Compressed Sensing of Simultaneous Low-Rank and Joint-Sparse Matrices

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

Abstract

In this paper we consider the problem of recovering a high dimensional data matrix from a set of incomplete and noisy linear measurements. We introduce a new model that can efficiently restrict the degrees of freedom of the problem and is generic enough to find a lot of applications, for instance in multichannel signal compressed sensing (*e.g.* sensor networks, hyperspectral imaging) and compressive sparse principal component analysis (s-PCA). We assume data matrices have a *simultaneous low-rank and joint sparse* structure, and we propose a novel approach for efficient compressed sensing (CS) of such data. Our CS recovery approach is based on a convex minimization problem that incorporates this restrictive structure by jointly regularizing the solutions with their nuclear (trace) norm and ℓ_2/ℓ_1 mixed norm. Our theoretical analysis uses a new notion of restricted isometry property (RIP) and shows that, for sampling schemes satisfying RIP, our approach can stably recover all low-rank and joint-sparse matrices. For a certain class of random sampling schemes satisfying a particular concentration bound (*e.g.* the subgaussian ensembles) we derive a lower bound on the number of CS measurements indicating the near-optimality of our recovery approach as well as a significant enhancement compared to the state-of-the-art. We introduce an iterative algorithm based on proximal calculus in order to solve the joint nuclear and ℓ_2/ℓ_1 norms minimization problem and, finally, we illustrate the empirical recovery phase transition of this approach by series of numerical experiments.

Index Terms

Compressed sensing, joint-sparsity, low-rank matrices, ℓ_2/ℓ_1 mixed norm, nuclear (trace) norm, convex optimization, proximal splitting method.

I. INTRODUCTION

Suppose we are given an $n_1 \times n_2$ matrix X which is *joint-sparse*, that is only $k \ll n_1$ rows of this matrix contain nonzero entries (a similar argument holds for joint-column-sparsity), and in the meantime the underlying submatrix of nonzero rows (and equivalently the whole matrix X) has a *low rank $r \ll$* $\min(k, n_2)$. Such matrices have few degrees of freedom: if one knows the indices of those k nonzero rows, the corresponding submatrix will only have $(k + n_2 - r)r$ degrees of freedom. Following the recent trend in the areas of compressed sensing [1]–[3], sparse approximation [4], [5] and low-rank matrix recovery [6]–[9], we would like to know how many non-adaptive linear measurements of this matrix are sufficient to robustly represent the whole data and, more importantly, how this compares to the state-of-the-art. The following questions are therefore natural:

- What are the efficient information/structure-preserving measurements?
- Is there any computationally tractable algorithm that incorporates the simultaneous low-rank and joint-sparse structures into data recovery and does it perform better than the state-of-the-art?
- And finally, how many measurements do we need to recover an exact low-rank and sparse matrix, and is the recovery algorithm *stable* with respect to noisy measurements and matrices that are approximately low-rank or not exactly joint-sparse but *compressible*?

This paper attempts to answer these questions. In this regard, we introduce a novel approach for CS reconstruction of simultaneous low-rank and joint-sparse matrices. Our approach is based on a convex problem that simultaneously uses low-rank and joint-sparsity promoting regularizations, namely, the *nuclear* (*trace*) norm and the ℓ_2/ℓ_1 mixed-norm of the solution matrix. As a result, the solution of this problem tends to have both desired structures. Further, we propose an efficient algorithm in order to solve this optimization problem.

Additionally, we establish a theoretical upper bound on the reconstruction error guaranteeing the *stability* of our recovery approach against noisy measurements, *non-exact* sparse and *approximately* low-rank data matrices. We prove that, if the sampling scheme satisfies a *specific* restricted isometry property (RIP), the proposed approach can stably recover *all* joint-sparse and low-rank matrices. In particular and for a certain class of *random* sampling schemes, we prove that this property implies the following lower bound on the number m of CS measurements:

$$m \gtrsim \mathcal{O}\big(k \log(n_1/k) + kr + n_2 r\big). \tag{1}$$

Compared to the necessary condition for recovery of such structured matrices, *i.e.* $m > (k + n_2 - r)r$, this measurement bound highlights the near-optimality of our recovery approach: its scaling order with

respect to k and n_2 has only an extra $\log$ factor due to the joint-sparsity pattern subset selection.

A. Application in Multichannel Signal CS

Multichannel signals are undoubtedly a perfect fit for the new compressed sensing paradigm: on one hand, they have numerous applications *e.g.* in computer vision, biomedical data, sensor networks, multispectral imaging and audio processing, just to name a few. On the other hand, the main present challenge is the tremendously large and increasing flow of data in those applications, which poses tight constraints on the available technologies at any of the acquisition, compression, transmission and data analysis steps. Despite their huge dimensionality, multichannel data are often highly redundant as a result of two categories of correlations: *inter-* and *intra*-channel correlations. Any efficient dimensionality reduction scheme requires comprehensive modeling of both types of correlations. A straightforward extension of the classical single-channel CS idea [1], [2] only considers inter-channel signal correlations through an *unstructured* sparsity assumption on the signal in each channel and the standard recovery approaches such as ℓ_1 minimization result in limited improvement.

The simultaneous low-rank and joint-sparse model turns out to be an ideal candidate to more restrictively model multichannel data and the underlying structures. In this case, the matrix X represents a multichannel signal where n_2 denotes the number of channels and n_1 is the signal dimensionality per channel. This model not only considers inter-channel signal correlations leading to a sparse representation, but also takes into account intra-channel correlations and is able to assign a particular structure for the sparsity : signals of different channels share a *unique* dominant sparsity pattern (support) and in addition, the values of those sparse coefficients are *linearly dependent* across the multiple channels. Beside its restrictiveness, this model is generic enough to hold for a wide range of applications where the entire multichannel signal can be decomposed into few sparse sources through a *linear source mixture model* as follows:

$$X = \mathbf{S}\mathbf{H}^T,$$

where $\mathbf{S}$ is an $n_1 \times \rho$ matrix whose columns contain $\rho \ll \min(n_1, n_2)$ source signals that are linearly mixed with an $n_2 \times \rho$ matrix $\mathbf{H}$ in order to generate the whole data across multiple channels. Indeed, for a sparse (w.l.o.g. in the canonical basis) source matrix $\mathbf{S}$ and an arbitrary mixing matrix $\mathbf{H}$, this factorization implies a low-rank ($r \leq \rho$) and joint-sparse X.

We have discussed in our previous articles [10]–[12] the application of this model in compressive hyperspectral imagery. The linear mixture model typically holds for this type of images where $\mathbf{H}$ contains

the spectral signatures of a few materials present in the region whose distributions are indicated by the source images, *i.e.* the columns of **S**. We have shown that leveraging both low-rank and sparse structures enables CS to achieve significantly low sampling rates: *A hyperspectral image can be subsampled by retaining only 3% of its original size and yet be robustly reconstructed.*

Alternatively, the same model can be deployed for a much more efficient CS acquisition in sensor network applications and it can eventually enhance the tradeoff between the number n_2 of sensors and the complexity (*e.g.* the number of measurements per channel m/n_2) of each sensor. Indeed, typical sensor network setups are monitoring highly correlated phenomena which are mainly generated by a small number of (sparse) sources distributed in a region, and **H** represents a possibly unknown Green's function accounting for the physical source-to-sensor signal propagation model.

B. Compressive Sparse PCA

Despite similarity in names, this paper considers a different problem than the robust principal component analysis (r-PCA) [13], [14]. We are not interested in decomposing a data matrix into separate low-rank L and sparse S components, *i.e.* $X = L + S$. In contrast, here X simultaneously have both structures and our problem is more related to the *sparse principal component analysis* (s-PCA) literature [15]–[17]. The s-PCA problem considers the following factorization of a data matrix:

$$X \approx UV^T$$

where, the columns of $V \in \mathbb{R}^{\rho \times n_2}$ contain the few $\rho \ll \min(n_1, n_2)$ principal components (PCs) of X and $U \in \mathbb{R}^{n_1 \times \rho}$ is a sparse matrix of the corresponding loadings of the principal components. The classical PCA [18], which is a key tool for multivariate data analysis and dimensionality reduction, numerically involves a singular value decomposition with orthogonal matrices U and V. However, the classical approach comes with a key drawback: in practical setups with n_1 variables and n_2 observations, most of the loadings are nonzero and thus, the main factors explaining data variance are linear combinations of *all* variables, making it often difficult to interpret the results. The idea of s-PCA has been mainly proposed to address this shortcoming at the cost of applying more complex algorithms and loosing orthogonality of U and V.

As previously described in Section I-A, this factorization (*i.e.* having a small number ρ of PCs with a sparse loading matrix U) implies a low-rank and joint-sparse structure for UV^T. Therefore an alternative approach to solve s-PCA is to first use our method to approximate X by a simultaneous low-rank and joint-sparse matrix, and then apply the conventional PCA to decompose the result into U and V. This

approach (*i.e.* s-PCA in the compressive domain) could be particularly useful for large-dimensional data analysis and visualization where data size is the main bottleneck. In addition, and to the best of our knowledge, the theoretical analysis provided in this paper is the first in this direction.

C. Paper Organization

This paper is organized as follows: in Section II we review the most related works in the compressed sensing literature and discuss the efficiency of different data priors for CS recovery. Section III presents our novel CS reconstruction scheme for simultaneous low-rank and joint-sparse data matrices. In Section IV we establish the theoretical guarantees of this recovery approach. The proofs of the main theorems are provided in Section V. In Section VI we introduce a set of iterative algorithms based on the proximal splitting method in order to solve the convex recovery problem. Finally, in Section VII we show the empirical recovery phase transition behavior of our method for different CS acquisition schemes and we compare them to prior art.

D. Notations

Throughout this paper we frequently use the following notations:

For the matrix $X \in \mathbb{R}^{n_1 \times n_2}$, $X_{i,.}$ and $X_{.,j}$ respectively denote its ith row and jth column, and $[X]_{i,j}$ is the element of X in the ith row and the jth column. Moreover, X_{vec} denotes the vector formed by stacking the columns of X below one another:

$$X_{vec} = \begin{bmatrix} X_{.,1} \\ X_{.,2} \\ \vdots \\ X_{.,n_2} \end{bmatrix}.$$

The vector space of matrices is equipped with an inner product defined for any pair of matrices $X, X' \in \mathbb{R}^{n_1 \times n_2}$ as follows:

$$\langle X, X' \rangle \triangleq \mathbf{Tr}(X^T X') = \sum_{i,j} [X]_{i,j} [X']_{i,j},$$

where, $\mathbf{Tr}(.)$ is the trace of a square matrix. Moreover, four matrix norms that we frequently use are:

1) The spectral norm $\|X\|$: the largest singular value of X.
2) The nuclear norm $\|X\|_* \triangleq \sum_i \sigma_i$ which is the sum of the singular values (σ_i) of X.
3) The Frobenius norm $\|X\|_F \triangleq \sqrt{\langle X, X \rangle} = \sqrt{\sum_{i,j} [X]_{i,j}^2} = \sqrt{\sum_i \sigma_i^2}$.
4) The ℓ_2/ℓ_1 mixed-norm $\|X\|_{2,1} \triangleq \sum_i \sqrt{\sum_j [X]_{i,j}^2}$ that is the sum of the ℓ_2 norms of the matrix rows.

