.

* M. G. and S. A. equally contributed to this work.

The authors are with the Signal Processing Laboratory LTS2, Electrical Engineering Department, École Polytechnique Fédérale de Lausanne (EPFL), Station 11, CH-1015 Lausanne, Switzerland. This work was supported in part by the EU FET program through projects SMALL (FET-225913) and UNLocX (FET-255931), and the Swiss National Science Foundation under grant 200021-117884.

E-mail:{mohammad.golbabaei,simon.arberet, pierre.vandergheynst}@epfl.ch.

I. INTRODUCTION

A Hyperspectral Image (HSI) is a collection of hundreds of images that have been acquired simultaneously in narrow and adjacent spectral bands, typically by airborne sensors. HSI are produced by expensive spectrometers that sample the light reflected from a two-dimensional area. An HSI data set is thus a "cube" with two spatial and one spectral dimensions. Hyperspectral imagery has many applications including environmental monitoring, agriculture planning or mineral exploration. The plurality of channels in HSI makes it possible to discriminate among the various materials that make up a geographical area: each of them is represented by a unique spectral signature. Accordingly, HSI are often processed via clustering or source separation methods to obtain segmentation maps locating and labeling the various materials appearing in the image. Unfortunately, having multiple channels comes at a price: the sheer volume of data makes acquisition, transmission, storage and analysis of HSI computationally very challenging. Therefore, the problem addressed in this paper is to reduce the complexity of manipulating HSI via a suitable compression or dimensionality reduction technique.

In this context the emerging *Compressive sensing* (CS) theory, which addresses the problem of recovering signals from few linear measurements, seems ideally suited [1], [2]. The main assumption underlying CS is that the signal is sparse or compressible when expressed in a convenient basis. A signal $x \in \mathbb{R}^n$ is said k-sparse in a basis Ψ if it is a linear combination of only k basis vectors of Ψ. The signal x is said *sparse* when $k \ll n$ and *compressible* if the coefficient's magnitudes, when sorted, have a fast power-law decay, meaning that the signal has few large coefficients and many small coefficients. The recent literature abounds with examples of sparse models for signals and images.

While the CS-community has mostly focused on 1D or 2D signals, few works have been done on higher dimensional signals, in particular multi-array signals such as HSI. Extensions of wavelets basis for 3D data have been proposed [3] and rather generic sparse models have been exploited in [4], [5] for designing innovative compressive hyperspectral imagers. However, multi-array signals such as HSI have usually some structures that go beyond the sparsity assumption. Indeed, HSI can be interpreted as a mixture of sources, each of them having a specific spectral signature. This model is widely used for unmixing HSI [6]–[10], that is extracting, form the HSI, each source and their respective spectral signatures.

The main focus of this paper is to exploit, beyond the sparsity assumption, an additional structured model, the *linear mixture* model, so as to reconstruct and separate the sources of multi-array signals assuming we know their spectra (or mixing parameters) as side information. Note that this hypothesis is

validated in many applications where the elements or materials composing the data are known and their spectra tabulated. This idea was first introduced in two of our conference papers [11], [12]. In this paper, we introduce and analyze a new sampling scheme, which exploits this structured model, and that has the following important properties:

- the number of measurements, or samples, does not scale with the number of channels,
- the recovery results do not depend on the conditioning of the mixing matrix (as long as the mixing spectra are linearly independent).

We propose new algorithms for HSI *compressive source separation* (CSS), that is source separation and data reconstruction from compressed measurements, which are based on exploiting the linear mixture structure and TV, ℓ_1 or ℓ_0 regularization. We establish that sources can be efficiently separated directly on the compressed measurements, i.e avoiding to run a source separation algorithm on this high-dimensional raw data, thereby eliminating this important bottleneck and providing a rather striking example of compressed domain data processing. We provide theoretical guaranties and intensive experiments which show that, with this approach, we can reconstruct a multi-array signal from compressed measurements with a far better accuracy than traditional CS approaches. For example, we are able to reconstruct HSI datasets with only 3% relative error from 3% of measurements and less than $1‰$ of data transmission, with an algorithm that is more than 40 times faster. While the main target application of this paper is HSI, our model and the theoretical analysis is general and could be applied to other multi-array signals like e.g. Positive Emission Tomography (PET) or distributed sensing.

The remainder of this paper is structured as follows. The necessary background and notations are first introduced in Section II. We then propose, in Section III, two acquisition schemes that exploit the prior knowledge of the mixing parameters so as to perform a decorrelation step. In Section IV, we provide theoretical guarantees for both source identification and data reconstruction. We determine the number of CS measurements sufficient for robust source identification and signal reconstruction as a function of the sparsity of the sources, sampling SNR and the conditioning of their corresponding mixture parameters. In Section V we discuss in further details the application of our acquisition and recovery schemes for HSI. We introduce different recovery algorithms that we compare with the classical methods, for various CS acquisition schemes on two sets of HSI. Finally, in the spirit of reproducible research, the code and data needed to reproduce the experimental sections of this paper is openly available at http://infoscience.epfl.ch/record/180911.

II. Background and Notations

A. CS of Multichannel Signals

We represent a multichannel signal with a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ where n_2 is the number of channels and n_1 is the dimension of signal in each channel. The CS acquisition protocol of a multichannel signal X is a linear mapping $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ of X into a CS measurement vector $y \in \mathbb{R}^m$ contaminated by the measurement noise $z \in \mathbb{R}^m$:

$$y = \mathcal{A}(X) + z.$$

When $m \ll n_1 n_2$ the signal is effectively compressed. The main goal of CS is to recover the signal X from the fewest amount of measurements m. Note that any linear mapping $\mathcal{A}(X)$ can be written in matrix form $AX_{vec} := \mathcal{A}(X)$, where $A \in \mathbb{R}^{m \times n_1 n_2}$ and $X_{vec} \in \mathbb{R}^{n_1 n_2}$ is the vectorized form of matrix $X \in \mathbb{R}^{n_1 \times n_2}$:

$$y = AX_{vec} + z. \tag{1}$$

In order to recover X_{vec}, we can search for the sparsest vector X_{vec} which is consistent with the measurement error, leading to the following ℓ_0-minimization problem:

$$\arg\min_{X_{vec}} \ \|X_{vec}\|_{\ell_0} \qquad s.t. \qquad \|y - AX_{vec}\|_2 \leq \varepsilon, \tag{2}$$

where ε is an upper bound on the norm of the noise vector (i.e. $\|z\|_2 \leq \varepsilon$), $\|\cdot\|_{\ell_0}$ denotes the ℓ_0 quasi-norm of a vector (*i.e.*, the number of its nonzero coefficients). Unfortunately, this combinatorial minimization problem is NP-hard in general [13], [14]. However, there are two tractable alternatives to solve problem (2): The convex relaxation leading to ℓ_1-minimization and greedy algorithms such as matching pursuits (MP) [13] or Iterative Hard Thresholding (IHT) [15]. Both types of approaches provide conditions on the matrix A and on the sparsity k such that the recovered solution coincides with the original signal X_{vec}, and consequently also with the solution of (2).

The ℓ_1 minimization approach consists in solving the following non-smooth convex optimization problem called Basis Pursuit DeNoising (BPDN):

$$\arg\min_{X_{vec}} \ \|X_{vec}\|_1 \qquad s.t. \qquad \|y - AX_{vec}\|_2 \leq \varepsilon, \tag{3}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm, which is equal to the sum of the absolute values of the vector entries, $\|\cdot\|_2$ denotes the ℓ_2 or Euclidean norm.

It has been shown in [1], [2], [16] that approximating the sparse recovery problem by the ℓ_1 minimization (3) can stably recover the kn_2-sparse original solution (i.e. k-sparse signal per channel) whenever A

satisfies the so-called *restricted isometry property* (RIP). This result guarantees that sparse signals can be perfectly recovered from noise-free measurements and that the recovery process is robust to the presence of noise. The computation of the isometry constants for a given matrix is prohibitive in practice, but certain classes of matrices, such as matrices with independent Gaussian or Bernoulli entries, obey the RIP condition with high probability (see Theorem 5.2 in [17]) as long as:

$$m \geq c\, n_2 k \log(n_1/k). \tag{4}$$

for a fixed constant c.

B. Sparse Regularization of a Multichannel Signal

Usually the data X_{vec} is not directly sparse, but sparse in a basis $\Psi \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$. In that case, the ℓ_1 regularization approach consists in solving the following problem which generalizes problem (3):

$$\underset{\Theta_{vec}}{\arg\min} \ \|\Theta_{vec}\|_1 \qquad s.t. \qquad \|y - A\Psi\Theta_{vec}\|_2 \leq \varepsilon, \tag{5}$$

with $X_{vec} = \Psi\Theta_{vec}$. Stable reconstruction by solving problem (5) is guaranteed as long as the $A\Psi$ matrix satisfies the RIP. When the data is a multichannel image, a classical basis is a block diagonal orthonormal basis $\Psi = \mathrm{Id}_{n_2} \otimes \Psi_{2\mathrm{D}}$ [1] where $\Psi_{2\mathrm{D}} \in \mathbb{R}^{n_1 \times n_1}$ denotes a proper 2-dimensional wavelet basis.

Another classical approach to regularize the data (specially images) is the total variation (TV) penalty [18], which tends to generate images with piecewise smooth regions and sharp boundaries. Replacing the ℓ_1 norm with the TV norm on each channel X_j of the multichannel in problem (5) leads to the Total Variation De-Noising (TVDN) problem:

$$\underset{X}{\arg\min} \ \sum_{j=1}^{n_2} \|X_j\|_{TV} \qquad s.t. \qquad \|y - AX_{vec}\|_2 \leq \varepsilon. \tag{6}$$

C. The Linear Mixture Model

One of the most practical setups of a multichannel signal is when the multichannel data matrix X is derived by a *sparse linear mixture* model as follows:

$$X = \mathbf{S}\mathbf{H}^T. \tag{7}$$

Here, $\mathbf{S} \in \mathbb{R}^{n_1 \times \rho}$ denotes the *source matrix* whose ith column contains the proportion of the source i at each pixel. Each source is mixed with the corresponding column of the *mixing matrix* $\mathbf{H} \in \mathbb{R}^{n_2 \times \rho}$ in order

[1] Id_{n_2} is the $n_2 \times n_2$ identity matrix and $\otimes$ denotes the matrix Kronecker product.

to generate the full multichannel data. Each column of $\mathbf{H}$ contains the spectrum of the corresponding source. The observed signal in any channel $j \in \{1, \ldots, n_2\}$ is thus a linear combination of ρ source signals:

$$X_j = \sum_{i=1}^{\rho} [\mathbf{H}]_{j,i} \, \mathbf{S}_i.$$

D. Mixing Parameters as Side Information for Multichannel CS Recovery

In certain multichannel signal acquisition setups the mixing parameters $\mathbf{H}$ are known at both decoder and encoder sides. In particular, this is the case in many remote sensing applications where the spectra of common materials are tabulated. Such prior efficiently restricts the degrees of freedom of the entire data matrix to the sparse coefficients of the underlying sources. Indeed, we will show that, when we know the mixing parameters $\mathbf{H}$, the inverse problem consisting in recovering the multichannel signal X from the measurements y in (1) is equivalent to the problem of recovering the sources $\mathbf{S}_{vec}$ from the following measurements:

$$y = A\Phi\mathbf{S}_{vec} + z, \tag{8}$$

with $\Phi = \mathbf{H} \otimes \mathrm{Id}_{n_1}$. The source coefficients can then be recovered by solving a convex optimization problem such as (5), where A is replaced by $A\Phi$ and the multichannel signal can be reconstructed by applying the mixing matrix to the recovered source matrix according to the linear mixture model (7). This approach has the advantage of solving two problems: i) source separation directly from the compressive measurements, ii) data compressive sampling via source separation or, equivalently, via a particular structured sparse model.