Since $\|.\|_*$ is a norm, it is *subadditive i.e.*, $\|X + X'\|_* \leq \|X\|_* + \|X'\|_*$ and the equality (*additivity*) holds if X and X' have row and column spaces that are orthogonal $X^T X' = X(X')^T = \mathbf{0}$ [19].

When X is a rank r matrix, we have the following inequalities:

$$\|X\| \leq \|X\|_F \leq \|X\|_* \leq \sqrt{r}\|X\|_F \leq r\|X\|. \tag{2}$$

In addition, the support of a vector $x \in \mathbb{R}^n$ is defined as the set containing the indices of the nonzero elements of x *i.e.*,

$$\text{supp}(x) \triangleq \Big\{ i \in \{1 \ldots n\} \, : \, x_i \neq 0 \Big\}. \tag{3}$$

This definition can be extended to a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ as follows:

$$\text{supp}(X) \triangleq \bigcup_j \text{supp}(X_{.,j}), \tag{4}$$

which is the union of the supports of all of its columns. We call X a *k-joint-sparse* matrix when there are no more than k rows of X containing nonzero elements, $\mathbf{Card}(\text{supp}(X)) \leq k$, where $\mathbf{Card}(.)$ denotes the cardinality of a set. Note that, if X is k-joint-sparse the following inequalities hold:

$$\|X\|_F \leq \|X\|_{2,1} \leq \sqrt{k}\|X\|_F. \tag{5}$$

Finally, throughout this paper we often use the notation $X_{\mathcal{T}}$ which is a matrix of the same size as X whose rows indexed by the set $\mathcal{T}$ are the same as X and zero elsewhere.

II. Compressed Sensing

Compressed sensing [1], [2] has been introduced in the past decade as an alternative to Shannon's sampling theorem. The original foundations of this new sampling paradigm were built over taking advantage of the *sparsity* of a wide range of natural signals in some properly-chosen orthonormal bases and representing large dimensional sparse vectors by very small number of linear measurements, mainly proportional to their sparsity levels.

1) CS Acquisition Protocol: $m \ll n_1 n_2$ linear and possibly noisy measurements are collected from the data matrix X by defining a linear mapping $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ and its adjoint $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{n_1 \times n_2}$ as follows:

$$y = \mathcal{A}(X) + z, \tag{6}$$

where $y \in \mathbb{R}^m$ is the vector containing the CS measurements and $z \in \mathbb{R}^m$ is the noise vector due to quantization or transmission errors. This formulation equivalently has the following explicit matrix

representation:

$$y = AX_{vec} + z, \tag{7}$$

because any linear mapping $\mathcal{A}(X)$ can be expressed in a matrix form $\mathcal{A}(X) \triangleq AX_{vec}$, where the inner product of the data matrix and each row of $A \in \mathbb{R}^{m \times n_1 n_2}$ gives a noise-free element of the measurement vector. In a similar way, we can express the adjoint operator as

$$\mathcal{A}^*(y) = \mathsf{reshape}(A^T y, n_1, n_2).$$

The reshape operator applies on an $n_1 n_2$ dimensional vector $A^T y$ in order to reshape it into an $n_1 \times n_2$ matrix such that the first n_1 elements fill the first column, the next n_1 elements fill the second column and so on.

The main goal of CS is to robustly recover X from the smallest number of measurements m. Despite the simplicity of using a linear sampling scheme and unlike Shannon's linear signal reconstruction, CS decoding applies more complex and nonlinear approximation methods. In the following we mention some of the most relevant approaches to our work:

2) Unstructured Sparse Recovery by ℓ_1 Minimization: The following ℓ_1 problem, also known as Basis Pursuit De-Noising (BPDN), was among the most successful solutions that have been initially proposed for CS recovery of a sparse vector X_{vec}:

$$\arg\min_X \ \|X_{vec}\|_{\ell_1} \qquad \text{subject to} \quad \|y - \mathcal{A}(X)\|_{\ell_2} \le \varepsilon. \tag{8}$$

Note that regularizing the solutions by their ℓ_1 norm, although known as the best sparsity-inducing convex functional, can only exploit *unstructured* sparsity in the data. It has been shown that if the sampling operator $\mathcal{A}$ satisfies the *restricted isometry property* (RIP), then the solution to (8) can stably recover the underlying data [1], [2]. For sampling matrices A (or equivalently $\mathcal{A}$ in (6)) whose elements are drawn independently at random from the Gaussian, Bernoulli or (in general) subgaussian distributions, this property implies the following lower bound on the number of measurements:

$$m \gtrsim \mathcal{O}\left(kn_2 \log(n_1/k)\right), \tag{9}$$

which indicates a suboptimal performance compared to the limited degrees of freedom of a simultaneous low-rank and joint-sparse data matrix.

3) Joint-Sparse Recovery by ℓ_2/ℓ_1 Minimization: To overcome this limitation, some alternative recovery approaches propose to consider a restrictive structure on the sparsity pattern of the data, namely, a *joint-sparsity* structure for the nonzero coefficients of X. Some approaches based on greedy heuristics

[20]–[22] attempt to identify the joint-sparsity support of X, whereas other methods based on convex relaxation [23]–[25] aim at recovering both the joint-sparsity pattern and the nonzero coefficients by penalizing other matrix norms promoting a joint-sparsity structure to their solutions. In particular, penalizing the ℓ_2/ℓ_1 mixed norm of the data matrix through the following convex problem is very popular and has been widely used in the literature:

$$\arg\min_X \ \|X\|_{2,1} \qquad \text{subject to} \quad \|y - \mathcal{A}(X)\|_{\ell_2} \le \varepsilon. \tag{10}$$

Despite significant improvements in support recovery[1], the joint-sparsity assumption fails to enhance the entire data reconstruction (*i.e.* recovering both the support and the nonzero coefficients). Eldar and Mishali [25] extend the notion of RIP to the *Block*-RIP and show that, if the sampling matrix A satisfies this property, the ℓ_2/ℓ_1 problem can stably reconstruct joint-sparse matrices. For a dense i.i.d. subgaussian sampling matrix A, the number of measurements sufficient for satisfying the Block-RIP is

$$m \gtrsim \mathcal{O}\left(k \log(n_1/k) + k n_2\right). \tag{11}$$

As we can see, for a large number $n_2 \gg k \log(n_1/k)$, this bound indicates only a minor enhancement of a $\log(n_1/k)$ factor compared to the measurement bound (9) corresponding to the classical ℓ_1 problem. This is a fundamental limitation arising from neglecting the underlying correlations among the values of nonzero coefficients *i.e.*, with such a loose assumption (the joint-sparsity model) the data matrix can have $k n_2$ nonzero elements freely taking independent values, and thus generally one requires $\mathcal{O}(k)$ measurements per column for recovery.

4) Low-Rank Recovery by the Nuclear Norm Minimization: Many works have been lately focusing on recovering a low-rank data matrix from an incomplete set of observations. In particular, the *matrix completion* problem [7]–[9] deals with recovering a data matrix given only a small fraction of its entries. The key assumption for a stable reconstruction is that the data matrix is highly redundant due to high linear dependencies among its columns/rows and therefore it has a *low rank*. Similarly, Candes and Plan [26] have shown that if the sampling operator $\mathcal{A}$ satisfies an extended notion of RIP, called the *Rank-RIP*, then the following *nuclear* norm minimization problem can stably recover the underlying low-rank matrix:

$$\arg\min_X \ \|X\|_* \qquad \text{subject to} \quad \|y - \mathcal{A}(X)\|_{\ell_2} \le \varepsilon. \tag{12}$$

---

[1]It has been shown in [22] that a simple one-stage p-Thresholding algorithm can correctly identify the support of an *exact* joint-sparse matrix provided $m \gtrsim \mathcal{O}(k \log(n_1))$.

In particular, if $\mathcal{A}$ is drawn from the i.i.d. subgaussian distribution, then the number of measurements sufficient for $\mathcal{A}$ to satisfy the Rank-RIP (and hence data recovery) is

$$m \gtrsim \mathcal{O}\left(r(n_1 + n_2)\right). \tag{13}$$

Compared to (11) and for a large number n_2, we require less number of measurements thanks to properly dealing with the high data correlations via nuclear norm regularization. As we can see in (13) and compared to the bound (11), the growth of m remains linear by increasing the number n_2 of columns, but scales as rn_2 instead of kn_2 with $r \ll k$. For a large number n_2, this bound indicates that $\mathcal{O}(r)$ measurements per column are sufficient which is proportional to the number of principal components of X rather than the sparsity level of each column $\mathcal{O}(k)$ in (11) or (9). Despite this improvement, for $n_1 \gg n_2$ ignoring sparsity would significantly decrease the efficiency of this approach compared to the few degrees of freedom of a low-rank and joint-sparse data matrix.

5) Rank Aware Joint-Sparse Recovery: Considering those previously described approaches and their weak and strong points, one can think of the idea of combining both data priors together and reconstruct matrices that are *simultaneously* low-rank and joint-sparse. Recently a few works [27], [28] have considered *rank awareness* in data recovery from multiple measurement vectors (MMV). More precisely, the joint-sparse MMV inverse problem [20], [29]–[31] focuses on recovering a joint-sparse matrix X from a set of CS measurements $Y \in \mathbb{R}^{\widetilde{m} \times n_2}$ acquired by $Y = \widetilde{A}X$. We refer to such acquisition model as the *uniform* sampling because a unique sampling matrix $\widetilde{A} \in \mathbb{R}^{m/n_2 \times n_1}$ applies on all columns of X.

Davies and Eldar [27] proposed a specific rank-aware *greedy* algorithm, that in case of using a random i.i.d. Gaussian sampling matrix $\widetilde{A}$ is able to recover (with a high probability) an *exact* k-joint-sparse and rank r X from the noiseless MMV provided

$$m = n_2\widetilde{m} \gtrsim \mathcal{O}\left(n_2 k \left(1 + \frac{\log n_1}{r}\right)\right). \tag{14}$$

It can be viewed from (14) that this approach is more suitable for applications with *high-rank* data matrices, where the columns of X are as independent as possible from each other. This is contrary to our assumption of having a redundant, low-rank data structure. This limitation comes from the fact that *the uniform sampling scheme is not generally suitable for acquisition of correlated data matrices*. Given highly redundant data, such sampling mechanism results in a redundant set of measurements (X is low-rank and so becomes Y, because $\mathrm{rank}(Y) \leq \mathrm{rank}(X)$). Think of the extreme example of having a joint-sparse rank one matrix whose columns are identical to each other. Using a uniform sampling scheme would result in multiple copies of the same measurements for all the columns and thus, applying any

joint decoding scheme exploiting the intra-column correlations would be pointless as the measurements do not carry any new information across the columns.

III. SIMULTANEOUS LOW-RANK AND JOINT-SPARSE MATRIX APPROXIMATION

Data reconstruction from CS measurements typically reduces to solving an underdetermined and noisy linear system of equations such as (6) where $m \ll n_1 n_2$. Without any further assumptions (6) would admit infinitely many solutions and thus, recovery of the original data matrix is not generally feasible. As previously discussed, however, data matrices of interest in this paper have a particular simultaneous low-rank and joint-sparse structure restricting data to have very few degrees of freedom, *i.e.* less than both the matrix dimension and the dimension of the underlying sparse coefficients. Our main goal in this part is to propose a robust recovery scheme which incorporates properly these constrains in order to make (6) invertible with as few measurements as possible.

Recovery of a joint-sparse and rank r data matrix from a set of noisy CS measurements can be casted as the solution to the following minimization problem:

$$\arg\min_{X} \quad \mathbf{Card}(\mathrm{supp}(X)) \tag{15}$$

$$\text{subject to} \quad \|y - \mathcal{A}(X)\|_{\ell_2} \leq \varepsilon,$$

$$\mathrm{rank}(X) \leq r,$$

where ε is an upper bound on the energy of the noise, $\|z\|_{\ell_2} \leq \varepsilon$. Note that finding the sparsest solution to a linear system of equations, in general, is an NP-hard problem [32]. Problems such as minimizing $\mathbf{Card}(\mathrm{supp}(X))$ or the rank minimization are similarly reducing to the cardinality minimization problem of finding the sparsest element in an affine space and thus in general, finding solution to a problem like (15) is computationally intractable. The original solution must be approximated by a polynomial-time method.

A trivial solution could be the *cascade* of two classical CS recovery schemes: One identifies the joint-support set and the other evaluates the rows of the matrix corresponding to the recovered support using a low-rank approximation scheme. Unfortunately, the application of this approach is mainly limited to the recovery of the *exact* low-rank and joint-sparse matrices. In practice signals are almost never exactly sparse, and using the cascade strategy (*e.g.* recovering first the dominant k-joint-sparse part of data, and then extracting a rank r matrix from it) for dense data matrices might not be always optimal in order to identify the most dominant component having simultaneously both structures.

In order to provide *robustness* against non-exact but approximately low-rank and joint-sparse data, we incorporate both data priors into a single recovery formulation. We relax (15) by replacing the non-convex terms $\mathrm{rank}(X)$ and $\mathbf{Card}(\mathrm{supp}(X))$ by the convex terms inducing similar structures to their solutions *i.e.*, the nuclear norm $\|X\|_*$ and the $\|X\|_{2,1}$ mixed-norm, correspondingly. We cast the CS recovery problem as solving one of the following three convex problems:

1) Having data nuclear norm as the regularization parameter, or provided with an estimation of it, we solve:

$$\underset{X}{\arg\min} \quad \|X\|_{2,1} \tag{16}$$

$$\text{subject to} \quad \|y - \mathcal{A}(X)\|_{\ell_2} \le \varepsilon,$$

$$\|X\|_* \le \tau.$$

2) Having the ℓ_2/ℓ_1 mixed-norm of data as the regularization parameter, or provided with an estimation if it, we solve:

$$\underset{X}{\arg\min} \quad \|X\|_* \tag{17}$$

$$\text{subject to} \quad \|y - \mathcal{A}(X)\|_{\ell_2} \le \varepsilon,$$

$$\|X\|_{2,1} \le \gamma.$$

3) Having a regularization factor λ balancing between the ℓ_2/ℓ_1 norm and the nuclear norm, we solve:

$$\underset{X}{\arg\min} \quad \|X\|_{2,1} + \lambda\|X\|_* \tag{18}$$

$$\text{subject to} \quad \|y - \mathcal{A}(X)\|_{\ell_2} \le \varepsilon.$$

We explain in Section VI that all three problems above can be solved iteratively (in polynomial time) using the parallel proximal splitting algorithm (PPXA) [33]. Note that all three formulations (16), (17) and (18) contain a regularization parameter (either τ, γ or λ) that should be correctly tuned. In the next section we show that, for properly-chosen parameters, the solutions of these convex problems can robustly approximate the original data provided by very few linear measurements

IV. THEORETICAL GUARANTEES

In this section we provide the theoretical analysis of the recovery problems (16)-(18). We prove that for certain linear mappings $\mathcal{A}$, these three convex relaxations can exactly reconstruct all simultaneous low-rank and joint-sparse data from their corresponding CS measurements. In addition, our results guarantee

the stability of our approach against noisy measurements and non-exact low-rank and joint-sparse data. For a certain class of random linear mappings our results establish a lower bound on the number m of CS measurements in terms of the joint-sparsity level k and the rank r of the underlying data matrix.

A. Sampling Operator RIP

We define a specific restricted isometry property (RIP) that later helps us to build our main theoretical results.

Definition 1. *For positive integers $k, r \in \mathbb{N}_+$, a linear map $\mathcal{A}$ satisfies the restricted isometry property, if for all simultaneous rank r and k-joint-sparse matrices X we have the following inequalities:*

$$(1 - \delta_{r,k})\|X\|_F^2 \leq \|\mathcal{A}(X)\|_{\ell_2}^2 \leq (1 + \delta_{r,k})\|X\|_F^2. \tag{19}$$

The RIP constant $\delta_{r,k}$ is the smallest constant for which the property above holds.

As we can see and compared to the two former definitions of the Block-RIP [25] and the Rank-RIP [19], [26], Definition 1 provides a more restrictive property which holds for the intersection set of the low-rank and the joint (block) sparse matrices. Verifying whether a linear map satisfies RIP is generally combinatorial and thus prohibitive for large size problems. However, for linear mappings drawn at random from a certain class of distributions and thanks to their well-established concentration properties, the following theorem and corollary guarantee RIP:

Theorem 1. *Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be a random linear map obeying the following concentration bound for any $X \in \mathbb{R}^{n_1 \times n_2}$ and $0 < t < 1$,*

$$\mathbf{Pr}\left(\left| \|\mathcal{A}(X)\|_{\ell_2}^2 - \|X\|_F^2 \right| > t\|X\|_F^2 \right) \leq C \exp\left(-c\,m\right), \tag{20}$$

where C and c are fixed constants given t. Then $\mathcal{A}$ satisfies RIP with constant $\delta_{r,k}$ with probability greater than $1 - Ce^{-\kappa_0 m}$, if the number of measurements is greater than

$$m \geq \kappa_1\Big(k \log(n_1/k) + (k+1)r + n_2 r\Big), \tag{21}$$

where κ_0 and κ_1 are fixed constants for a given $\delta_{r,k}$.

Corollary 1. *For a linear map $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ corresponding to a matrix $A \in \mathbb{R}^{m \times n_1 n_2}$ whose elements are drawn independently at random from the Gaussian $\mathcal{N}(0, 1/m)$, Bernoulli $\{\pm 1/\sqrt{m}\}$ or subgaussian (with a proper normalization) distributions, RIP holds whenever the number of measurements satisfies* (21).

The proof of Theorem 1 is provided in Section V-A and Corollary 1 is a straightforward application of Theorem 1. Indeed, for the Gaussian distribution since $\|\mathcal{A}(X)\|_{\ell_2}^2$ is a chi-squared random variable with m degrees of freedom multiplied by $\|X\|_F^2/m$, the following concentration bound holds:

$$\mathbf{Pr}\left(\left|\|\mathcal{A}(X)\|_{\ell_2}^2 - \|X\|_F^2\right| > t\|X\|_F^2\right) \leq 2\exp\left(-\frac{m}{2}(t^2/2 - t^3/3)\right). \tag{22}$$

Similar bounds hold for the Bernoulli and subgaussian ensembles with different constants. Regarding Theorem 1, random linear maps drawn from these distributions indeed satisfy the deviation bound (20) and thus RIP holds for them provided (21).

B. Main Reconstruction Error Bounds

Our main performance bounds here are derived based on the new notion of RIP defined above. Given a set of linear measurements that are collected from a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ (which is not necessarily low-rank or joint-sparse) through the sampling model (6) and by using a sampling operator $\mathcal{A}$ satisfying RIP, the following theorems then bound the approximation error of the recovery problems (16)-(18):

Theorem 2. *Suppose $\delta_{6r,2k} < 1/7$ and $\|z\|_{\ell_2} \leq \varepsilon$. Let $X_{r,k}$ be any rank r and k-joint-sparse matrix and $\mathcal{T}_0 = \mathrm{supp}(X_{r,k})$.*

(a) For $\tau = \|X\|_*$, the solution $\widehat{X}$ to (16) obeys the following bound:

$$\|\widehat{X} - X\|_F \leq \kappa_0'\left(\frac{\|X - X_{\mathcal{T}_0}\|_{2,1}}{\sqrt{k}} + \frac{\|X - X_{r,k}\|_*}{\sqrt{2r}}\right) + \kappa_1'\varepsilon. \tag{23}$$

Constants κ_0' and κ_1' are fixed for a given $\delta_{r,k}$.

(b) For $\gamma = \|X\|_{2,1}$, the solution $\widehat{X}$ to (17) obeys the bound (23).

(c) For $\lambda = \sqrt{\frac{k}{2r}}$, the solution $\widehat{X}$ to (18) obeys the bound (23).

Corollary 2. *Suppose $\delta_{6r,2k} < 1/7$ and the measurements are noiseless. Then, all exact (simultaneous) k-joint-sparse and rank r matrices X can be exactly recovered by (16), (17) or (18) (with proper regularization parameters) i.e., $\widehat{X} = X$.*

The proof of Theorem 2 is provided in Section V-B. The proof of the corollary is trivial by setting $X_{r,k} = X_{\mathcal{T}_0} = X$ and $\varepsilon = 0$ in (23). Given noisy measurements and a data matrix X that is not exactly low-rank or joint-sparse, the reconstruction error (23) is bounded by two terms. First, it is proportional (with a constant factor) to the power of the sampling noise. Second, it is determined by the best rank r and k-joint-sparse approximation of X that minimizes the first term of (23). This term can be interpreted

as a measure of data *compressibility i.e.*, how accurately data can be approximated by a simultaneous low-rank and joint-sparse matrix.

According to Corollary 1, for sampling matrices A drawn independently at random from the Gaussian, Bernoulli or subgaussian distributions (and more generally, all random linear maps $\mathcal{A}$ satisfying (20)), Theorem 2 guarantees that our proposed reconstruction approach based on joint nuclear and ℓ_2/ℓ_1 norms minimization can stably recover all simultaneous rank r and k-joint-sparse matrices X (with an overwhelming probability) provided

$$m \gtrsim \mathcal{O}\big(k\log(n_1/k) + kr + n_2r\big),$$

which, as previously mentioned, indicates a near-optimal recovery performance regarding the degrees of freedom of those data matrices *i.e.*, $r(k + n_2 - r)$. In addition, this bound reveals that our proposed recovery approach requires significantly less number of CS measurements to reconstruct such structured data compared to the prior art methods (see (11), (13) and (14)). Unlike the ℓ_2/ℓ_1 approach (10) and for a large number n_2 of columns, our approach requires $\mathcal{O}(r)$ measurements per column, which is proportional to the number of principal components of this data. In addition, compared to the nuclear norm minimization scheme (12), we achieve a shaper lower bound on m that is proportional to the joint-sparsity level k of the data matrix and scales sub-linearly with the column dimensionality n_1. Finally, our bound is of a different nature than (14): the lower the rank, less measurements are required. This reflects the importance of a good design for the sampling scheme $\mathcal{A}$ along with a recovery approach benefiting efficiently from data structures.