III. Compressive Multichannel Signal Acquisition Schemes

If the multichannel signal follows the linear mixture model (7), the knowledge of the mixing matrix can be used efficiently. The sparse source coefficients can be directly recovered from the measurements. In this section we introduce a decorrelation mechanism, applied at the acquisition process or as a post-processing step, which has two main advantages: first it leads to strong dimensionality reduction and secondly it improves the conditioning of the recovery problem.

A. Multichannel Recovery via Source Recovery

When we know the mixing matrix $\mathbf{H}$, and thanks to the property ($(BCD)_{vec} = (D^T \otimes B)C_{vec}$) of the Kronecker product, the sampling equation (1) (in the noise free case) can be written as:

$$AX_{vec} = A(\mathbf{S}\mathbf{H}^T)_{vec} = A\underbrace{(\mathbf{H} \otimes \mathrm{Id}_{n_1})}_{\triangleq \Phi}\mathbf{S}_{vec} = A\Phi\mathbf{S}_{vec}. \tag{9}$$

Then, the ℓ_1 regularization approach for the recovery of the whole data consists in finding the sparsest coefficients vector $\mathbf{\Theta}_{vec} \in \mathbb{R}^{\rho n_1}$ of the sources vector $\mathbf{S}_{vec} = \mathbf{\Psi}\mathbf{\Theta}_{vec}$ in a basis $\mathbf{\Psi} \in \mathbb{R}^{\rho n_1 \times \rho n_1}$, where e.g. $\mathbf{\Psi} = \mathrm{Id}_\rho \otimes \Psi_{\text{2D}}$ is a block diagonal orthonormal basis, through the following minimization:

$$\underset{\mathbf{\Theta}_{vec}}{\arg\min} \ \|\mathbf{\Theta}_{vec}\|_1 \qquad s.t. \qquad \|y - A\Phi\mathbf{\Psi}\mathbf{\Theta}_{vec}\|_2 \leq \varepsilon. \tag{10}$$

This corresponds to a "synthesis" formulation of BPDN using a basis $\mathbf{\Psi}$. The "analysis" formulation, which is equivalent to the synthesis one when $\mathbf{\Psi}$ is a basis but different when $\mathbf{\Psi}$ is a redundant dictionary, consists in solving the following problem with respect to the sources instead of its coefficients:

$$\underset{\mathbf{S}_{vec}}{\arg\min} \ \|\mathbf{\Psi}^*\mathbf{S}_{vec}\|_1 \qquad s.t. \qquad \|y - A\Phi\mathbf{S}_{vec}\|_2 \leq \varepsilon, \tag{11}$$

where $\mathbf{\Psi}^*$ is the adjoint of the operator $\mathbf{\Psi}$.

The data X can then be recovered via the mixture model $\widehat{X} = \widehat{\mathbf{S}}\mathbf{H}^T$, with $\widehat{\mathbf{S}}_{vec}$ being either the solution of the analysis problem (11) or $\widehat{\mathbf{S}}_{vec}$ being equal to $\widehat{\mathbf{S}}_{vec} = \mathbf{\Psi}\widehat{\mathbf{\Theta}}_{vec}$ with $\widehat{\mathbf{\Theta}}_{vec}$, solution of the synthesis problem (10).

B. Decorrelation Scheme

We have seen in section II-A, that the conditions to recover the signal from the noisy measurements $y = AX_{vec} + z$ depend on properties (such as RIP) of the sensing matrix A. We introduce a particular structure for the sampling matrix A which benefits from the available knowledge of the mixture parameters $\mathbf{H}$ and incorporates data decorrelation into the compressive acquisition.

1) Decorrelating Multichannel CS Acquisition: The decorrelation mechanism consists of applying the Moore-Penrose pseudo inverse matrix $\mathbf{H}^\dagger = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ in order to remove the underlying dependencies among CS measurements. We therefore propose the following sampling matrix:

$$A = \mathbf{H}^\dagger \otimes \widetilde{A}. \tag{12}$$

The main sampling matrix is generated from a smaller-size $\widehat{m} \times n_1$ *core sampling matrix* $\widetilde{A}$. Note that CS imposes $\widehat{m} \ll n_1$.

The total number of measurements is $m = \rho\,\widehat{m}$. Applying the sampling matrix A of (12) on multi-channel data results in the following CS measurements:

$$\begin{aligned} y &= A\Phi\mathbf{S}_{vec} + z \qquad (13)\\ &= \underbrace{(\mathbf{H}^{\dagger}\otimes\widetilde{A})}_{A}\underbrace{(\mathbf{H}\otimes \mathrm{Id}_{n_1})}_{\Phi}\,\mathbf{S}_{vec} + z,\\ &= \underbrace{(\mathrm{Id}_{\rho}\otimes\widetilde{A})}_{\triangleq\widetilde{A}_{\rho}}\,\mathbf{S}_{vec} + z. \qquad (14)\end{aligned}$$

The third equality comes from the following property: $(B\otimes C)(D\otimes F) = BD\otimes CF$, and $\widetilde{A}_{\rho}$ is a block diagonal matrix whose ρ diagonal blocks are populated with $\widetilde{A}$: $\widetilde{A}_{\rho}\triangleq \mathrm{Id}_{\rho}\otimes\widetilde{A}$.

As we can observe in (14) and thanks to the specific structure of the sampling matrix, the mixing parameters $\mathbf{H}$ are discarded from the formulation and each source (each column of $\mathbf{S}$) is directly subsampled by the same matrix $\widetilde{A}$.

2) Uniform Multichannel CS Acquisition: In many practical setups the acquisition scheme can not be arbitrarily chosen and is rather determined by various constraints posed by the physics of the signals and the implementation technology. Certain acquisition systems such as Rice's single-pixel hyperspectral imager [4] are using a universal random matrix to sample independently data in each channel. In this case, acquisition models such as (12), which require inter-channel interactions for compressed sampling, simply cannot be implemented. Here, the sampling matrix A in (1) is block diagonal with n_2 blocks (each applies on a certain channel) that are populated by a unique $\widehat{m}\times n_1$ matrix (similarly as $\widetilde{A}$ in (14)):

$$A = \widetilde{A}_{n_2}\triangleq \mathrm{Id}_{n_2}\otimes\widetilde{A}. \qquad (15)$$

The total number of measurements is then $m = n_2\,\widehat{m}$. Reshaping y and z correspondingly into $\widehat{m}\times n_2$ matrices Y (the measurement matrix) and Z (the noise matrix) leads to the following equivalent formulation:

$$Y = \widetilde{A}X + Z.$$

3) Decorrelation-based Uniform Sampling: A decorrelation step similar to the one introduced in Section III-B1 can be applied on the CS measurements. It consists in multiplying the rows of the measurement matrix by $(\mathbf{H}^{\dagger})^T$ and reducing the dimensionality of Y to an $\widehat{m}\times\rho$ matrix as follows:

$$\begin{aligned} Y^{*} &= Y(\mathbf{H}^{\dagger})^T\\ &= \widetilde{A}\,\mathbf{S} + Z^{*},\end{aligned}$$

where, $Z^* = Z(\mathbf{H}^\dagger)^T$. By reshaping Y^* and Z^* into the vectors y^* and z^*, we can observe that the outcome of such *decorrelation-based uniform sampling* leads to an expression similar to (14) *i.e.*,

$$y^* = \widetilde{A}_\rho \mathbf{S}_{vec} + z^*. \tag{16}$$

This decorrelating scheme favorably reduces the dimension of the data: at the acquisition stage, the total number of samples is $n_2\,\widehat{m}$ but at the transmission and decoding stages the number of samples is only $\rho\,\widehat{m} \ll n_2\,\widehat{m}$.

For the *decorrelating* sampling schemes described in section III-B1 and III-B3, the ℓ_1 minimization (e.g. the "synthesis" problem (10)) of section III-A takes the following form:

$$\underset{\mathbf{\Theta}_{vec}}{\arg\min}\ \|\mathbf{\Theta}_{vec}\|_1 \qquad s.t. \qquad \|y - \widetilde{A}_\rho\,\mathbf{\Psi}\mathbf{\Theta}_{vec}\|_2 \leq \varepsilon, \tag{17}$$

which, in the noiseless case can be decoupled into ρ independent ℓ_1 minimizations, each of them corresponding to a certain source compressed by a universal matrix $\widetilde{A}$. In Section IV we provide the theoretical analysis of such recovery scheme for various acquisition schemes.

IV. Main Theoretical Analysis

Compressive sparse source recovery is closely related to the problem of compressed sensing with *redundant dictionaries* [19], [20]. Indeed, the later problem has the same formulation as in (10) by replacing Φ by an overcomplete dictionary matrix. The first part of this section provides an overview of the CS literature on redundant dictionaries. In the second part, we derive new performance bounds that extend the former results for a larger class of dictionaries. In the third part, we cast the sparse source separation problem as a particular case of CS recovery using redundant dictionaries and we give a bound on the performance of the ℓ_1 minimization for each of the considered CS acquisition schemes (dense, uniform and decorrelated).

A. Compressed Sensing and Redundant Dictionaries

Let $x \in \mathbb{R}^n$ be a vector that is sparse in a dictionary $\mathbf{D} \in \mathbb{R}^{n\times d}$ (*i.e.*, $x = \mathbf{D}\,\theta$ with, $\theta \in \mathbb{R}^d$). The ℓ_1 minimization approach for recovering θ (equivalently x) from the compressive measurements $y = Ax + z$ consists in solving:

$$\underset{\theta}{\arg\min}\ \|\theta\|_1 \qquad s.t. \qquad \|y - A\mathbf{D}\theta\|_2 \leq \varepsilon, \tag{18}$$

where, $\|z\|_2 \leq \varepsilon$. Note that in this section A is a sampling matrix of size $m \times n$ and the dictionary $\mathbf{D}$ typically contains a large number of columns ($d \gg n$).

It has been shown in [1], [2] that the ℓ_1 minimization (18) can stably recover the original solution whenever $A\mathbf{D}$ satisfies the *restricted isometry property* (RIP). More precisely, if for all k-sparse vectors θ the following RIP property holds:

$$\Big(1-\delta_k(A\mathbf{D})\Big)\|\theta\|_2^2 \leq \|A\mathbf{D}\,\theta\|_2^2 \leq \Big(1+\delta_k(A\mathbf{D})\Big)\|\theta\|_2^2 \tag{19}$$

with the RIP constant of order k, $\delta_k(A\mathbf{D}) \leq \sqrt{2}-1$, then the solution $\widehat{\theta}$ to (18) satisfies the following error bound:

$$\|\theta-\widehat{\theta}\|_2 \leq c_0\, k^{-1/2}\,\|\theta-\theta_k\|_1 + c_1\varepsilon, \tag{20}$$

for some positive constants c_0, c_1, and where θ_k is the best k-sparse approximation of θ. Now the question is how many CS measurements are sufficient so that $A\mathbf{D}$ satisfies the RIP ? It has been shown in [19] that, for a certain class of random sampling matrices A (*e.g.*, with i.i.d. Gaussian, Bernoulli or subgaussian elements), with very high probability the RIP constant $\delta_k(A\mathbf{D})$ is bounded by:

$$\delta_k(A\mathbf{D}) \leq \delta_k(A) + \delta_k(\mathbf{D}) + \delta_k(A)\delta_k(\mathbf{D}). \tag{21}$$

If $\mathbf{D}$ is an orthonormal basis, then $\delta_k(\mathbf{D}) = 0$ and $A\mathbf{D}$ becomes another subgaussian matrix with a similar distribution as for A and thus (21) holds with equality *i.e.*, $\delta_k(A\mathbf{D}) = \delta_k(A)$.