Note that Theorem 2 precisely characterizes the proper values for the regularization parameters in (16)-(18). If there is no prior knowledge available about the nuclear norm or the ℓ_2/ℓ_1 norm of the underlying data (and a blind tuning might be challenging), (18) brings more flexibility to the reconstruction process. For a given λ, Theorem 2-c guarantees that the recovery error is bounded by the minimum of (23) for the best pair r, k satisfying both RIP (*i.e.* $\delta_{6r,2k} < 1/7$ and $\lambda = \sqrt{k/2r}$.

V. Proofs

A. Proof of Theorem 1

Let $\widetilde{X}_{\mathcal{T}}$ denote the $k \times n_2$ submatrix formed by selecting the rows of X indexed by a set $\mathcal{T} \subset \{1, \ldots, n_1\}$ with the cardinality k. Correspondingly, for a given $\mathcal{T}$ we define the restricted map $\mathcal{A}_{\mathcal{T}} : \mathbb{R}^{k \times n_2} \to \mathbb{R}^m$ which applies on $\widetilde{X}_{\mathcal{T}}$, and is related to $\mathcal{A}$ through its explicit matrix formulation (7) as

follows:

$$\mathcal{A}_{\mathcal{T}}(\widetilde{X}_{\mathcal{T}}) \triangleq \widetilde{A}_{\mathcal{T}}\,(\widetilde{X}_{\mathcal{T}})_{vec},$$

where, $\widetilde{A}_{\mathcal{T}} \in \mathbb{R}^{m \times k n_2}$ is the submatrix containing the columns of A indexed by $\mathcal{T}$. If we assume X is an exact k-joint-sparse matrix whose rows are supported over the set $\mathcal{T}$ and zero elsewhere, we will have

$$\mathcal{A}(X) = \mathcal{A}_{\mathcal{T}}(\widetilde{X}_{\mathcal{T}}).$$

In this case and according to our assumption (*i.e.*, $\mathcal{A}$ satisfies (20)) we can write a similar concentration bound for a given $\mathcal{T}$ and $\widetilde{X}_{\mathcal{T}}$:

$$\mathbf{Pr}\left(\left|\|\mathcal{A}_{\mathcal{T}}(\widetilde{X}_{\mathcal{T}})\|_{\ell_2}^2 - \|\widetilde{X}_{\mathcal{T}}\|_F^2\right| > t\|\widetilde{X}_{\mathcal{T}}\|_F^2\right) \le C\exp\left(-c\,m\right). \tag{24}$$

According to the Rank-RIP property defined in [26], if a linear map $\mathcal{A}_{\mathcal{T}}$ satisfies the bound (24) then, for a given $\mathcal{T}$ and *all* rank r realizations of $\widetilde{X}_{\mathcal{T}}$, we have

$$\mathbf{Pr}\left(\left|\|\mathcal{A}_{\mathcal{T}}(\widetilde{X}_{\mathcal{T}})\|_{\ell_2}^2 - \|\widetilde{X}_{\mathcal{T}}\|_F^2\right| > \delta\|\widetilde{X}_{\mathcal{T}}\|_F^2\right) \le C\exp\left((k + n_2 + 1)r\log(36\sqrt{2}/\delta) - c\,m\right), \tag{25}$$

where $0 < \delta < 1$ is a fixed constant. The proof of this statement mainly uses a *covering number* argument for low-rank matrices and it can be found in [26] (see theorem 2.3). Now, by considering *all* possible combinations for the subset $\mathcal{T}$ *i.e.*, $\binom{n_1}{k} \le \left(\frac{n_1}{k}e\right)^k$ and applying a union bound on (25), we can deduce (19) for *all* k-joint-sparse and rank r matrices X with probability greater than

$$1 - C\exp\left(k + k\log(n_1/k) + (k + n_2 + 1)r\log(36\sqrt{2}/\delta) - c\,m\right).$$

Equivalently, with probability exceeding $1 - Ce^{-\kappa_0 m}$, $\mathcal{A}$ satisfies RIP with constant $\delta_{r,k} \le \delta$, whenever

$$m \ge \kappa_1\Big(k\log(n_1/k) + (k + n_2 + 1)r\Big),$$

where κ_0 and κ_1 are fixed constants for a given δ *i.e.*, $\kappa_0 = c - \log(36\sqrt{2}/\delta)/\kappa_1$, and $\kappa_1 > \log(36\sqrt{2}/\delta)/c$.

B. Proof of Theorem 2

In this part we first provide the proof of Theorem 2-a and later, we use similar techniques to prove Theorems 2-b and 2-c. For this purpose, we prove the following two lemmas that are the key elements for the proof of Theorem 2.

An important implication of the RIP is that for linear maps $\mathcal{A}$ having small RIP constants, the measurement outcomes of any two k-joint-sparse and rank r orthogonal matrices become *nearly* orthogonal. This statement is precisely characterized by the following lemma:

Lemma 1. $\forall X, X'$ *such that,* $\mathrm{rank}(X) \leq r$, $\mathrm{rank}(X') \leq r'$, **Card**$(\mathrm{supp}(X)) \leq k$, **Card**$(\mathrm{supp}(X')) \leq k'$ *and* $\langle X, X' \rangle = 0$, *we have:*

$$\left| \langle \mathcal{A}(X), \mathcal{A}(X') \rangle \right| \leq \delta_{r+r', k+k'} \|X\|_F \|X'\|_F \tag{26}$$

Proof: Without loss of generality assume X and X' have unit Frobenius norms. Since $X \pm X'$ is at most $k + k'$ joint-sparse and $\mathrm{rank}(X \pm X') \leq r + r'$, we can write

$$(1 - \delta_{r+r', k+k'})\|X \pm X'\|_F^2 \leq \|\mathcal{A}(X \pm X')\|_{\ell_2}^2 \leq (1 + \delta_{r+r', k+k'})\|X \pm X'\|_F^2.$$

Moreover, X and X' are orthogonal to each other which implies that $\|X \pm X'\|_F^2 = \|X\|_F^2 + \|X'\|_F^2 = 2$. Now, by using the parallelogram identity we can deduce:

$$\left| \langle \mathcal{A}(X), \mathcal{A}(X') \rangle \right| = \frac{1}{4} \left| \|\mathcal{A}(X + X')\|_{\ell_2}^2 - \|\mathcal{A}(X - X')\|_{\ell_2}^2 \right| \leq \delta_{r+r', k+k'},$$

which completes the proof. ∎

Lemma 2. *Let A and B be matrices of the same size and* $\mathrm{rank}(A) \leq r$, $\mathrm{supp}(A) = \mathcal{T}$. $\exists B_1, B_2$ *(matrices of the same size as A and B) such that:*

1) $B = B_1 + B_2$,
2) $\mathrm{supp}(B_1) = \mathcal{T}$, $\mathrm{rank}(B_1) \leq 2r$,
3) $A^T B_2 = \mathbf{0}$, $A(B_2)^T = \mathbf{0}$,
4) $\langle B_1, B_2 \rangle = 0$.

Proof: Suppose $\widetilde{A}_{\mathcal{T}}$ and $\widetilde{B}_{\mathcal{T}}$ are submatrices correspondingly containing the rows of A and B indexed by $\mathcal{T}$. Recht *et al.*, (see Lemma 3.4 in [19]) propose a decomposition $\widetilde{B}_{\mathcal{T}} = \widetilde{B}_1 + \widetilde{B}_2$ so that

- $\mathrm{rank}(\widetilde{B}_1) \leq 2\,\mathrm{rank}(\widetilde{A}) \leq 2r$,
- $(\widetilde{A}_{\mathcal{T}})^T \widetilde{B}_2 = \mathbf{0}$, $\widetilde{A}_{\mathcal{T}}(\widetilde{B}_2)^T = \mathbf{0}$,
- $\langle \widetilde{B}_1, \widetilde{B}_2 \rangle = 0$.

On the other hand, lets decompose $B = B_{\mathcal{T}} + B_{\mathcal{T}^c}$ where $B_{\mathcal{T}}$ and $B_{\mathcal{T}^c}$ are respectively the same as B on the index sets $\mathcal{T}$ and $\mathcal{T}^c$ (the set complement of $\mathcal{T}$) and zero elsewhere. Suppose $B_{1_{\mathcal{T}}}$ and $B_{2_{\mathcal{T}}}$ are matrices of the same size as B, whose rows indexed by $\mathcal{T}$ are respectively populated by the submatrices $\widetilde{B}_1$ and $\widetilde{B}_2$, and the remaining elements are zero. We can write $B_{\mathcal{T}} = B_{1_{\mathcal{T}}} + B_{2_{\mathcal{T}}}$. Now, by letting $B_1 = B_{1_{\mathcal{T}}}$ and $B_2 = B_{2_{\mathcal{T}}} + B_{\mathcal{T}^c}$ one can easily verifies that B_1 and B_2 satisfy properties (1)-(4). ∎

1) Decomposition of The Error Term: Here, by using Lemma 2 we decompose the error term $H = X - \widehat{X}$ into the summation of rank $2r$ and k-joint-sparse matrices that are mutually orthogonal to each other. This decomposition later appears to be essential to bound the error using Lemma 1.

Let us fix the notation so that $X_{\mathcal{T}} \in \mathbb{R}^{n_1 \times n_2}$ is a joint-sparse matrix which is identical to X for the rows indexed by $\mathcal{T}$, and zero elsewhere. Let $\mathcal{T}_0$ denote the joint support of $X_{r,k}$ *i.e.*, $\mathcal{T}_0 = \mathrm{supp}(X_{r,k})$ with $\mathbf{Card}(\mathcal{T}_0) \leq k$. Since $X_{r,k}$ is additionally a rank r matrix, we can use Lemma 2 to decompose H into orthogonal matrices $H = H_{\mathcal{T}_0}^0 + \overline{H}$ such that,

- $\mathrm{rank}(H_{\mathcal{T}_0}^0) \leq 2r$,
- $\langle H_{\mathcal{T}_0}^0, \overline{H} \rangle = 0$,
- $(X_{r,k})^T \overline{H} = X_{r,k} \overline{H}^T = \mathbf{0}$.

Further, we decompose $\overline{H}$ into the summation of disjoint k-joint-sparse matrices. For this purpose we write $\overline{H} = \overline{H}_{\mathcal{T}_0} + \overline{H}_{\mathcal{T}_0^c}$, where $\mathcal{T}_0^c$ denotes the set complement of the support $\mathcal{T}_0$. Now we decompose $\overline{H}_{\mathcal{T}_0^c}$ into the following terms:

$$\overline{H}_{\mathcal{T}_0^c} = \sum_{i \geq 1} \overline{H}_{\mathcal{T}_i},$$

where $\mathcal{T}_1$ is the set containing the indices of the k rows of $\overline{H}_{\mathcal{T}_0^c}$ with the largest ℓ_2 norm, $\mathcal{T}_2$ contains the indices of the next k high energy rows and so on. By construction $\overline{H}_{\mathcal{T}_i}$s are supported on disjoint sets. We simplify the notations by using $\overline{H}_i$ instead of $\overline{H}_{\mathcal{T}_i}$.