Considering the recovery condition using ℓ_1 minimization (*i.e.*, $\delta_k(A\mathbf{D}) \leq \sqrt{2}-1$) and the bound in (21), we can conclude that A must satisfy RIP with the following constant:

$$\delta_k(A) \leq \frac{\sqrt{2}-1-\delta_k(\mathbf{D})}{1+\delta_k(\mathbf{D})}. \tag{22}$$

Moreover, using the Johnson-Lindenstrauss lemma, it has been shown that (see Theorem 5.2 in [17]) a random matrix A whose elements are drawn independently at random from Gaussian, Bernoulli or subgaussian distributions satisfies RIP as long as we have:

$$m \geq c\,k\log(n/k), \tag{23}$$

for a constant c depending on the RIP constant of A *i.e.*, the higher $\delta_k(A)$, the smaller c. If $\mathbf{D}$ is not a unitary matrix, $\delta_k(\mathbf{D})$ becomes a positive constant and the more coherent the columns of $\mathbf{D}$, the larger its RIP constant. Therefore, there is a tradeoff for compressed sensing using redundant dictionaries: redundancy can result in a more compact representations of the signals *i.e.*, smaller k, and thus less measurements are required for CS recovery using (18). Meanwhile, too much redundancy can lead to an awfully large constant in (23) implying that more CS measurements are required to overcome the uncertainties brought by over completeness.

B. Performance Bounds for Compressed Sensing using Asymmetric-RIP Dictionaries

In Section IV-C we will show that applying the classical RIP based analysis results in conditions that are too restrictive to guaranty the source recovery. Therefore in this part and in order to overcome such limitations, we derive a new theoretical performance bound that uses different notions of RIP. We begin by introducing the notions of the *asymmetric restricted isometry property* (A-RIP) and the *restricted condition number* of a dictionary $\mathbf{D}$.

Definition 1. *For a positive integer $k \in \mathbb{N}$, an $n \times d$ matrix $\mathbf{D}$ satisfies the asymmetric restricted isometry property, if for all k-sparse $x \in \mathbb{R}^d$ the following inequalities hold:*

$$\mathcal{L}_k(\mathbf{D})\|x\|_2 \leq \|\mathbf{D}x\|_2 \leq \mathcal{U}_k(\mathbf{D})\|x\|_2, \tag{24}$$

where, $\mathcal{L}_k(\mathbf{D})$ and $\mathcal{U}_k(\mathbf{D})$ are correspondingly the largest and the smallest constants for which the inequalities above hold. The restricted condition number of $\mathbf{D}$ is defined as:

$$\xi_k(\mathbf{D}) \triangleq \frac{\mathcal{U}_k(\mathbf{D})}{\mathcal{L}_k(\mathbf{D})}. \tag{25}$$

In addition, we use a different notion of RIP for the compression matrix A, namely, the *Dictionary Restricted Isometry Property* (D-RIP), proposed by Candes *et al*. in [20]:

Definition 2. *For a positive integer $k \in \mathbb{N}$, a matrix A satisfies the D-RIP adapted to a dictionary $\mathbf{D}$ as long as for all k-sparse vectors x the following inequalities hold:*

$$(1 - \delta_k^*)\|\mathbf{D}x\|_2^2 \leq \|A\mathbf{D}x\|_2^2 \leq (1 + \delta_k^*)\|\mathbf{D}x\|_2^2. \tag{26}$$

The D-RIP constant δ_k^* is the smallest constant for which the property above holds.

This definition extends the classical RIP (which deals with signals that are sparse in the canonical basis) to linear mappings that are able to stably embed all low dimensional subspaces spanned by every k columns of a redundant dictionary $\mathbf{D}$.

As in [20], we suppose that A is an $m \times n$ matrix drawn at random from certain distributions that satisfy the following concentration bound for any vector x:

$$\mathbf{Pr}\left(\left|\|Ax\|_2^2 - \|x\|_2^2\right| > t\|x\|_2^2\right) \leq C\exp\left(-c\,m\right), \tag{27}$$

for some constants C and $c > 0$ that are only depending on t. Then, A will satisfy the D-RIP for *any* $n \times d$ dictionary $\mathbf{D}$ with overwhelming probability if

$$m \gtrsim \mathcal{O}(k\log(d/k)).$$

Remark 1. *Matrices $A \in \mathbb{R}^{m\times n}$ whose elements are independently drawn at random from Gaussian, Bernoulli or (in general) subgaussian distributions satisfy the concentration bound in* (27) *and therefore satisfy D-RIP for any $n \times d$ dictionary as long as $m \gtrsim \mathcal{O}(k\log(d/k))$.*

Based on these definitions we establish the following theorem in order to bound the performance of the ℓ_1 minimization in (18):

Theorem 1. *Given a matrix A that satisfies the D-RIP adapted to a dictionary $\mathbf{D}$, with the constant $\delta^*_{\gamma k} < 1/3$ where $\gamma \geq 1 + 2\xi^2_{\gamma k}(\mathbf{D})$, then the solution $\widehat{\theta}$ to* (18) *obeys the following bound:*

$$\|\theta - \widehat{\theta}\|_2 \leq c'_0\, k^{-1/2}\, \|\theta - \theta_k\|_1 + c'_1\varepsilon, \tag{28}$$

for some positive constants c'_0, c'_1.

The proof of this theorem is given in Appendix. Using Remark 1, the following result is straightforward:

Corollary 1. *For A whose elements are drawn independently at random from Gaussian, Bernoulli or subgaussian distributions, the solution to* (18) *obeys the error bound* (28) *with an overwhelming probability and for any dictionary with a finite Restricted Condition Number $\xi_{\gamma k}(\mathbf{D})$, if*

$$m \gtrsim \gamma k\, \log\left(d/\gamma k\right). \tag{29}$$

Comparing to the bound (23) based on the classical RIP analysis, we see that (29) features the same scaling-order for the number of measurements. In addition, for both types of analysis the constant factors grow as the atoms of the dictionary become more coherent and therefore, more CS measurement are required.

Note that this result requires neither $A\mathbf{D}$ nor the dictionary $\mathbf{D}$ to satisfy the *classical* RIP. In the next section, we apply these results to guaranty the performance of the ℓ_1 minimization approach (10) for source identification and in particular, for the case where $\mathbf{H}$ is not well-conditioned.

C. Theoretical Guaranties for Source Recovery using ℓ_1 Minimization

Sparse source recovery from compressive measurements using ℓ_1 minimization (10) is a particular case of the compressed sensing problem using dictionaries (18). Indeed, for the source recovery problem, θ and the dictionary matrix $\mathbf{D}$ are replaced respectively with $\mathbf{\Theta}_{vec}$ and $\Phi' \triangleq \Phi\mathbf{\Psi} = (\mathbf{H} \otimes \mathrm{Id}_{n_1})\mathbf{\Psi}$, and consequently, $n = n_1 n_2$ and $d = \rho n_1$. The only difference here is that Φ' is a tall matrix (*i.e.*, $d \leq n$) due to its specific construction and the assumption of having few number of sources (*i.e.*, $\rho \leq n_2$).

Though there is no redundancy in Φ' in terms of the number of columns, there is uncertainty at the sparse decoder because of *coherent* columns. The following lemma which has been proven in [3] (see Lemma 2 in [3]) shows that the conditioning of Φ' is directly related to the conditioning of the underlying mixture parameters *i.e.*, intuitively, if the columns of $\mathbf{H}$ become coherent, so become the columns of Φ'.

Lemma 1. *For matrices $V_1, V_2, \ldots, V_\ell$ with restricted isometry constants $\delta_k(V_1), \delta_k(V_1), \ldots, \delta_k(V_\ell)$ respectively, we have:*

$$\delta_k(V_1 \otimes V_2 \otimes \ldots \otimes V_\ell) \leq \prod_{i=1}^{\ell} \Big(1 + \delta_k(V_i)\Big) - 1. \tag{30}$$

Since the RIP constant of any orthonormal basis is zero (*e.g.*, $\delta_k(\mathrm{Id}_{n_1}) = 0$), and since $\mathbf{\Psi}$ is an orthogonal matrix, we can deduce the following bound on the RIP constant of $\Phi' = (\mathbf{H} \otimes \mathrm{Id}_{n_1})\mathbf{\Psi}$ by applying Lemma 1:

$$\begin{aligned} \delta_k(\Phi') &= \delta_k(\Phi) \\ &\leq \delta_k(\mathbf{H}) \qquad (31) \\ &\leq \eta \triangleq \max\Big(1 - \sigma^2_{\min}(\mathbf{H}),\ \sigma^2_{\max}(\mathbf{H}) - 1\Big). \qquad (32) \end{aligned}$$

For $k \leq \rho$ one can use (31) (which then holds with equality), and more generally (32) for any k. Note that (32) follows by the definition of the RIP constant and it only holds if $\mathbf{H}$ is properly normalized so that $1 \leq \sigma_{\max}(\mathbf{H}) < 2$ and $0 < \sigma_{\min}(\mathbf{H}) \leq 1$. [2]

Moreover, due to the properties of the extreme singular values of the Kronecker product of two matrices:

$$\sigma_{\max}(V_1 \otimes V_2) = \sigma_{\max}(V_1)\, \sigma_{\max}(V_2),$$

$$\sigma_{\min}(V_1 \otimes V_2) = \sigma_{\min}(V_1)\, \sigma_{\min}(V_2),$$

and according to Definition 1, we can bound the restricted condition number of Φ' as follows:

$$\xi_k(\Phi') \leq \frac{\sigma_{\max}(\Phi')}{\sigma_{\min}(\Phi')} = \frac{\sigma_{\max}(\mathbf{H})}{\sigma_{\min}(\mathbf{H})} \triangleq \xi(\mathbf{H}), \tag{33}$$

where, $\xi(.)$ (without subscript) denotes the standard definition of the condition number of a matrix. With those descriptions, the performance of the sparse source recovery using (10) can be easily characterized by any of the previous types of performance bound of sections IV-A and IV-B.

[2] This can be done by dividing $\mathbf{H}$ and multiplying $\mathbf{S}$ by $\big(\sigma_{\max}(\mathbf{H}) + \sigma_{\min}(\mathbf{H})\big)/2$, respectively.

According to the standard definition of the RIP for the matrix Φ', we can bound its restricted condition number $\xi_k(\Phi')$ as follows:

$$\xi_k(\Phi') \leq \sqrt{\frac{1+\delta_k(\Phi')}{1-\delta_k(\Phi')}}.$$

Recall that, the classical RIP based analysis in section IV-A requires $\delta_k(\Phi') < \sqrt{2}-1$ (in order to have $\delta_k(A) > 0$ in (22)), which implies $\xi_k(\Phi') < \sqrt{\sqrt{2}+1}$, or consequently $\xi(\mathbf{H}) < \sqrt{\sqrt{2}+1}$. This severely restricts the application of such analysis for a limited class of relatively well-conditioned mixture parameters.

To address this limitation, we use the second theoretical analysis based on the D-RIP of the compression matrix presented in section IV-B. The following theorem is a corollary of Theorem 1:

Theorem 2. *Given a mixture matrix* $\mathbf{H}$ *whose condition number is* $\xi(\mathbf{H})$*, and a matrix* A *that satisfies the D-RIP adapted to* $\mathbf{H} \otimes \mathrm{Id}_{n_1}$ *with the constant* $\delta^*_{\gamma' k} < 1/3$ *where* $\gamma' = 1 + 2\xi^2(\mathbf{H})$*, then the solution* $\widehat{\mathbf{\Theta}}_{vec}$ *to* (10) *obeys the following bound for the same constants* c'_0, c'_1 *as in* (28)*:*

$$\|\mathbf{\Theta}_{vec} - \widehat{\mathbf{\Theta}}_{vec}\|_2 \leq c'_0\, k^{-1/2} \|\mathbf{\Theta}_{vec} - (\mathbf{\Theta}_{vec})_k\|_1 + c'_1 \varepsilon. \tag{34}$$

Comparing to Theorem 1, $\mathbf{D}$ is replaced by Φ' and γ is set to γ' which satisfies the requirement of Theorem 1 *i.e.*, according to (33) we have $\gamma' \geq 1 + 2\xi^2_{\gamma' k}(\mathbf{H})$. As we can see, this analysis is valid for a much wider range of condition number namely, $\xi(\mathbf{H}) \leq \sqrt{\frac{n_1 n_2/k - 1}{2}}$. [3]

Now, if we use this approximation to recover the multichannel data *i.e.*, $\widehat{X} = \widehat{\mathbf{S}}\mathbf{H}^T$, the reconstruction error can be bounded using (34) and the following inequality:

$$\begin{aligned} \|X - \widehat{X}\|_F &\leq \sigma_{\max}(\mathbf{H}) \|\mathbf{S} - \widehat{\mathbf{S}}\|_F \\ &= \sigma_{\max}(\mathbf{H}) \|\mathbf{\Theta} - \widehat{\mathbf{\Theta}}\|_F. \end{aligned} \tag{35}$$

Theorem 2 indicates $\delta^*_{\gamma' k} \leq 1/3$ as the sufficient condition for the sparse source recovery. In the following we investigate the implication of this condition for the previously mentioned acquisition schemes to bound the number of CS measurements.

[3] As for $\gamma' k \geq n_1 n_2$ an $n_1 n_2 \times n_1 n_2$ identity matrix A always satisfies $\delta^*_{\gamma' k} = 0$ (*i.e.* there is no advantage by replacing the full Nyquist sampling with CS), Theorem 2 becomes useful only when we have $\gamma' k < n_1 n_2$ which for the value of γ' in the theorem implies $\xi(\mathbf{H}) \leq \sqrt{\frac{n_1 n_2/k - 1}{2}}$.

1) Dense Random Sampling: Assume the compression matrix A that is used for subsampling data in (1) is an $m \times n_1 n_2$ matrix whose elements are drawn independently at random from the Gaussian, Bernoulli or subgaussian distributions. According to Remark 1, such matrices satisfy D-RIP adapted to Φ (with the constant $\delta^*_{\gamma' k} \leq 1/3$) provided by:

$$m \gtrsim \gamma' k \, \log(\rho n_1/\gamma' k)). \tag{36}$$

2) Uniform Random Sampling: The same type of analysis indicates a very poor performance for the uniform random acquisition scheme described in section III-B2. The corresponding sampling matrix has a block-diagonal form $A = \mathrm{Id}_{n_2} \otimes \widetilde{A}$. Here, we assume that the core compression matrix $\widetilde{A}$ that separately applies to each channel is an $\widehat{m} \times n_1$ matrix whose elements are drawn independently at random from Gaussian, Bernoulli or subgaussian distributions.

According to the theoretical analysis provided in section IV-A, the sufficient condition for source recovery via (10) is $\delta_k(A) \leq \frac{\sqrt{2}-1-\delta_k(\Phi')}{1+\delta_k(\Phi')}$ which, by considering (32) can be rephrased as:

$$\delta_k(A) \leq \frac{\sqrt{2}-1-\eta}{1+\eta}.$$

For a compression matrix with this structure and by using Lemma 1 we can deduce $\delta_k(A) \leq \delta_k(\widetilde{A})$. Now similarly as for the bound (23), $\widetilde{A}$ satisfies the RIP with the constant above (and so does A) as long as $\widehat{m} \geq c\, k \, \log(n_1/k))$ or equivalently,

$$m \geq c\, n_2 \, k \, \log(n_1/k)). \tag{37}$$

The constant c depends on the conditioning of the mixture matrix $\mathbf{H}$. When the columns of $\mathbf{H}$ are very coherent, the extreme singular values spread away from each other and η becomes large. As a consequence, $\widetilde{A}$ (or equivalently A) must satisfy RIP for a smaller constant which, as discussed earlier in section IV-A, implies c to be large and more CS measurements are required for source recovery.

3) Decorrelating Random Sampling: When a decorrelation step is incorporated into the compressive acquisition process, $\mathbf{H}$ is discarded in the recovery formulation, and then we can use the standard RIP analysis in [1], [2] to evaluate the source recovery performance. Therefore, if $A = \mathrm{Id}_\rho \otimes \widetilde{A}$ satisfies the RIP with a constant $\delta_k(A) \leq \sqrt{2}-1$, then the solution $\widehat{\mathbf{\Theta}}$ to (17) obeys the following error bound:

$$\|\mathbf{\Theta}_{vec} - \widehat{\mathbf{\Theta}}_{vec}\|_2 \leq c_0 \, k^{-1/2} \|\mathbf{\Theta}_{vec} - (\mathbf{\Theta}_{vec})_k\|_1 + c_1 \varepsilon,$$

where the constants c_0, c_1 are the same as in (20).

Now, since A is a block diagonal matrix, we can proceed along the exact same steps as for the uniform sampling scheme (Section IV-C2) to bound the minimum number of CS measurements such that

CS Acquisition Scheme	Dense	Dense	Uniform	Decorrelating
CS Recovery Approach	BPDN	SS-ℓ_1	SS-ℓ_1	SS-ℓ_1
CS measurements $m \gtrsim$	$\mathcal{O}\Big(n_2 k \log(n_1/k)\Big)$	$\mathcal{O}\Big(k \log(\rho n_1/k)\Big)$	$\mathcal{O}\Big(n_2 k \log(n_1/k)\Big)$	$\mathcal{O}\Big(k \log(\rho n_1/k)\Big)$
Constant depends on $\mathbf{H}$	-	Yes	Yes	No

TABLE I: Measurement bounds for random sampling schemes: dense, uniform and decorrelating, and for recovery approaches: BPDN and SS-ℓ_1 (*i.e*. source separation based recovery using (10) or (17)). The last row shows if the bounds for the SS-ℓ_1 are sensitive to the conditioning of the mixing matrix $\mathbf{H}$.

A satisfies the RIP:

$$\widehat{m} \geq \overline{c}\, k\, \log(n_1/k).$$

Unlike the previous measurement bounds for the non-decorrelating sampling schemes, here $\overline{c}$ is a fixed constant independent of the mixture matrix $\mathbf{H}$. Consequently, the total number of CS measurements used for source recovery is:

$$m \geq \overline{c}\, \rho\, k\, \log(n_1/k)). \tag{38}$$

Note that, for a noiseless sampling scenario ($\varepsilon = 0$) the minimization (17) can be decoupled into ρ independent ℓ_1 minimizations, each of them corresponding to a sparse recovery of a certain source. Now, if we assume that each source has exactly $k' = k/\rho$ nonzero coefficients, then a perfect recovery can be guaranteed as long as $\delta_{k'}(\widetilde{A}) \leq \sqrt{2} - 1$ which, for a matrix $\widetilde{A}$ drawn form the previously-mentioned distributions, implies that $\widehat{m} \geq \overline{c}\, k' \log(n_1/k')$ and consequently:

$$m = \rho \widehat{m} \geq \overline{c}\, k \log(\rho n_1/k). \tag{39}$$

Comparing to (38) where m is roughly proportional to ρk, here the measurement bound improves by a factor ρ and it is mainly proportional to the sparsity level k of all sources.

D. Conclusions on the Theoretical Bounds

Consider a multichannel data derived by the linear mixture (7) of ρ sources, each having a k'-sparse representation *i.e*. $\mathbf{S}$ is $k = \rho k'$ sparse. Table I summarizes the scaling-orders of the number of CS measurements sufficient for an exact data reconstruction for different noiseless random acquisition schemes and sparse recovery approaches. As we can observe, compressed sensing via source recovery using (10) once it is coupled with a proper CS acquisition (*i.e*., Dense i.i.d. subgaussian A, or a random decorrelating sampling scheme as in sections III-B1 and III-B3) leads to a significantly improved bound

compared to standard methods such as BPDN. More remarkably, the number of CS measurements turns out to be independent of the number n_2 of channels.

Finally note that the measurement bound for the source-separation-based reconstruction approach, which uses a non-decorrelating random compression matrix, depends on the conditioning of the mixture parameters via the constant factor γ' in (36). Therefore, when the columns of $\mathbf{H}$ are highly coherent, the condition number of $\mathbf{H}$ becomes relatively large, and so does γ'. This limitation can be circumvented thanks to the decorrelating acquisition scheme.

V. Applications in Compressive Hyperspectral Imagery

Compressed sensing is particularly promising for hyperspectral imagery where the acquisition procedure is very costly. This type of images can be approximated by a linear mixture model as in (7) where each spatial pixel is populated with a very few number of materials (i.e. sources). In this regard, $\mathbf{S} \in [0,1]^{n_1 \times \rho}$ is a matrix whose ρ columns are *source images* (vectorized 2D images) indicating the percentage of each material in one of the n_1 spatial pixels, and therefore

$$\sum_{j=1}^{\rho} [\mathbf{S}]_{i,j} = 1 \qquad \forall i \in \{1, \dots, n_1\}. \tag{40}$$

Moreover, $\mathbf{H} \in \mathbb{R}_+^{n_2 \times \rho}$ is a matrix whose columns contain the spectral signatures of the corresponding sources of $\mathbf{S}$. Note that in some particular applications and specially when the spatial resolution is high enough, the source images become *disjoint*, meaning that each spatial pixel contains only one material and $[\mathbf{S}]_{i,j} \in \{0,1\}$.

The two key priors that will be essential for compressive source identification are the following: i) Each source image contains piecewise smooth variations along the spatial domain, implying a sparse representation in a wavelet basis, or sparsity of its gradient, and ii) each spatial pixel is a non-negative linear combination of a *small* number of sources.

In the next two sections we introduce two classes of source separation based recovery approaches that are particularly adapted to hyperspectral compressive imagery.

A. Compressive HSI Source Separation via Convex Minimization

According to our earlier assumptions, source images are spatially piecewise smooth, which means the coefficients $\mathbf{\Theta}$ of $\mathbf{S} = \Psi_{\text{2D}}\mathbf{\Theta}$ are sparse in a 2-dimensional wavelet basis $\Psi_{\text{2D}} \in \mathbb{R}^{n_1 \times n_1}$. We conveniently rephrase this representation in a vectorized form $\mathbf{S}_{vec} = \mathbf{\Psi\Theta}_{vec}$ with $\mathbf{\Psi} = \text{Id}_\rho \otimes \Psi_{\text{2D}}$ as described in Section II-B.