Next, we decompose each $\overline{H}_i$ into the summation of rank $2r$ matrices:

$$\overline{H}_i = \sum_{j \geq 1} \overline{H}_{i,j}.$$

We use here a similar decomposition proposed in [26]: let $\overline{H}_i = U \Sigma V^T$ be the ordered singular value decomposition of $\overline{H}_i$ *i.e.*, $\mathrm{diag}(\Sigma)$ is the ordered sequence of the singular values ($\Sigma_{i,i} \geq \Sigma_{i+1,i+1}$). For integers $j \geq 1$, define the index sets $J_j := \{2r(j-1) + 1, ..., 2rj\}$. Then, the decomposition terms will be

$$\overline{H}_{i,j} := U_{J_j} \Sigma_{J_j} (V_{J_j})^T,$$

i.e. $\overline{H}_{i,1}$ corresponds to the first $2r$ largest singular values of $\overline{H}_i$, $\overline{H}_{i,2}$ corresponds to the next $2r$ largest singular values of $\overline{H}_i$ and so on.

Those two steps suggest a decomposition $\overline{H} = \sum_{i \geq 0, j \geq 1} \overline{H}_{i,j}$, whose terms $\overline{H}_{i,j}$ are k-joint-sparse and rank $2r$ matrices mutually orthogonal to each other *i.e.*, $\forall (i,j) \neq (i', j')$

$$\langle \overline{H}_{i,j}, \overline{H}_{i',j'} \rangle = 0.$$

Additionally, $H_{\mathcal{T}_0}^0$ is orthogonal to all $\overline{H}_{i,j}$ as a result of either having disjoint supports (*i.e.*, for $i \geq 1$), or by the specific constructions of $H_{\mathcal{T}_0}^0$ (see Lemma 2 in this paper, and Lemma 3.4 in [19]) and $\overline{H}_{0,j}$.

Standard computations show that the following inequality holds for all i:

$$\sum_{j \geq 2} \|\overline{H}_{i,j}\|_F \leq \frac{\|\overline{H}_i\|_*}{\sqrt{2r}}. \tag{27}$$

Also, we have

$$\sum_{i \geq 2} \|\overline{H}_i\|_F \leq \frac{\|\overline{H}_{\mathcal{T}_0^c}\|_{2,1}}{\sqrt{k}}. \tag{28}$$

Inequalities (27) and (28) are respectively the consequences of

$$\|\overline{H}_{i,j}\|_F \leq (2r)^{-1/2} \|\overline{H}_{i,j-1}\|_*$$

and for $i > 1$

$$\|\overline{H}_i\|_F \leq k^{-1/2} \|\overline{H}_{i-1}\|_{2,1}.$$

As a result, the following inequalities hold:

$$\mathcal{S} \triangleq \sum_{\substack{i \geq 0, j \geq 1 \\ (i,j) \neq (1,1),(0,1)}} \|\overline{H}_{i,j}\|_F$$

$$\leq \sum_{i \geq 2} \|\overline{H}_i\|_F + \sum_{i \geq 0, j \geq 2} \|\overline{H}_{i,j}\|_F \tag{29}$$

$$\leq \frac{1}{\sqrt{k}} \|\overline{H}_{\mathcal{T}_0^c}\|_{2,1} + \frac{1}{\sqrt{2r}} \sum_{i \geq 0} \|\overline{H}_i\|_* \tag{30}$$

$$\leq \frac{1}{\sqrt{k}} \|\overline{H}_{\mathcal{T}_0^c}\|_{2,1} + \frac{1}{\sqrt{2r}} \|\overline{H}\|_*. \tag{31}$$

2) Proof of the Main Error Bounds: We start with the proof of Theorem 2-a. Since $\mathcal{A}(X)$ is corrupted by a noise whose ℓ_2 norm is assumed to be bounded by ε, and $\widehat{X}$ is a feasible point for (16) we can write,

$$\|\mathcal{A}(H)\|_{\ell_2} = \|\mathcal{A}(X - \widehat{X})\|_{\ell_2}$$

$$\leq \|y - \mathcal{A}(X)\|_{\ell_2} + \|y - \mathcal{A}(\widehat{X})\|_{\ell_2} \leq 2\varepsilon. \tag{32}$$

By assumption $\|X\|_* = \tau$, and since $\widehat{X}$ is a feasible point of (16) satisfying $\|\widehat{X}\|_* \leq \tau$ we have

$$\|X\|_* \geq \|X + H\|_*$$
$$\geq \|X_{r,k} + \overline{H}\|_* - \|\overline{X}_{r,k}\|_* - \|H_{\mathcal{T}_0}^0\|_*$$
$$= \|X_{r,k}\|_* + \|\overline{H}\|_* - \|\overline{X}_{r,k}\|_* - \|H_{\mathcal{T}_0}^0\|_*, \tag{33}$$

which yields the following inequality:

$$\|\overline{H}\|_* \leq \|H_{\mathcal{T}_0}^0\|_* + 2\|\overline{X}_{r,k}\|_*. \tag{34}$$

Note that equality (33) holds because both row and column spaces of $X_{r,k}$ and $\overline{H}$ are orthogonal to each other. On the other hand, as a result of the minimization (16) we have

$$\|X\|_{2,1} \geq \|X + H\|_{2,1}$$
$$= \|X_{\mathcal{T}_0} + H_{\mathcal{T}_0}\|_{2,1} + \|X_{\mathcal{T}_0^c} + H_{\mathcal{T}_0^c}\|_{2,1}$$
$$\geq \|X_{\mathcal{T}_0}\|_{2,1} - \|H_{\mathcal{T}_0}\|_{2,1} - \|X_{\mathcal{T}_0^c}\|_{2,1} + \|H_{\mathcal{T}_0^c}\|_{2,1}.$$

Consequently, we can deduce:

$$\|H_{\mathcal{T}_0^c}\|_{2,1} \leq \|H_{\mathcal{T}_0}\|_{2,1} + 2\|X_{\mathcal{T}_0^c}\|_{2,1}. \tag{35}$$

Note that $H_{\mathcal{T}_0^c} = \overline{H}_{\mathcal{T}_0^c}$. Now, dividing both sides of (35) by $\sqrt{k}$ and using (5) yields:

$$\frac{1}{\sqrt{k}}\|\overline{H}_{\mathcal{T}_0^c}\|_{2,1} \leq \|H_{\mathcal{T}_0}\|_F + 2e$$
$$\leq \|H_{\mathcal{T}_0}^0\|_F + \|\overline{H}_{0,1}\|_F + \sum_{j \geq 2}\|\overline{H}_{0,j}\|_F + 2e$$
$$\leq \|H_{\mathcal{T}_0}^0\|_F + \|\overline{H}_{0,1}\|_F + \frac{1}{\sqrt{2r}}\|\overline{H}\|_* + 2e \tag{36}$$
$$\leq 2\|H_{\mathcal{T}_0}^0\|_F + \|\overline{H}_{0,1}\|_F + 2e' + 2e, \tag{37}$$

where $e = \frac{\|X_{\mathcal{T}_0^c}\|_{2,1}}{\sqrt{k}}$ and $e' = \frac{\|\overline{X}_{r,k}\|_*}{\sqrt{2r}}$. Inequality (36) is the consequence of (27), and for deriving (37) we use (34) and (2).

Using (37) and (34), we can continue to upper bound the summation of the Frobenius norms in (31) as follows:

$$\mathcal{S} \leq 3\|H_{\mathcal{T}_0}^0\|_F + \|\overline{H}_{0,1}\|_F + 4e' + 2e.$$
$$\leq 3\sqrt{2}\|H_{\mathcal{T}_0}^0 + \overline{H}_{0,1}\|_F + e'', \tag{38}$$

where $e'' = 4e' + 2e$. Note that for the last inequality we use

$$\|H^0_{\mathcal{T}_0}\|_F + \|\overline{H}_{0,1}\|_F \leq \sqrt{2}\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1}\|_F,$$

which holds because $H^0_{\mathcal{T}_0}$ and $\overline{H}_{0,1}$ are orthogonal to each other.

Meanwhile, by such specific error decomposition we have

$$\mathrm{rank}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}) \leq 6r,$$

$$\mathrm{supp}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}) \leq 2k,$$

and thus according to Definition 1 we can write

$$(1 - \delta_{6r,2k})\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|^2_F \leq \|\mathcal{A}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1})\|^2_{\ell_2}. \tag{39}$$

On the other hand,

$$\|\mathcal{A}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1})\|^2_{\ell_2} = \left| \langle \mathcal{A}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}), \mathcal{A}(H - \widetilde{H}) \rangle \right|,$$

where,

$$\widetilde{H} = \sum_{\substack{i \geq 0, j \geq 1 \\ (i,j) \neq (1,1),(0,1)}} \overline{H}_{i,j}.$$

Using the Cauchy-Schwarz inequality and (32) yields:

$$\left| \langle \mathcal{A}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}), \mathcal{A}(H) \rangle \right|$$

$$\leq \|\mathcal{A}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1})\|_{\ell_2} \|\mathcal{A}(H)\|_{\ell_2}$$

$$\leq 2\varepsilon \sqrt{1 + \delta_{6r,2k}} \|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|_F. \tag{40}$$

Moreover, because $H^0_{\mathcal{T}_0} + \overline{H}_{0,1}$ is at most a rank $4r$, k-joint-sparse matrix orthogonal to the summation terms in $\widetilde{H}$, and so is $\overline{H}_{1,1}$ (a rank $2r$ and k-joint-sparse matrix), we can use Lemma 1 to write

$$\left| \langle \mathcal{A}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}), \mathcal{A}(\widetilde{H}) \rangle \right| \leq \left(\delta_{6r,2k}\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1}\|_F + \delta_{4r,2k}\|\overline{H}_{1,1}\|_F \right) \times \left(\sum_{\substack{i \geq 0, j \geq 1 \\ (i,j) \neq (1,1),(0,1)}} \|\overline{H}_{i,j}\|_F \right)$$

$$\leq \sqrt{2}\,\delta_{6r,2k}\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|_F \times \mathcal{S}$$

$$\leq 6\,\delta_{6r,2k}\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|^2_F + \sqrt{2}\,e''\,\delta_{6r,2k}\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|_F. \tag{41}$$

The second inequality is the outcome of the orthogonality of $H^0_{\mathcal{T}_0} + \overline{H}_{0,1}$ and $\overline{H}_{1,1}$, and replacing the sum of the Frobenius norms by its upper bound (38) gives the last inequality. Inequalities (40) and (41) are providing an upper bound on $\|\mathcal{A}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1})\|^2_{\ell_2}$:

$$\|\mathcal{A}(H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1})\|^2_{\ell_2} \quad \leq 6\,\delta_{6r,2k}\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|^2_F$$
$$+ \left(\sqrt{2}\,e'' \delta_{6r,2k} + 2\varepsilon\,\sqrt{1 + \delta_{4r,2k}}\right)\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|_F.$$

Now by using the lower bound derived in (39), we can deduce the following inequality:

$$\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|_F \leq \alpha e'' + \beta\epsilon, \tag{42}$$

where,

$$\alpha = \sqrt{2}\frac{\delta_{6r,2k}}{1 - 7\delta_{6r,2k}}$$

and

$$\beta = 2\frac{\sqrt{1 + \delta_{6r,2k}}}{1 - 7\delta_{6r,2k}}.$$

Finally, we upper bound the whole error term H as follows:

$$\|H\|_F \leq \|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|_F + \mathcal{S}$$
$$\leq (3\sqrt{2} + 1)\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|_F + e'' \tag{43}$$
$$\leq \kappa'_0\left(\frac{\|X_{\mathcal{T}^c_0}\|_{2,1}}{\sqrt{k}} + \frac{\|\overline{X}_{r,k}\|_*}{\sqrt{2r}}\right) + \kappa'_1\varepsilon. \tag{44}$$

Inequality (43) uses the bound derived in (38). Replacing $\|H^0_{\mathcal{T}_0} + \overline{H}_{0,1} + \overline{H}_{1,1}\|_F$ by its upper bound in (42) results in the last line where the constants $\kappa'_0 = 4 + 4(3\sqrt{2}+1)\alpha$ and $\kappa'_1 = (3\sqrt{2}+1)\beta$ are fixed for a given RIP constant $\delta_{6r,2k} < 1/7$. This completes the proof of Theorem 2-a.