Taking into account the sparsity of $\mathbf{\Theta}_{vec}$ and by incorporating specific assumptions such as (40) and non-negativity we can extend the ℓ_1 minimization approach in (10) as follows:

$$\begin{aligned} \underset{\mathbf{\Theta}}{\arg\min} \quad & \|\mathbf{\Theta}_{vec}\|_1 \\ \text{subject to} \quad & \|y - A\Phi\mathbf{\Psi}\mathbf{\Theta}_{vec}\|_2 \leq \varepsilon \\ & \Psi_{\text{2D}}\,\mathbf{\Theta}\,\mathbb{I}_\rho = \mathbb{I}_{n_1} \\ & \mathbf{\Psi}\mathbf{\Theta}_{vec} \geq 0. \end{aligned} \tag{41}$$

Where, $\mathbb{I}_n$ denotes an all one n-dimensional vector. The first constraint is the same as the fidelity constraint in (10). The last two constraints impose the element-wise non-negativity of $\mathbf{S}$ and the "percentage" normalization (40) *i.e.*, each row of $\mathbf{S}$ belongs to the positive face of the simplex in $\mathbb{R}^\rho$. Minimizing the ℓ_1 norm together with the last two constraints (that is equivalent to an additional ℓ_1 norm constraint) gives solutions that contain both desired sorts of sparsity: i) along the 2D wavelet coefficients of $\mathbf{S}$ and, ii) along each row of $\mathbf{S}$.

Note that the theoretical analysis given in Section IV-C can also apply here to bound the performance of (41). Although we bound the error similarly as for (10), one can naturally expect a much better performance for (41) thanks to the two additional constrains.

Alternatively, problem (41) can be formulated in a more general "analysis" formulation with an analysis sparsity prior $\mathcal{P}(\mathbf{S})$:

$$\begin{aligned} \underset{\mathbf{S}}{\arg\min} \quad & \mathcal{P}(\mathbf{S}) \\ \text{subject to} \quad & \|y - A\Phi\mathbf{S}_{vec}\|_2 \leq \varepsilon \\ & \mathbf{S}\,\mathbb{I}_\rho = \mathbb{I}_{n_1} \\ & \mathbf{S}_{vec} \geq 0. \end{aligned} \tag{42}$$

which is equivalent to (41) when $\mathcal{P}(\mathbf{S}) = \|\mathbf{\Psi}^*\mathbf{S}_{vec}\|_1$ and $\mathbf{\Psi}$ is a square and invertible operator. Another efficient analysis prior for image regularization is the Total Variation which can be applied on each source image of the HSI with the prior: $\mathcal{P}(\mathbf{S}) = \sum_j \|\mathbf{S}_j\|_{TV}$. The problem formulation (42) is general and includes the decorrelating schemes discussed in sections III-B1 and III-B3. Indeed inserting the matrix A of (12) in (42) leads to the following fidelity term $\|y - \widetilde{A}_\rho\,\mathbf{S}_{vec}\|_2 \leq \varepsilon$ while the other terms remain unchanged.

In the next Section we provide an iterative algorithm for solving problem (42). When sources are

disjoint, it is also possible to add a *hard thresholding* post-processing step that sets the maximum coefficient of each row of $\widehat{\mathbf{S}}$ equal to one and set to zero the other coefficients.

B. The PPXA Algorithm for Compressive Source Separation

The Parallel Proximal Splitting Algorithm (PPXA) [21] is an iterative method for minimizing an arbitrarily finite sum of lower semi-continuous (l.s.c.) convex functions. Each of the iteration consists in computing the *proximity* operator of all functions (which can be done in parallel), averaging their results and updating the solution until convergence. The proximity operator of a function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is defined as $\mathit{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ [21]:

$$\underset{\widetilde{x}\in\mathbb{R}^n}{\arg\min}\ f(\widetilde{x}) + \frac{1}{2}\|x-\widetilde{x}\|_2^2. \tag{43}$$

For solving (42) with PPXA, we rewrite it as the minimization of the sum of three l.s.c. convex functions:

$$\underset{\mathbf{S}}{\arg\min}\ \ f_1(\mathbf{S}) + f_2(\mathbf{S}) + f_3(\mathbf{S}), \tag{44}$$

with $f_1(\mathbf{S}) = \mathcal{P}(\mathbf{S})$, $f_2(\mathbf{S}) = i_{\mathcal{B}_2}(\mathbf{S})$ and $f_3(\mathbf{S}) = i_{\mathcal{B}_{\Delta+}}(\mathbf{S})$ and where $i_{\mathcal{C}}$ is the indicator function of a convex set $\mathcal{C}$ defined as:

$$i_{\mathcal{C}}(\mathbf{S}) = \begin{cases} 0 & \text{if } \mathbf{S}\in\mathcal{C} \\ +\infty & \text{otherwise,} \end{cases}$$

and the convex sets $\mathcal{B}_2, \mathcal{B}_{\Delta+} \subset \mathbb{R}^{n_1\times\rho}$ are respectively, the set of matrices that satisfy the fidelity constraint $\|y - A\Phi\mathbf{S}_{vec}\|_2 \leq \varepsilon$, and the set of matrices whose rows belong to the standard simplex in $\mathbb{R}^\rho$. The template of the PPXA algorithm that solves (44) and hence (42) is given in Algorithm 1. We now derive the proximity operator of each function f_i. Note that the definition of the proximity operator in (43) naturally extends for matrices by replacing the ℓ_2 norm with the Frobenius norm.

For $\mathcal{P}(\mathbf{S}) = \|\mathbf{\Psi}^*\mathbf{S}_{vec}\|_1$, a standard calculation shows that

$$(\mathit{prox}_{\alpha\mathcal{P}})_i = \text{sign}\big((\mathbf{\Psi}^*\mathbf{S}_{vec})_i\big)\,.\,\big(|(\mathbf{\Psi}^*\mathbf{S}_{vec})_i| - \alpha\big)_+, \tag{45}$$

which is the *soft thresholding* operator applied on the wavelet coefficients of $\mathbf{S}$. The proximity operator of $\mathcal{P}(\mathbf{S}) = \sum_{j=1}^{\rho}\|\mathbf{S}_j\|_{TV}$ can be decoupled and computed in parallel for each of the ρ sources via an efficient implementation proposed by [22]. By definition, the proximal operator of an indicator function $i_{\mathcal{C}}(\mathbf{S})$ is the orthogonal projection of $\mathbf{S}$ onto the corresponding set $\mathcal{C}$. The projection onto the standard simplex $\mathcal{B}_{\Delta+}$ can be done in one iteration using the method proposed by Duchi et al. [23]. For a general implicit operator $L \triangleq A\Phi$, the projector onto $\mathcal{B}_2$ can be computed using a *forward backward* scheme as

Algorithm 1: The Parallel Proximal Algorithm to solve (42).

Input: y, A, Φ, ε, $\beta > 0$.

Initializations:

$n = 0$, $\mathbf{S}_0 = \Gamma_{1,0} = \Gamma_{2,0} = \Gamma_{3,0} \in \mathbb{R}^{n_1 \times n_2}$

repeat

  for $(i = 1:3)$ **do**

    $P_{i,n} = prox_{3\beta f_i}(\Gamma_{i,n})$

  end

  $\mathbf{S}_{n+1} = (P_{1,n} + P_{2,n} + P_{3,n})/3$

  for $(i = 1:3)$ **do**

    $\Gamma_{i,n+1} = \Gamma_{i,n} + 2\mathbf{S}_{n+1} - \mathbf{S}_n - P_{i,n}$

  end

until *convergence*;

proposed in [24]. This projection usually has the dominant computational complexity of the algorithm because of costly sub-iterations. However if the decorrelating sampling scheme is used and $L = \widetilde{A}_\rho$ is a tight frame (*i.e.*, $\forall x \in \mathbb{R}^{\widehat{m}}\ LL^*x = \nu\, x$ for a constant ν), then according to the *semi-orthogonal linear transform* property of proximity operators [21], the orthogonal projection onto $\mathcal{B}_2$ has the following explicit form:

$$\left(prox_{\alpha f_2}(\mathbf{S})\right)_{vec} = \mathbf{S}_{vec} + \frac{1}{\nu}(\widetilde{A}_\rho)^*\mathbf{r}\left(1 - \frac{\varepsilon}{\|\mathbf{r}\|_2}\right)_+, \tag{46}$$

with $\mathbf{r} = y - \widetilde{A}_\rho \mathbf{S}_{vec}$.

C. Compressive HSI Source Separation via Iterative Hard Thresholding

If the source images are disjoint, the following non-convex minimization can be alternatively used for recovering the sparse wavelet coefficients of the sources:

$$\begin{aligned} \arg\min_{\mathbf{\Theta}} \quad & \|y - A\Phi\mathbf{\Psi}\mathbf{\Theta}_{vec}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{\Theta}_{vec}\|_0 \leq k \\ & \text{Off diag}(\mathbf{\Theta}^*\mathbf{\Theta}) = 0 \\ & \Psi_{\text{2D}}\,\mathbf{\Theta}\,\mathbb{I}_\rho = \mathbb{I}_{n_1} \\ & \mathbf{\Psi}\mathbf{\Theta}_{vec} \geq 0. \end{aligned} \tag{47}$$

Algorithm 2: The Iterative Hard Thresholding Algorithm to approximate solution of (47)

Input: y, A, Φ, $\gamma = 1/\|A\Phi\mathbf{\Psi}\|^2 = 1/\|A\Phi\|^2$ and k.

Initializations:

$n = 0$, $\mathbf{\Theta}^0 \in \mathbb{R}^{n_1 \times \rho}$

repeat

1- Gradient descent:

$$\mathbf{\Theta}_{vec}^{n+1} = \mathbf{\Theta}_{vec}^{n} - \gamma \nabla F(\mathbf{\Theta}^n)$$

2- Hard thresholding:

$$\mathbf{\Theta}_{vec}^{n+1} = \mathrm{Th}_k(\mathbf{\Theta}_{vec}^{n+1})$$

3- Orthogonal matrix procrustes:

Update Ω : $[\Omega]_{i,i} = \sqrt{n_1}\, \frac{\|\mathbf{\Theta}_{.,i}^{n+1}\|_2}{\|\mathbf{\Theta}^{n+1}\|_F}$

Singular value decomposition: $U\Sigma V^* = \mathbf{\Theta}^{n+1}\Omega$

$$\mathbf{\Theta}^{n+1} = UV^*\Omega$$

4- Simplex projection:

$$\mathbf{\Theta}^{n+1} = \Psi_{\mathrm{2D}}^* \ \mathrm{Project}_{\mathcal{B}_{\Delta_+}}(\Psi_{\mathrm{2D}}\mathbf{\Theta}^{n+1})$$

until *convergence*;

where the operator Off diag(B) returns the off-diagonal elements of matrix B, and the ℓ_0 norm constraint on $\mathbf{\Theta}_{vec}$ imposes the wavelet coefficients to be k-sparse. The second constraint imposes the orthogonality of the wavelet coefficients which is a consequence of the source disjointness. The two last constraints are the same as in (41).

Despite its convex objective term, (47) has multiple non-convex constraints and is therefore a non-convex problem. We propose an algorithm similar to the *Iterative Hard Thresholding* (IHT) algorithm [15] to approximate the solution of (47). At each iteration the current solution is updated by a gradient descent step followed by a hard thresholding step $\mathrm{Th}_k(\cdot)$ that selects the k largest wavelet coefficients of $\widehat{\mathbf{\Theta}}_{vec}$. In addition the three last constraints of (47) are applied sequentially:

- First, a procedure inspired by the *orthogonal matrix procrustes* is applied to diagonalize $\widehat{\mathbf{\Theta}}^*\widehat{\mathbf{\Theta}}$. Let Ω be a $\rho \times \rho$ diagonal matrix where for $1 \leq i \leq \rho$ we have

$$[\Omega]_{i,i} = \sqrt{n_1}\, \frac{\|\widehat{\mathbf{\Theta}}_{.,i}\|_2}{\|\widehat{\mathbf{\Theta}}\|_F}.$$

  Since for disjoint sources we have $\|\mathbf{S}\|_F = \|\mathbf{\Theta}\|_F = \sqrt{n_1}$, then a good orthogonal matrix that would approximate $\widehat{\mathbf{\Theta}}$ and keeps the energy of the current estimate of each source image proportional to

that of the previous estimate would be $UV^*\Omega$ through the following singular value decomposition $U\Sigma V^* = \widehat{\mathbf{\Theta}}\Omega$.

- Second, the current solution $\widehat{\mathbf{S}} = \Psi_{2\mathrm{D}}\widehat{\mathbf{\Theta}}$ is projected onto the standard simplex as in [23].

The description of the this algorithm can be found in Algorithm 2. Note that the gradient of the objective functional $F(\mathbf{\Theta}) = \|y - A\Phi\mathbf{\Psi}\mathbf{\Theta}_{vec}\|_2^2$ is:

$$\nabla F(\mathbf{\Theta}) = -(A\Phi\mathbf{\Psi})^*\big(y - A\Phi\mathbf{\Psi}\mathbf{\Theta}_{vec}\big). \tag{48}$$

Using the decorrelating scheme, the objective function in (47) becomes $F(\mathbf{\Theta}) = \|y - \widetilde{A}_\rho\mathbf{\Psi}\mathbf{\Theta}_{vec}\|_2^2$ with gradient :

$$\nabla F(\mathbf{\Theta}) = -(\widetilde{A}_\rho\mathbf{\Psi})^*\big(y - \widetilde{A}_\rho\mathbf{\Psi}\mathbf{\Theta}_{vec}\big). \tag{49}$$

The rest of Algorithm 2 remains unchanged.

In the next section, we evaluate the performances of these algorithms on HSI.

VI. Experiments

In this section, we evaluate the ability of the methods presented in Section V, (called "*SS methods*" and summed up in table II) to separate the sources and recover HSI in various scenarios: various noise levels (from noiseless to 10 dB SNR), various sampling ratios (from $m/(n_1n_2) = 1/4$ to $1/32$ sampling rates), various sampling mechanisms (uniform and dense sampling), on two different HSI (Geneva and Urban). We also compare the *SS methods* with the classical methods for CS, such as the BPDN problem (5) `BPDN`, the TVDN problem (6) `TVDN`, both solved with a Douglas-Rachford (DR) splitting algorithm.

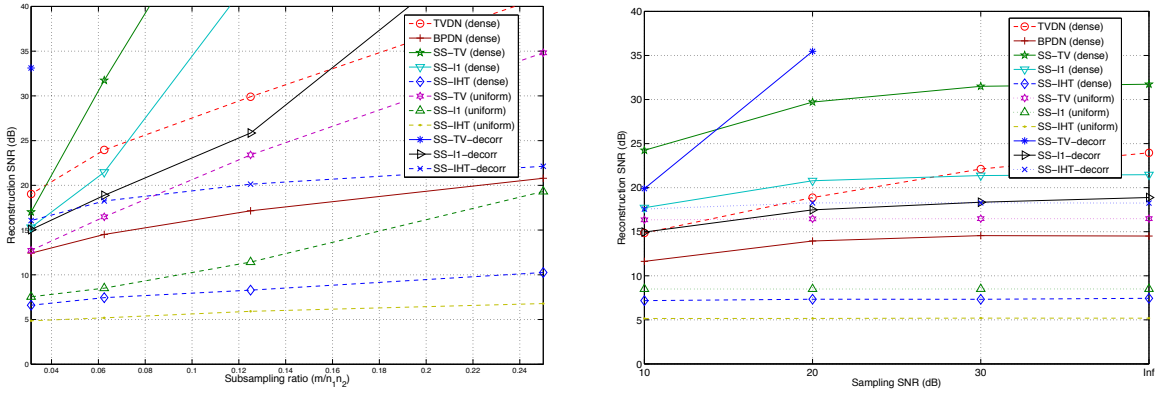
A. Sampling Mechanism

We used two different sampling schemes: i) the sensing matrix A is *dense* (and the methods implementing the decorrelation step cannot be applied), and ii) *uniform* sampling where the sensing matrix is block diagonal with identical blocks as in (15). In the latter, the decorrelation step can be applied as explained in section III-B.

So as to generate the random sampling matrices A and $\widetilde{A}$ that can be used in practical applications, we used the Random Convolution (RC) measurement scheme proposed by Romberg [25] that convolves the image with a random pattern using few optical blocks. More remarkably, sampling matrices generated by RC are tight frames and thus for decorrelating schemes, they benefit from a closed form expression (46) for computing $prox_{\alpha f_2}(\cdot)$ that can massively accelerates the recovery procedure.

TABLE II: Description of the proposed *SS methods*.

Method name	Description
`SS-IHT`	Problem (47) solved with Algorithm 2 with gradient $\nabla F(\boldsymbol{\Theta})$ of Eq. (48).
`SS-l1`	Problem (42) solved with Algorithm 1, with $\mathcal{P}(\mathbf{S}) = \|\boldsymbol{\Psi}^* \mathbf{S}_{vec}\|_1$ and $prox_{\alpha f_2}(\cdot)$ computed using a forward-backward scheme as proposed in [24].
`SS-TV`	Problem (42) solved with Algorithm 1, with $\mathcal{P}(\mathbf{S}) = \sum_{j=1}^{\rho} \|\mathbf{S}_j\|_{TV}$ and $prox_{\alpha f_2}(\cdot)$ computed using a forward-backward scheme as proposed in [24].
`SS-IHT-decorr`	Problem (47) solved with Algorithm 2 with gradient $\nabla F(\boldsymbol{\Theta})$ of Eq. (49).
`SS-l1-decorr`	Problem (42) solved with Algorithm 1, with $\mathcal{P}(\mathbf{S}) = \|\boldsymbol{\Psi}^* \mathbf{S}_{vec}\|_1$, and $prox_{\alpha f_2}(\cdot)$ computed with the closed form Eq. (46).
`SS-TV-decorr`	Problem (42) solved with Algorithm 1, with $\mathcal{P}(\mathbf{S}) = \sum_{j=1}^{\rho} \|\mathbf{S}_j\|_{TV}$ and $prox_{\alpha f_2}(\cdot)$ computed with (46).



(a) Reconstruction SNR vs. subsampling ratio (noiseless sampling)

(b) Reconstruction SNR vs. sampling SNR (subsampling ratio:$1/16$)

Fig. 1: Geneva HSI reconstruction performance for different sampling mechanisms and recovery methods. Points with ∞ reconstruction SNR (exact recovery) are not plotted.

B. The Geneva HSI

We evaluate the different methods, for different sampling rates (Fig. 1(a)), and different noise levels (Fig. 1(b)), on a HSI generated from a ground truth map image [4] of farms in a suburb of Geneva. The source spectra (i.e. columns of $\mathbf{H}$) are chosen form the USGS digital spectral library [26].The HSI cube has spatial slices of the resolution $N = 256 \times 256$ that are taken over $J = 224$ frequency bands.

[4]We acknowledge Xavier Gigandet and Meritxell Bach Cuadra for providing this ground truth map.

TABLE III: Source separation performance (Accuracy) of *SS methods*. Methods with the highest accuracy are highlighted in each column.

Noise SNR	$+\infty$ dB				30 dB				10 dB			
Sampling rate	1/4	1/8	1/16	1/32	1/4	1/8	1/16	1/32	1/4	1/8	1/16	1/32
SS-IHT*(dense sampling)*	0.69	0.61	0.57	0.48	0.71	0.6	0.57	0.48	0.7	0.6	0.57	0.48
SS-l1*(dense sampling)*	**1.0**	**1.0**	0.95	0.81	**1.0**	**1.0**	0.95	0.8	**1.0**	0.98	0.91	0.73
SS-TV*(dense sampling)*	**1.0**	**1.0**	**1.0**	0.92	**1.0**	**1.0**	**1.0**	0.91	**1.0**	**1.0**	**0.98**	0.88
SS-IHT*(uniform sampling)*	0.43	0.38	0.31	0.25	0.43	0.37	0.31	0.26	0.43	0.37	0.3	0.26
SS-l1*(uniform sampling)*	0.97	0.73	0.45	0.31	0.95	0.73	0.48	0.3	0.96	0.75	0.42	0.3
SS-TV*(uniform sampling)*	**1.0**	0.98	0.9	0.76	**1.0**	0.97	0.89	0.74	**1.0**	0.97	0.88	0.74
SS-IHT-decorr	0.98	0.98	0.96	0.94	0.99	0.98	0.96	0.94	0.98	0.97	0.95	0.92
SS-l1-decorr	**1.0**	0.99	0.97	0.92	**1.0**	0.99	0.96	0.91	0.98	0.95	0.92	0.87
SS-TV-decorr	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.99	**0.98**	**0.96**

1) Performance of the SS methods: Concerning the performance of the *SS methods*, we observe in Fig. 1 that:

- The dense sampling scheme is always better than the uniform sampling scheme.
- The decorrelated scheme is always better than dense sampling for TV-based and IHT methods, but is not for the ℓ_1-based method.
- The decorrelating method SS-TV-decorr results in perfect reconstruction in the cases where the sampling ratio is higher or equal to $1/16$ and performs better than all the other methods in all regimes, except in high noise of 10 dB SNR, where the dense approach SS-TV (*dense sampling*) performs slightly better.

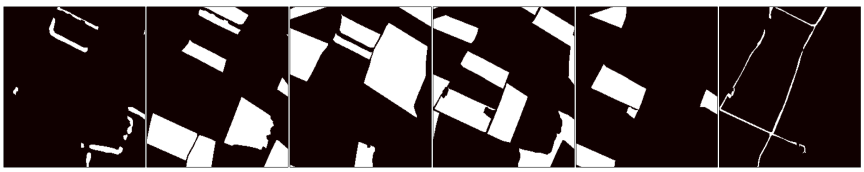
2) Comparison with Classical CS Methods: We observed that SS-TV-decorr always obtained significantly better results than the classical CS methods in all regimes.

3) Source Reconstruction: We reported in Tab. III the source separation performance of the *SS methods*. Since source images are disjoint, the quality was measured by the source recovery *accuracy* indicating the percentage of correctly classified pixels in the spatial domain. The method SS-TV-decorr, based on TV regularization and decorrelation, which achieved the best performance for HSI reconstruction also obtain the best performance for source separation. Figure 2 illustrates the reconstructed sources of different *SS methods* for various sampling schemes (dense, uniform, decorrelating).

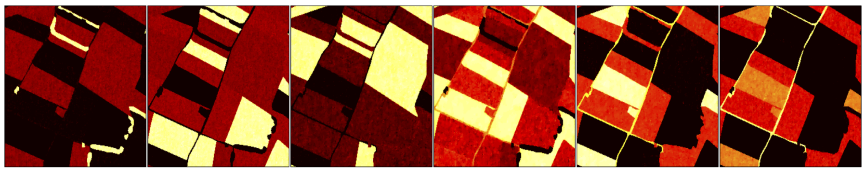
C. The Urban HSI

In order to evaluate different approaches on a real HSI, we consider the Urban HSI of size $256 \times 256 \times 171$ which was obtained from the site [27] of the US Army Topographic Engineering Center.