Proof of Theorem 2-b: we follow the same steps as for Theorem 2-a. Here, inequalities (33) and (34) are the outcome of the nuclear norm minimization in (17) i.e., $\| X \|_* \geq \|X + H\|_*$, and (35) holds since $\gamma = \|X\|_{2,1} \geq \|X + H\|_{2,1}$. Thus, one can easily verifies that all bounds (27)-(44) are valid and the solution of (17) obeys the error bound in (23).

Proof of Theorem 2-c: it is also similar to Theorem 2-a and almost all the previous bounds hold. We just need to develop a new bound on $\frac{1}{\sqrt{k}}\|\overline{H}_{\mathcal{T}^c_0}\|_{2,1} + \frac{1}{\sqrt{2r}}\|\overline{H}\|_*$ i.e. inequality (31). From minimization

(18) we have

$$\|X\|_{2,1} + \lambda\|X\|_*$$

$$\geq \|X+H\|_{2,1} + \lambda\|X+H\|_*$$

$$\geq \|X_{\mathcal{T}_0}\|_{2,1} - \|H_{\mathcal{T}_0}\|_{2,1} - \|X_{\mathcal{T}_0^c}\|_{2,1} + \|H_{\mathcal{T}_0^c}\|_{2,1}$$

$$+ \lambda\left(\|X_{r,k}\|_* + \|\overline{H}\|_* - \|\overline{X}_{r,k}\|_* - \|H_{\mathcal{T}_0}^0\|_*\right).$$

Thus, we can write the following bound:

$$\|H_{\mathcal{T}_0^c}\|_{2,1} + \lambda\|\overline{H}\|_* \leq \|H_{\mathcal{T}_0}\|_{2,1} + 2\|X_{\mathcal{T}_0^c}\|_{2,1} + \lambda\|H_{\mathcal{T}_0}^0\|_* + 2\lambda\|\overline{X}_{r,k}\|_*.$$

Dividing both sides by $\sqrt{k}$ yields:

$$\frac{1}{\sqrt{k}}\|\overline{H}_{\mathcal{T}_0^c}\|_{2,1} + \frac{\lambda}{\sqrt{k}}\|\overline{H}\|_* \leq \|H_{\mathcal{T}_0}\|_F + \lambda\sqrt{\frac{2r}{k}}\|H_{\mathcal{T}_0}^0\|_F + 2e + 2\lambda\sqrt{\frac{2r}{k}}e'$$

$$\leq \left(2 + \lambda\sqrt{\frac{2r}{k}}\right)\|H_{\mathcal{T}_0}^0\|_F + \|\overline{H}_{0,1}\|_F + 2e + 2\left(1 + \lambda\sqrt{\frac{2r}{k}}\right)e'.$$

Note that the last inequality uses the bound

$$\|H_{\mathcal{T}_0}\|_F \leq 2\|H_{\mathcal{T}_0}^0\|_F + \|\overline{H}_{0,1}\|_F + 2e',$$

which has been previously developed for obtaining (37). Now, by letting $\lambda = \sqrt{\frac{k}{2r}}$ we achieve the following upper bound:

$$\frac{1}{\sqrt{k}}\|\overline{H}_{\mathcal{T}_0^c}\|_{2,1} + \frac{1}{\sqrt{2r}}\|\overline{H}\|_* \leq 3\|H_{\mathcal{T}_0}^0\|_F + \|\overline{H}_{0,1}\|_F + 4e' + 2e,$$

which gives the same bound as in (38) for sum of the Frobenius norms. All other steps and inequalities that are derived for the proof of Theorem 2-a are valid as well for Theorem 2-c. Therefore, solving (18) with $\lambda = \sqrt{\frac{k}{2r}}$ leads to the same error bound as in (23).

VI. THE PARALLEL PROXIMAL ALGORITHM

Problems (16)-(18) are nonlinearly constrained non-differentiable convex problems. There are numerous methods for solving problems of this kind. We discuss here our approach based on the Parallel Proximal Algorithm (PPXA) proposed in [33]. A remarkable advantage of using this method is the ability of having a parallel implementation. In a nutshell, PPXA is an iterative method for minimizing an arbitrary finite sum of lower semi-continuous (l.s.c.) convex functions, and each iteration consists of computing the *proximity* operators of all functions (which can be done in parallel), averaging their results and updating

---

Algorithm 1: The Parallel Proximal Algorithm to solve (16)-(18)

Input: y, $\mathcal{A}$, the regularization parameters (ε and τ, γ or λ), the convergence parameter $\beta > 0$.

Initializations:

$n = 0$, $X_0 = \Gamma_{1,0} = \Gamma_{2,0} = \Gamma_{3,0} \in \mathbb{R}^{n_1 \times n_2}$

repeat

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

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

    end

    $X_{n+1} = (P_{1,n} + P_{2,n} + P_{3,n})/3$

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

      | $\Gamma_{i,n+1} = \Gamma_{i,n} + 2X_{n+1} - X_n - P_{i,n}$

    end

until *convergence*;

---

the solution until the convergence. The proximity operator of a l.s.c. function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is defined as $prox_f : \mathbb{R}^n \to \mathbb{R}^n$:

$$\underset{\widetilde{x} \in \mathbb{R}^n}{\arg\min}\ f(\widetilde{x}) + \frac{1}{2}\|x - \widetilde{x}\|_{\ell_2}^2, \tag{45}$$

which may have an analytical expression or can be solved iteratively.

In this part we rewrite each of the problems (16)-(17) as a minimization of the sum of three l.s.c. convex functions

$$\underset{X \in \mathbb{R}^{n_1 \times n_2}}{\arg\min}\ f_1(X) + f_2(X) + f_3(X). \tag{46}$$

The template of the PPXA algorithm that solves this minimization is demonstrated in Algorithm 1. In the following we define those functions f_i and we derive their corresponding proximity operators. Note that because f_i is defined on $\mathbb{R}^{n_1 \times n_2}$, the definition of the proximity operator in (45) is naturally extended for matrices by replacing the ℓ_2 norm with the Frobenius norm. Finally, $\beta > 0$ (in Algorithm 1) is a parameter controling the speed of convergence.

A. Constrained ℓ_2/ℓ_1 Norm Minimization

We can rewrite the minimization problem (16) in the form (46) where

$$f_1(X) = \|X\|_{2,1}\ ,\ f_2(X) = i_{\mathcal{B}_{\ell_2}}(X),\ \text{and}\ f_3(X) = i_{\mathcal{B}_*}(X).$$

The convex sets $\mathcal{B}_{\ell_2}, \mathcal{B}_* \subset \mathbb{R}^{n_1 \times n_2}$ correspond to the matrices satisfying the constraints $\|y - \mathcal{A}(X)\|_{\ell_2} \leq \varepsilon$ and $\|X\|_* \leq \tau$, respectively. The indicator function $i_\mathcal{C}$ of a convex set $\mathcal{C}$ is defined as follows:

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

Standard calculations show that $\forall i \in \{1, \ldots, n_1\}$

$$\left(prox_{\alpha f_1}(X)\right)_{i,.} = \frac{\left(\|X_{i,.}\|_{\ell_2} - \alpha\right)_+}{\|X_{i,.}\|_{\ell_2}} X_{i,.} \tag{47}$$

which is the *soft thresholding* operator applied on the rows of X to shrink their ℓ_2 norms. By definition, the proximal operator of an indicator function $i_\mathcal{C}(X)$ is the the orthogonal projection of X onto the corresponding convex set $\mathcal{C}$. Thus, $prox_{\alpha f_2}(X)$ and $prox_{\alpha f_3}(X)$ are respectively the orthogonal projection of X onto the sets $\mathcal{B}_{\ell_2}$ and $\mathcal{B}_*$. For a general implicit operator $\mathcal{A}$, the projector onto $\mathcal{B}_{\ell_2}$ can be computed using an iterative *forward-backward* scheme as proposed in [34]. However, once $\mathcal{A}$ corresponds to a tight frame (*i.e.* $\forall y \in \mathbb{R}^m \ \mathcal{A}(\mathcal{A}^*(y)) = \nu y$ for a constant ν) then according to the *semi-orthogonal linear transform* property of proximity operators [33], the orthogonal projection onto $\mathcal{B}_{\ell_2}$ has the following explicit formulation:

$$prox_{\alpha f_2}(X) = X + \nu^{-1} \mathcal{A}^*(\mathbf{r})\left(1 - \varepsilon\|\mathbf{r}\|_{\ell_2}^{-1}\right)_+ \tag{48}$$

where, $\mathbf{r} = y - \mathcal{A}(X)$.

Finally, let $X = U\Sigma V^T$ be the singular value decomposition of X with $\Sigma = \text{diag}(\sigma_1, \sigma_2 \ldots)$, then the projection of X onto $\mathcal{B}_*$ corresponds to the projection of its singular values onto the following positive simplex (ℓ_1 ball):

$$\mathcal{B}_{\ell_1} = \{\sigma_1, \sigma_2, \ldots, \in \mathbb{R}_+ : \sum_i \sigma_i \leq \tau\}.$$

More precisely, we have

$$prox_{\alpha f_3}(X) = U \, \text{diag}(\overline{\sigma}_1, \ldots, \overline{\sigma}_r) V^T,$$

where $[\overline{\sigma}_1, \overline{\sigma}_2, \ldots] = \mathcal{P}_{\mathcal{B}_{\ell_1}}([\sigma_1, \sigma_2, \ldots])$ and $\mathcal{P}_{\mathcal{B}_{\ell_1}}$ denotes the orthogonal projection of the singular values onto the set $\mathcal{B}_{\ell_1}$. This projection can be computed within few low-complex iterations using the ℓ_1 projection Matlab implementation of the SPGL1 software package [35].

B. Constrained Nuclear Norm Minimization

We rewrite the minimization (17) as (46) where

$$f_1(X) = \|X\|_* , \ f_2(X) = i_{\mathcal{B}_{\ell_2}}(X), \text{ and } f_3(X) = i_{\mathcal{B}_{\ell_2/\ell_1}}(X).$$

The function f_2 and its proximity operator are the same as in the previous part. The convex set $\mathcal{B}_{\ell_2/\ell_1}$ corresponds to the matrices satisfying $\|X\|_{2,1} \leq \gamma$, and $prox_{\alpha f_3}(X)$ is the orthogonal projection of X onto $\mathcal{B}_{\ell_2/\ell_1}$.[2]

The proximity operator of the nuclear norm is the *soft thresholding* shrinkage operator applying on the singular values of X:

$$prox_{\alpha f_1}(X) = U \operatorname{diag}(\overline{\sigma}_1, \overline{\sigma}_2, \ldots)V^T \tag{49}$$

where, $\forall i\ \overline{\sigma}_i = (\sigma_i - \alpha)_+$.