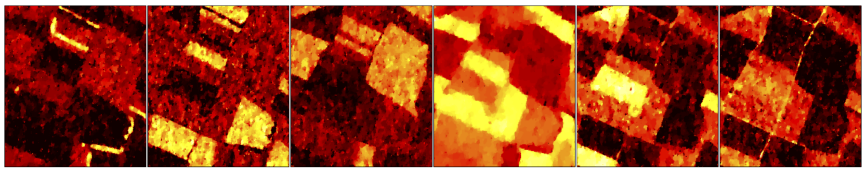
As the ground truth of this image (*i.e.*, the true source images and their corresponding spectral



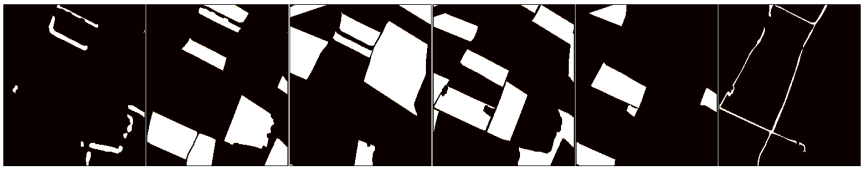
(a) True sources



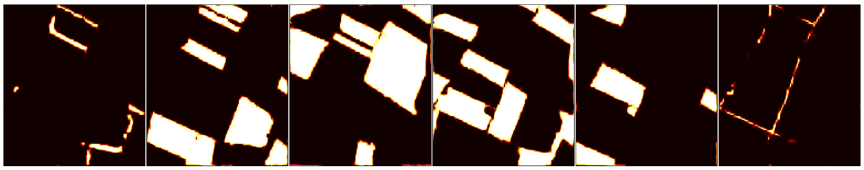
(b) `SS-TV` *(dense sampling)*



(c) `SS-TV` *(uniform non-decorrelating sampling)*



(d) `SS-TV-decorr` *(uniform decorrelating sampling)*



(e) `SS-IHT-decorr` *(uniform decorrelating sampling)*

Fig. 2: Estimated source images of Geneva HSI for different sampling schemes and recovery methods (subsampling ratio: 1/16, noiseless sampling).

signatures) is not available, we first separate the underlying sources using a *blind* source separation algorithm for fully-sampled HSI [10] and later, use these separated sources, depicted in Fig. 3(a), as a reference. Figure 3 demonstrates the reconstructed sources of Urban using our proposed SS approaches based on convex minimization, for different noiseless sampling mechanisms (dense, uniform, uniform-

decorrelating) and for a fixed subsampling ratio.[5] Moreover, Figure 4 shows the reconstructed Urban HSI for a certain spectral band, using the source images estimated by the SS methods based on TV minimization (*i.e.*, `SS-TV` and `SS-TV-decorr`).

a) Results: Similar to our previous experiment, we observe that for a uniform (non-decorrelating) sampling scheme `SS-TV` has very poor recovery performance. Meanwhile, adding a decorrelation step results in a significant improvement in source recovery. As we can see in Figure 3(b), the estimated source images using `SS-TV` for a dense sampling scheme have better spatial resolutions, but are not as well separated as with the `SS-TV-decorr` method.

b) Computational Performance: Decorrelation step massively decreases the computational complexity. `SS-TV-decorr` performs within 20 minutes whereas `SS-TV` for a dense sampling scheme requires more than 80 hours of computations! The classical `BPDN` and `TVDN` methods take between 3 to 14 hours, as the corresponding ℓ_1 or TV minimization runs over a large number of channels (rather than few underlying sources). We ran all the codes on a Mac Pro 2.26 GHz Intel CPU, 16 GB RAM computer.

D. Conclusion on the Experiments

The decorrelation step is of great benefit, and the proposed method `SS-TV-decorr`, based on TV regularization and decorrelation, outperforms significantly all the other methods for HSI reconstruction and source estimation for all tested SNRs and sampling rates. Moreover `SS-TV-decorr` is clearly the fastest method and is more than 40 times faster than the classical `TVDN`.

While finalizing this work we became aware of a recent paper [28] that proposes a source recovery approach similar to (42), albeit for the particular case of uniform sampling and TV regularization. The authors also use a "SVD preprocessing" step for dimensionality reduction and denoising that, contrary to our decorrelation step, does not cancel the effects of the conditioning of the mixing matrix. Comparing the fourth source image in Figure 3(e), corresponding to the "roads", with the similar source recovered in Figure 6 in [28] indicates that `SS-TV-decorr` achieves similar or better separation performance with only half the measurements rate. Additionally we provide a theoretical analysis for the compressive source separation problem, considering various sampling schemes and multiple recovery methods.

VII. Conclusion

In this paper, we exploited a linear mixture of sources model into a Compressed Sensing (CS) scheme for multichannel signal acquisition and source separation with a particular focus on hyperspectral images

[5] As the source images of Urban are not spatially disjoint, we do not apply Algo. 2.

(HSI). We study three different acquisition schemes (dense, uniform and decorrelated) theoretically and experimentally, and showed that the decorrelating scheme enhances drastically the recovery of the spectral data and its sources. Indeed, our theoretical analysis showed that, using this scheme, and contrary to the traditional CS approach, the number of measurements does not scale with the number of channels and does not depend on the conditioning of the mixing matrix, as long as the mixed spectra are linearly independent. This leads to a strong reduction in the number of needed measurements for a given reconstruction error. We also provided algorithms that reconstruct the multichannel signal (more particularly HSI) and its sources, by exploiting both sparsity of the signal at each channel and the correlation of the signals along the channels. We provided experiments on HSI and showed that we can reconstruct both the HSI and its sources with far fewer measurements and less computational effort than traditional CS approaches. Finally, we showed that it is possible to accurately recover the sources directly from the compressed measurements, avoiding to run a source separation algorithm on the high-dimensional raw data.

Extension of this work includes dealing with non-linear mixture of sources as well as dealing with the difficult problem of recovering the sources and the mixing system from the compressed measurements.

VIII. Appendix

A. Proof of Theorem 1

Let $\theta \in \mathbb{R}^d$ be the original vector we aim to recover from its CS measurements y with $y = A\mathbf{D}\theta + z$ and $\|z\|_2 \leq \varepsilon$, and let $\widehat{\theta}$ be the solution of the ℓ_1 minimization (18). The reconstruction error is denoted $h = \widehat{\theta} - \theta$. Let $\mathcal{T}_0 \subseteq \{0, \ldots, d\}$ be the set that contains the indices of the k coefficients of θ having the largest magnitudes and, $\mathcal{T}_0^c$ the complement set of $\mathcal{T}_0$. Let $\theta_{\mathcal{T}}$ denote a vector of the same size as θ whose elements indexed by the set $\mathcal{T}$ are identical to that of θ and zero elsewhere.

Minimizing the ℓ_1 norm in (18) implies

$$\begin{aligned} \|\theta\|_1 &\geq \|\theta + h\|_1 \\ &= \|\theta_{\mathcal{T}_0} + h_{\mathcal{T}_0}\|_1 + \|\theta_{\mathcal{T}_0^c} + h_{\mathcal{T}_0^c}\|_1 \\ &\geq \|\theta_{\mathcal{T}_0}\|_1 - \|h_{\mathcal{T}_0}\|_1 - \|\theta_{\mathcal{T}_0^c}\|_1 + \|h_{\mathcal{T}_0^c}\|_1, \end{aligned}$$

and therefore,

$$\|h_{\mathcal{T}_0^c}\|_1 \leq \|h_{\mathcal{T}_0}\|_1 + 2\|\theta_{\mathcal{T}_0^c}\|_1. \tag{50}$$

Let $\mathcal{T}_1$ be the set that contains the indices of the τk coefficients of $\theta_{\mathcal{T}_0^c}$ having the largest magnitudes, $\mathcal{T}_2$ the set containing the indices of the second τk largest coefficients of $\theta_{\mathcal{T}_0^c}$, and so on. With this

decomposition, $\forall j \geq 2$ we have:

$$\|h_{\mathcal{T}_j}\|_2 \leq (\tau k)^{-1/2} \|h_{\mathcal{T}_{j-1}}\|_1,$$

and thus,

$$\sum_{j\geq 2} \|h_{\mathcal{T}_j}\|_2 \leq (\tau k)^{-1/2} \|h_{\mathcal{T}_0^c}\|_1.$$

Now according to (50) and since $h_{\mathcal{T}_0}$ is k-sparse we have

$$\sum_{j\geq 2} \|h_{\mathcal{T}_j}\|_2 \leq \tau^{-1/2}\|h_{\mathcal{T}_0}\|_2 + 2(\tau k)^{-1/2}\|\theta_{\mathcal{T}_0^c}\|_1. \tag{51}$$

On the other hand, since both θ and $\widehat{\theta}$ satisfy the fidelity constraint of (18), we have

$$\|A\mathbf{D}h\|_2 \leq \|y - A\mathbf{D}\theta\|_2 + \|y - A\mathbf{D}\widehat{\theta}\|_2 \leq 2\varepsilon.$$

Let's define $\mathcal{T}_{01} := \mathcal{T}_0 \cup \mathcal{T}_1$ and $\gamma = \tau + 1$. According to the last inequality we can write

$$\begin{aligned}
2\varepsilon &\geq \|A\mathbf{D}h\|_2 \\
&\geq \|A\mathbf{D}h_{\mathcal{T}_{01}}\|_2 - \sum_{j\geq 2} \|A\mathbf{D}h_{\mathcal{T}_j}\|_2 \\
&\geq \sqrt{1-\delta^*_{\gamma k}}\,\|\mathbf{D}h_{\mathcal{T}_{01}}\|_2 - \sqrt{1+\delta^*_{\tau k}} \sum_{j\geq 2} \|\mathbf{D}h_{\mathcal{T}_j}\|_2 \\
&\geq \mathcal{L}_{\gamma k}(\mathbf{D})\sqrt{1-\delta^*_{\gamma k}}\,\|h_{\mathcal{T}_{01}}\|_2 - \mathcal{U}_{\tau k}(\mathbf{D})\sqrt{1+\delta^*_{\tau k}} \sum_{j\geq 2} \|h_{\mathcal{T}_j}\|_2 \\
&\geq \mathcal{L}_{\gamma k}(\mathbf{D})\sqrt{1-\delta^*_{\gamma k}}\|h_{\mathcal{T}_{01}}\|_2 - \mathcal{U}_{\tau k}(\mathbf{D})\sqrt{1+\delta^*_{\tau k}}\Big(\tau^{-1/2}\|h_{\mathcal{T}_0}\|_2 + 2(\tau k)^{-1/2}\|\theta_{\mathcal{T}_0^c}\|_1\Big).
\end{aligned}$$

The third inequality follows from definition of the D-RIP (see Definition 2) which holds for the matrix A, together with the fact that $h_{\mathcal{T}_{01}}$ and $h_{\mathcal{T}_j}$ ($\forall j \geq 2$) are respectively γk and τk sparse. The fourth inequality follows from the definition of the A-RIP that holds for matrix $\mathbf{D}$ (see Definition 1), and finally the last inequality uses (51). We apply the bounds $\delta^*_{\tau k} \leq \delta^*_{\gamma k}$, $\mathcal{U}_{\tau k}(\mathbf{D}) \leq \mathcal{U}_{\gamma k}(\mathbf{D})$ and $\|h_{\mathcal{T}_0}\|_2 \leq \|h_{\mathcal{T}_{01}}\|_2$ in the last inequality and we deduce the following bound:

$$\|h_{\mathcal{T}_{01}}\|_2 \leq \alpha k^{-1/2}\|\theta_{\mathcal{T}_0^c}\|_1 + \beta\varepsilon, \tag{52}$$

where the constants α, β are $\alpha = \frac{2}{\xi^{-1}_{\gamma k}(\mathbf{D})\sqrt{\tau\left(\frac{1-\delta^*_{\gamma k}}{1+\delta^*_{\gamma k}}\right)}-1}$, and $\beta = \frac{2\mathcal{U}_{\gamma k}(\mathbf{D})\sqrt{\tau(1+\delta^*_{\gamma k})}}{\xi^{-1}_{\gamma k}(\mathbf{D})\sqrt{\tau\left(\frac{1-\delta^*_{\gamma k}}{1+\delta^*_{\gamma k}}\right)}-1}$.
Now if we set $\tau \geq 2\xi^2_{\gamma k}(\mathbf{D})$ (equivalently, $\gamma \geq 1 + 2\xi^2_{\gamma k}(\mathbf{D})$), it is sufficient to have $\delta^*_{\gamma k} < 1/3$ so that α and β remain positive. Finally we conclude the proof of Theorem 1 by using the inequalities (51) and

(52) to bound the whole error term as follows:

$$\begin{aligned}\|h\|_2 &\leq \|h_{\mathcal{T}_{01}}\|_2 + \sum_{j\geq 2}\|h_{\mathcal{T}_j}\|_2 \\ &\leq (1+\tau^{-1/2})\|h_{\mathcal{T}_{01}}\|_2 + 2(\tau k)^{-1/2}\|\theta_{\mathcal{T}_0^c}\|_1 \\ &\leq c_0' k^{-1/2}\|\theta_{\mathcal{T}_0^c}\|_1 + c_1'\varepsilon,\end{aligned}$$

where, the constants of the error bound are $c_0' = \alpha + (2+\alpha)\tau^{-1/2}$ and $c_1' = \beta(1+\tau^{-1/2})$.