C. Constrained, Nuclear norm plus ℓ_2/ℓ_1 Norm Minimization

Here, we rephrase problem (18) as (46) where the functions in the summation are

$$f_1(X) = \|X\|_{2,1}\ ,\ f_2(X) = i_{\mathcal{B}_{\ell_2}}(X),\ \text{and}\ f_3(X) = \lambda\|X\|_*.$$

According to our previous descriptions, $prox_{\alpha f_2}$ is the orthogonal projection onto the set $\mathcal{B}_{\ell_2}(X)$, $prox_{\alpha f_3}$ is the singular value soft thresholding as in (49) and $prox_{\alpha f_1}$ is equivalent to the soft thresholding of the ℓ_2 norms of the rows of X as in (47).

Note that, for all three above-mentioned implementations using the "economized" singular value decomposition is very useful as the data matrix is assumed to be low-rank and thus it avoids a huge computation at each iteration (particularly for a high-dimensional data).[3] Since the rank r of the solution at each iteration is unknown, at the first iteration of Algorithm 1, r is initially set to some value and later, incremented by doubling its value until achieving the zero singular value in VI-A, or the singular value soft thresholding parameter α in VI-B and VI-C. For the next iterations, r is initialized by the value of the last iteration and so on. This idea has been used in the SVT software package [36] for the singular values soft thresholding and our experiments indicates that in the early iterations of Algorithm 1 and within at most 2 or 3 sub-iterations, r is set to its correct value, and remains fixed for the forthcoming iterations.

---

[2]The Matlab code to perform such projection can be found in the SPGL1 package and is a natural extension of the vector projection onto the ℓ_1 ball.

[3]The "economized" singular value decomposition of a rank r matrix, here, refers to computing only the r largest (*i.e.* nonzero) singular values.

VII. NUMERICAL EXPERIMENTS

In this section we conduct a series of experiments in order to demonstrate the average behavior of the proposed simultaneous low-rank and joint-sparse recovery approach for different problem setups, *i.e.* different matrix sizes $n_1 \times n_2$, joint-sparsity levels k, ranks r and number of CS measurements m. Among the three formulations suggested in this paper, we choose the nuclear norm plus the ℓ_2/ℓ_1 mixed-norm minimization scheme in (18) (referred to as `Nuclear`$-\ell_2/\ell_1$) as it requires the minimum prior knowledge about the underlying data, and the regularization parameter is set to $\lambda = \sqrt{\frac{k}{2r}}$ as suggested by Theorem 2-c. A rank r and k-joint sparse matrix $X \in \mathbb{R}^{n_1 \times n_2}$ is generated by multiplying two independent i.i.d. Gaussian matrices $X_1 \in \mathbb{R}^{k \times r}$ and $X_2 \in \mathbb{R}^{r \times n_2}$, and populating uniformly at random k rows of X by the rows of $X_1 X_2$. The figures that are used for performance evaluation demonstrate the empirical reconstruction error phase transition: for a certain problem setup (n_1, n_2, k, r, m), the normalized reconstruction error $\|X - \widehat{X}\|_F / \|X\|_F$ has been averaged over twenty independent random realizations of X and $\mathcal{A}$. Figures are plotted in gray scale and white pixels correspond to high reconstruction errors while black pixels are indicating the recovery regions.

Figure 1(a) demonstrates the performance of our proposed scheme at recovering 40×40 random matrices of rank $r = 2$, and for different joint-sparsity ratios ($\frac{k}{n_1}$) and subsampling factors ($\frac{m}{n_1 n_2}$). A dense i.i.d. random Gaussian sampling scheme $\mathcal{A}$ has been applied for compressed sampling of data. We can observe that for recovering data matrices with high joint-sparsity ratios more CS measurements are required. Meanwhile, it turns out that `Nuclear`$-\ell_2/\ell_1$ is able to recover even dense (non-sparse) data matrices (*i.e.*, $k = n_1$) subsampled by retaining less than 40% of their original sizes. This should not come as a surprise : despite loosing sparsity, `Nuclear`$-\ell_2/\ell_1$ still takes advantage of data correlations due to the low-rank structure, which ultimately results in a successful reconstruction for an oversampling factor nearly four times the degrees of freedom of the original data.

Figure 1(b) illustrates the error phase transition behavior of the ℓ_2/ℓ_1 minimization (10) for the same setup. As we expect, for such highly structured data, our proposed scheme significantly outperforms the ℓ_2/ℓ_1 approach by having a much wider recovery region.

Figures 1(c) and 1(d) repeat similar experiments but for rank $r = 4$ matrices. The recovery region of the joint `Nuclear`$-\ell_2/\ell_1$ norm minimization is narrowed compared to the Figure 1(a) as more measurements are required to compensate higher degrees of freedom of data in this scenario. However, this method still keeps outperforming the ℓ_2/ℓ_1 recovery approach.
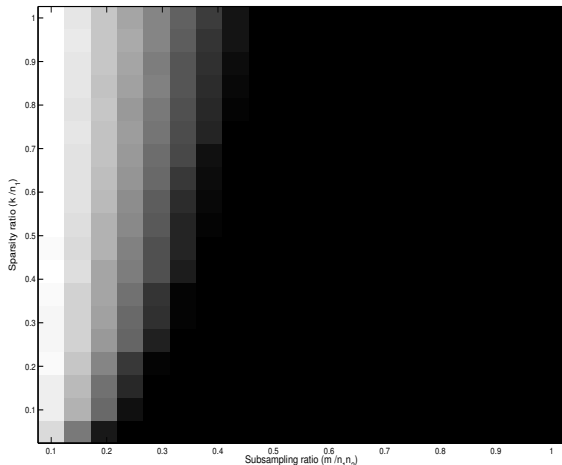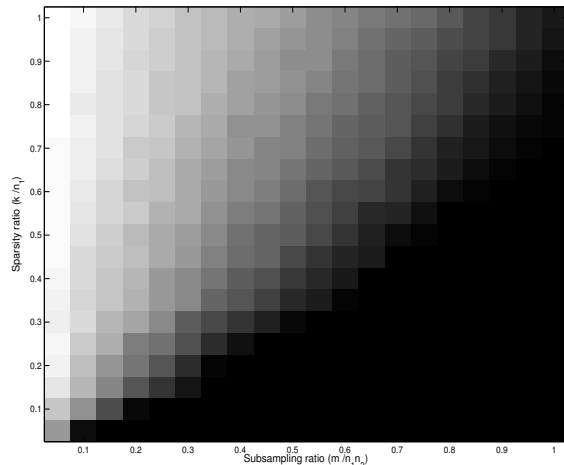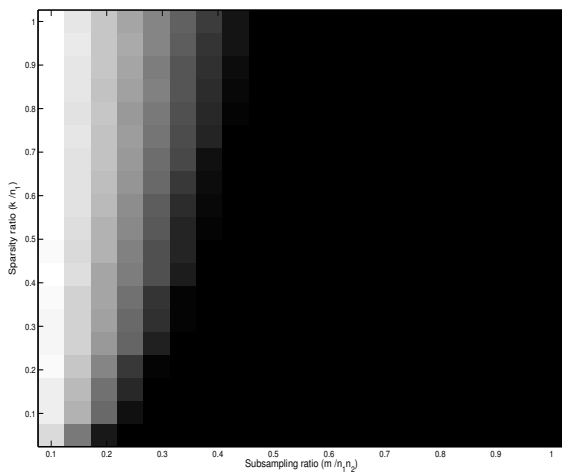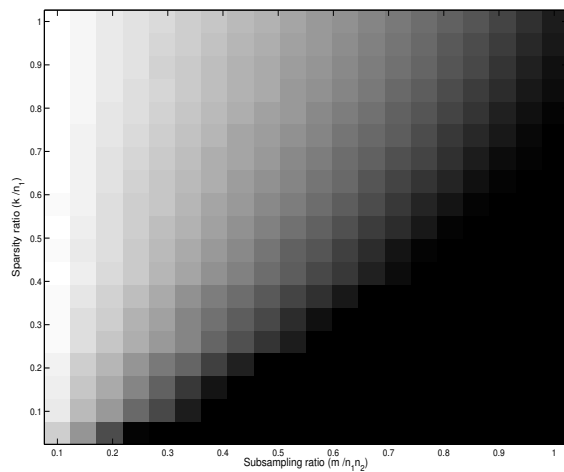
(a) Nuclear-ℓ_2/ℓ_1, $r = 2$

(b) ℓ_2/ℓ_1 minimization, $r = 2$

(c) Nuclear-ℓ_2/ℓ_1, $r = 4$

(d) ℓ_2/ℓ_1 minimization, $r = 4$

Fig. 1.   Reconstruction error phase transitions for 40×40 matrices of rank r acquired by a dense i.i.d. Gaussian sampling matrix: Sparsity ratio (k/n_1) vs. Subsampling ratio $(\frac{m}{n_1 n_2})$.

Reconstruction Using Distributed Sampling Schemes

In another experimental setup, we demonstrate the influence of different sampling schemes on our recovery approach. Besides *dense sampling*, corresponding to the use of dense i.i.d. random Gaussian matrices A in (7), we use the two following schemes particularly suited for distributed sensing applications:

- *Distributed Uniform Sampling*: In this case all columns of the data matrix are sampled by a *unique* i.i.d. Gaussian random matrix $\widetilde{A} \in \mathbb{R}^{\widetilde{m} \times n_1}$ and thus the sampling operator $\mathcal{A}$ corresponds to a block
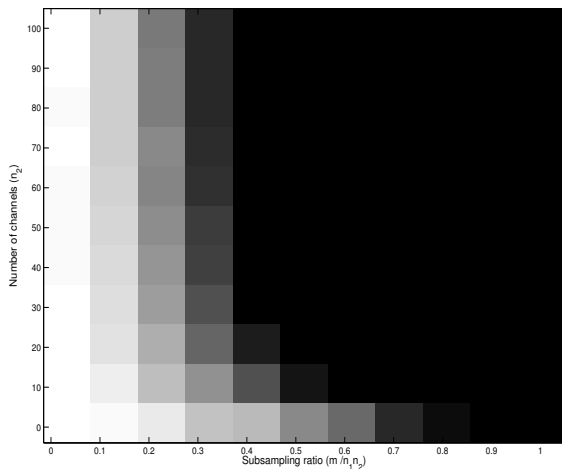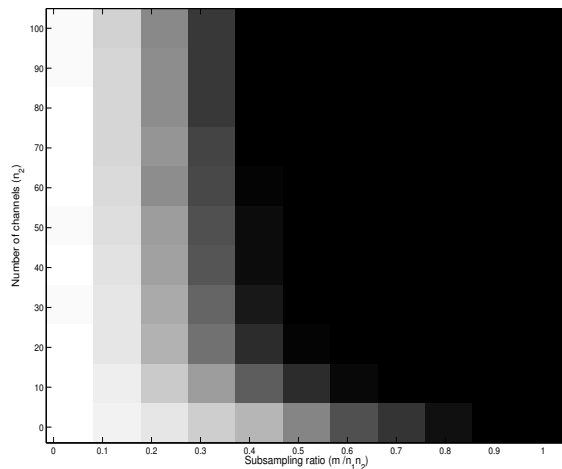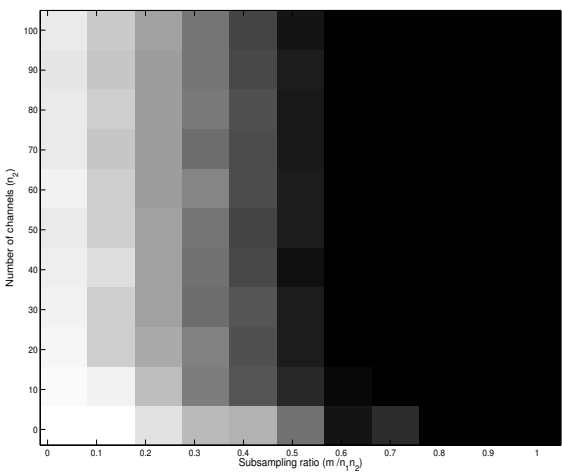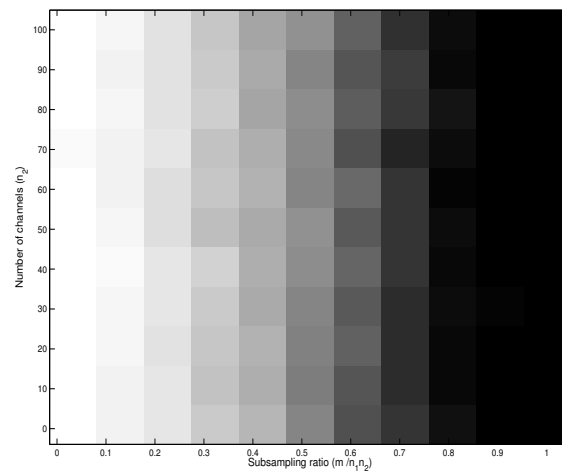
(a) Nuclear-ℓ_2/ℓ_1, dense sampling

(b) Nuclear-ℓ_2/ℓ_1, distributed independent sampling

(c) ℓ_2/ℓ_1 minimization, dense sampling

(d) Nuclear-ℓ_2/ℓ_1, distributed uniform sampling

Fig. 2. Reconstruction error phase transitions for $30 \times n_2$ matrices having $r = 3$ and $k = 10$: Number of columns (n_2) vs. Subsampling ratio $\left(\frac{m}{n_1 n_2}\right)$.

diagonal matrix given below:

$$A = \begin{bmatrix} \widetilde{A} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \widetilde{A} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \widetilde{A} \end{bmatrix}, \tag{50}$$

where $\widetilde{m} = m/n_2$ denotes the number of measurements collected per column. As previously mentioned in Section II-5, this sensing model is analogous to the multiple measurement vectors (MMV) model. For multichannel signal applications such as sensor networks where each column of

the data matrix represents the observation of a sensor node, this scheme enables all nodes to use a unified mechanism for sampling their own *local* observations in a distributed manner (*i.e.* sampling without intra-channel collaboration).

- *Distributed Independent Sampling*: In this case, an independent random Gaussian matrix $\widetilde{A}_j \in \mathbb{R}^{\widetilde{m} \times n_1}$ is used to collect $\widetilde{m}$ samples from each column j of the data matrix, and thus the sampling operator corresponds to a block diagonal matrix with distinct blocks:

$$A = \begin{bmatrix} \widetilde{A}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \widetilde{A}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \widetilde{A}_{n_2} \end{bmatrix}, \tag{51}$$

where $\widetilde{m} = m/n_2$. For multichannel sensing applications with strong intra channel correlations, this sampling strategy gives the freedom of designing the blocks so that the measurements of different channels carry diverse and non-redundant information to the decoder.

Figures 2(a)-2(d) demonstrate the reconstruction error phase transitions for various numbers of columns n_2 and CS measurements m. Data matrices of size $30 \times n_2$ with a fixed rank $r = 3$ and joint-sparsity level $k = 10$ are drawn at random and subsampled by dense, distributed independent or distributed uniform random sampling matrices. These graphs suggest a useful tradeoff for practical multichannel sensing setups, *e.g.* in sensor network applications, in order to design low cost/complex sensors but distribute many of them (high number of channels n_2). As we can see in Figure 2(a), the recovery region of $\texttt{Nuclear}-\ell_2/\ell_1$ using dense sampling matrices follows the behavior predicted by the theoretical analysis (see Corollaries 1 and 2) *i.e.*,

$$\frac{m}{n_1 \, n_2} \gtrsim \mathcal{O}\left(\frac{k \log(n_1/k) + kr}{n_1 \, n_2} + \frac{r}{n_1}\right),$$

that is a constant plus a decaying term. A very similar recovery performance can be observed in Figure 2(b) using distributed independent sampling. In both cases by increasing n_2, $\texttt{Nuclear}-\ell_2/\ell_1$ requires less CS measurements per column (up to a minimal measurement limit).

Figure 2(c) demonstrates the performance of the ℓ_2/ℓ_1 recovery scheme using dense random sampling matrices. Compared to $\texttt{Nuclear}-\ell_2/\ell_1$, the recovery region shrinks and more remarkably for high values of n_2, the minimal measurement limit becomes larger *i.e.* more measurements per column is required. As a result, for multichannel sensing applications $\texttt{Nuclear}-\ell_2/\ell_1$ achieves better tradeoff between the number of sensor nodes and their complexity compared to ℓ_2/ℓ_1 recovery approach.

Finally, we repeat the same experiment for the $\texttt{Nuclear}-\ell_2/\ell_1$ approach and using distributed uniform random sampling matrices. We can observe in Figure 2(d) that by adding extra channels recovery does

not improve much, which is consistent with our early explanations: The uniform sampling is not a proper strategy for CS acquisition of low-rank data matrices because the resulting measurements become highly redundant and they do not carry diverse information across channels.

VIII. CONCLUSION

In this paper we showed that combining the two key priors in the fields of compressive sampling and matrix completion, namely joint-sparsity and low-rank structures, can tremendously enhance CS dimensionality reduction performance. We proposed a joint $\texttt{Nuclear-}\ell_2/\ell_1$ convex minimization approach that is able to efficiently exploit such restrictive structure. Our theoretical analysis indicates that the $\texttt{Nuclear-}\ell_2/\ell_1$ scheme is robust against measurement noise and non-exact low-rank and joint-sparse data matrices. It can achieve a tight lower bound on the number of CS measurements improving significantly the state-of-the-art. We provided algorithms to solve this type of minimization problem enabling us to validate these improvements by several numerical experiments.

Extension of this work includes theoretical analysis of the $\texttt{Nuclear-}\ell_2/\ell_1$ approach for structured random matrices (*e.g.*, distributed independent random sampling matrices), as well as considering alternative structures such as simultaneous low-rank, sparse (rather than joint-sparse) and positive semidefinite matrices which may find applications in high-dimensional sparse covariance matrix estimation. We are also interested in further applications of this approach in multichannel signal compressive sampling.

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] E. Candès, "Compressive sampling," in *Proceedings of the International Congress of Mathematicians*, vol. 3.   Citeseer, 2006, p. 14331452.

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

[5] S. S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[6] M. Fazel, H. Hindi, and S. Boyd, "Rank minimization and applications in system theory," in *Proc. American Control Conference*, 2004, pp. 3273–3278.

[7] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009-12-01. [Online]. Available: http://dx.doi.org/10.1007/s10208-009-9045-5

[8] E. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925 –936, june 2010.

[9] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *Information Theory, IEEE Transactions on*, vol. 56, no. 6, pp. 2980 –2998, june 2010.

[10] M. Golbabaee and P. Vandergheynst, "Guaranteed recovery of a low-rank and joint-sparse matrix from incomplete and noisy measurement," in *Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS11)*, 2011.

[11] ——, "Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[12] ——, "Joint Trace/TV norm minimization: A new efficient approach for spectral compressive imaging," in *IEEE International Conference on Image Processing (ICIP)*, 2012.

[13] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, 2011.

[14] A. Ganesh, K. Min, J. Wright, and Y. Ma, "Principal component pursuit with reduced linear measurements," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, july 2012, pp. 1281 –1285.

[15] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[16] A. d Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, "A direct formulation for sparse pca using semidefinite programming," *SIAM Review*, vol. 49, no. 3, pp. 434–448, 2007.

[17] A. d'Aspremont, F. Bach, and L. El Ghaoui, "Optimal solutions for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 9, pp. 1269–1294, 2008.

[18] I. Jolliffe, *Principal component analysis*. Springer Verlag, New York, 1986.

[19] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[20] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572 – 588, 2006.

[21] M. Duarte, S. Sarvotham, D. Baron, M. Wakin, and R. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Asilomar Conf. Signals, Sys., Comput*, 2005, pp. 1537–1541.

[22] M. Golbabaee and P. Vandergheynst, "Distributed Compressed Sensing for Sensor Networks Using Thresholding," in *Proceedings of the SPIE 2009*, vol. 7446, 2009, pp. 7446–50, invited Paper.

[23] J. Tropp, "Algorithms for simultaneous sparse approximation. part II: Convex relaxation," *Signal Processing*, vol. 86, no. 3, pp. 589 – 602, 2006.

[24] Y. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *Information Theory, IEEE Transactions on*, vol. 56, no. 1, pp. 505 –519, jan. 2010.

[25] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *Information Theory, IEEE Transactions on*, vol. 55, no. 11, pp. 5302 –5316, nov. 2009.

[26] E. Candes and Y. Plan, "Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements." *IEEE Trans. on Inf. Theory*, 2009.

[27] M. Davies and Y. Eldar, "Rank awareness in joint sparse recovery," *Information Theory, IEEE Transactions on*, vol. 58, no. 2, pp. 1135 –1146, feb. 2012.

[28] J. Kim, O. Lee, and J. Ye, "Compressive music: A missing link between compressive sensing and array signal processing," *CoRR*, vol. abs/1004.4398, 2010.

[29] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *Signal Processing, IEEE Transactions on*, vol. 53, no. 7, pp. 2477 – 2488, july 2005.

[30] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *Signal Processing, IEEE Transactions on*, vol. 54, no. 12, pp. 4634 –4643, dec. 2006.

[31] Y. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *Signal Processing, IEEE Transactions on*, vol. 58, no. 6, pp. 3042 –3054, june 2010.

[32] B. Natarajan, "Sparse Approximate Solutions to Linear Systems," *SIAM Journal on Computing*, vol. 24, p. 227, 1995.

[33] 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.

[34] M. J. Fadili and J. L. Starck, "Monotone operator splitting for optimization problems in sparse recovery," in *ICIP*, 2009, pp. 1461–1464.

[35] "http://www.cs.ubc.ca/labs/scl/spgl1/."

[36] "http://www-stat.stanford.edu/ candes/svt/."