REFERENCES

[1] D. Donoho, "Compressed sensing," *IEEE Trans. on Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] E. J. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements." *Pure Appl. Math.*, vol. 59, pp. 1207–1223, 2005.

[3] M. Duarte and R. Baraniuk, "Kronecker Compressive Sensing," *to appear in the IEEE Trans. on Image Processing*, 2009.

[4] T. Sun and K. Kelly, "Compressive sensing hyperspectral imager," *Comp. Optical Sensing and Imaging (COSI), San Jose, CA, Oct. 2009.*

[5] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Applied Optics*, vol. 47, pp. B44–B51, 2008.

[6] N. Keshava and J. Mustard, "Spectral unmixing," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 44–57, 2002.

[7] J. Nascimento and J. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 4, pp. 898–910, 2005.

[8] J. Wang and C.-I. Chang, "Applications of independent component analysis in endmember extraction and abundance quantification for hyperspectral imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 44, no. 9, sept. 2006.

[9] H. Ren and C.-I. Chang, "Automatic spectral target recognition in hyperspectral imagery," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 39, no. 4, pp. 1232 – 1249, oct. 2003.

[10] S. Arberet, "Hyper-demix: Blind source separation of hyperspectral images using local ml estimates," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1393–1396.

[11] M. Golbabaee, S. Arberet, and P. Vandergheynst, "Multichannel compressed sensing via source separation for hyperspectral images," in *Eusipco*, 2010.

[12] ——, "Distributed compressed sensing of hyperspectral images via blind source separation," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*. IEEE, 2010.

[13] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[14] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, p. 227, 1995.

[15] T. Blumensath and M. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 629–654, 2008.

[16] E. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[17] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, pp. 253–263, 2008.

[18] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, pp. 259 – 268, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016727899290242F

[19] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *Information Theory, IEEE Transactions on*, vol. 54, no. 5, pp. 2210 –2219, may 2008.

[20] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59 – 73, 2011.

[21] P. L. Combettes and J. C. Pesquet, "Proximal splitting methods in signal processing," *in: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer-Verlag*, vol. 49, pp. 185–212, 2011.

[22] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, pp. 89–97, 2004.

[23] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l1-ball for learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*, ser. ICML '08, 2008, pp. 272–279.

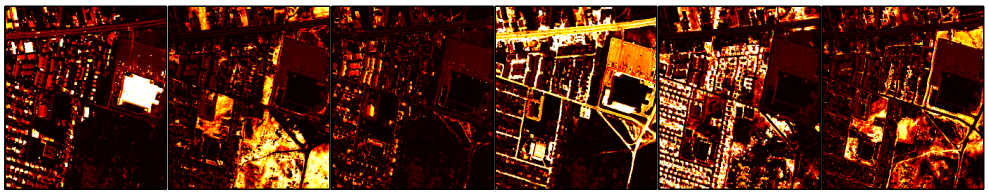
[24] M. Fadili and J. Starck, "Monotone operator splitting for optimization problems in sparse recovery," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 1461–1464.

[25] J. Romberg, "Compressive sensing by random convolution," *SIAM J. Imaging Sciences*, 2009.

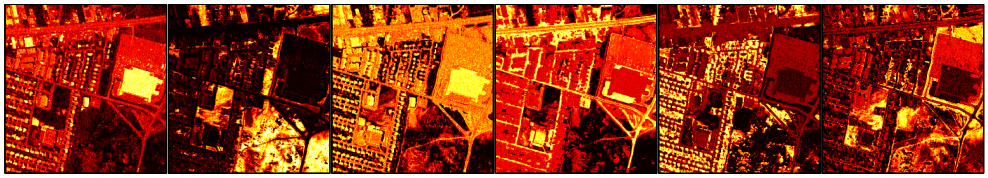
[26] [Online]. Available: http://speclab.cr.usgs.gov/spectral.lib06

[27] [Online]. Available: http://www.agc.army.mil/hypercube/

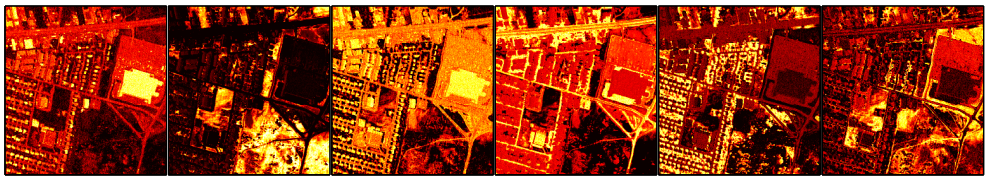
[28] C. Li, T. Sun, K. Kelly, and Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing," *Image Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 1200 –1210, march 2012.



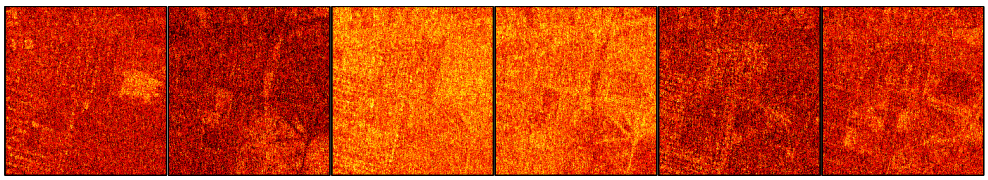
(a) Reference: Sources estimated with a BSS algorithm.



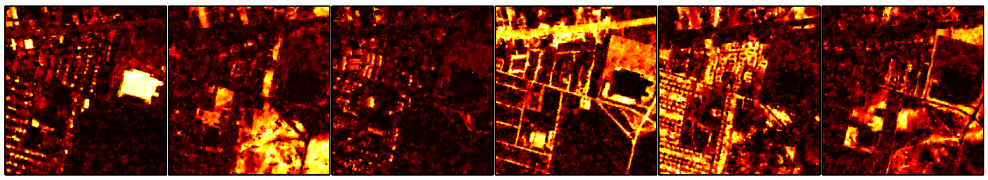
(b) `SS-TV` *(dense sampling)*, source reconstruction SNR: 6.34 dB



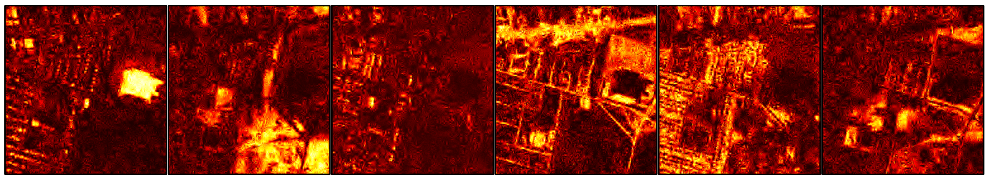
(c) `SS-l1` *(dense sampling)*, source reconstruction SNR: 6.29 dB



(d) `SS-TV` *(uniform non-decorrelating sampling)*, source reconstruction SNR: 1.88 dB



(e) `SS-TV-decorr` *(uniform decorrelating sampling)*, source reconstruction SNR: 8.64 dB



(f) `SS-l1-decorr` *(uniform decorrelating sampling)*, source reconstruction SNR: 5.65 dB

Fig. 3: Estimated source images of Urban HSI using different recovery methods (*i.e.*, TV or wavelet ℓ_1 minimization), and for different sampling mechanisms (subsampling ratio: 1/8, noiseless sampling).

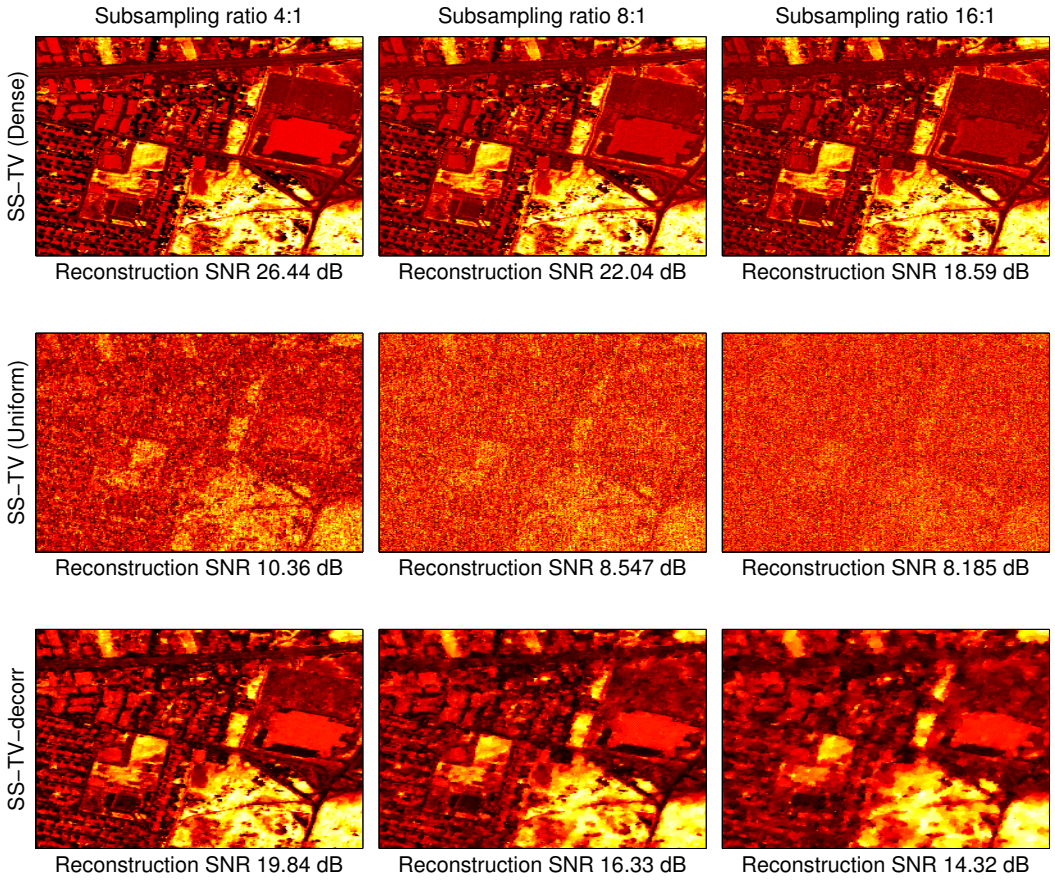


Fig. 4: Reconstructed Urban HSI at spectral band 33, using SS methods based on TV minimization, for various sampling mechanisms (Dense, Uniform non-decorrelating, Uniform decorrelating) and subsampling ratios.