Joint Sensing Matrix and Sparsifying Dictionary Optimization for Tensor Compressive Sensing

Xin Ding, *Student Member, IEEE*, Wei Chen, *Member, IEEE*, and Ian J. Wassell

***Abstract*—Tensor Compressive Sensing (TCS) is a multidimensional framework of Compressive Sensing (CS), and it is advantageous in terms of reducing the amount of storage, easing hardware implementations and preserving multidimensional structures of signals in comparison to a conventional CS system. In a TCS system, instead of using a random sensing matrix and a predefined dictionary, the average-case performance can be further improved by employing an optimized multidimensional sensing matrix and a learned multilinear sparsifying dictionary. In this paper, we propose an approach that jointly optimizes the sensing matrix and dictionary for a TCS system. For the sensing matrix design in TCS, an extended separable approach with a closed form solution and a novel iterative non-separable method are proposed when the multilinear dictionary is fixed. In addition, a multidimensional dictionary learning method that takes advantages of the multidimensional structure is derived, and the influence of sensing matrices is taken into account in the learning process. A joint optimization is achieved via alternately iterating the optimization of the sensing matrix and dictionary. Numerical experiments using both synthetic data and real images demonstrate the superiority of the proposed approaches.**

***Keywords*—Multidimensional system, compressive sensing, tensor compressive sensing, dictionary learning, sensing matrix optimization.**

I. Introduction

The traditional signal acquisition-and-compression paradigm removes the signal redundancy and preserves the essential contents of signals to achieve savings on storage and transmission, where the minimum sampling ratio is restricted by the Shannon-Nyquist Theorem at the signal sampling stage. The wasteful process of sensing-then-compressing is replaced by directly acquiring the compressed version of signals in Compressive Sensing (CS) [1]–[3], a new sampling paradigm that leverages the fact that most signals have sparse representations (i.e., there are only a few non-zero coefficients) in some suitable basis. Successful reconstruction of such signals is guaranteed for a sufficient number of randomly taken samples that are far fewer in number than that required in the Shannon-Nyquist Theorem. Therefore CS is very attractive for applications such as medical imaging and wireless sensor networks where data acquisition is expensive [4], [5].

Xin Ding and Ian J. Wassell are with the Computer Lab, University of Cambridge, UK (e-mail: xd225, ijw24@cam.ac.uk).

Wei Chen is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China (e-mail: weich@bjtu.edu.cn).

Corresponding author: Wei Chen.

This work is supported by EPSRC Research Grant EP/K033700/1; the Natural Science Foundation of China (61671046,61401018, U1334202); the State Key Laboratory of Rail Traffic Control and Safety (RCS2016ZT014) of Beijing Jiaotong University.

Successful CS reconstruction has been achieved by characterizing a number of properties of the sampling operator, e.g., the Restricted Isometry Property (RIP) [1], the mutual coherence [6] and the null space property [2]. These properties have been used to provide sufficient conditions on the equivalent sensing matrix (i.e., the product of the sensing matrix and the sparsifying basis) and to quantify the worst-case reconstruction performance [2], [6], [7]. Random matrices such as Gaussian or Bernoulli matrices have been shown to fulfill these conditions for any given sparsifying basis, i.e., a dictionary, and hence are widely used as the sensing matrix in CS applications.

In view of the fact that the mainstream view in the signal processing community considers the average-case performance rather than the worst-case performance, it is shown in [8]–[12] that the average-case reconstruction performance can be enhanced by optimizing the sensing matrix such that the sensing matrix is more incoherent with the given sparsifying basis than a random sensing matrix. Also, instead of using a fixed signal-sparsifying basis, e.g., a Discrete Wavelet Transform (DWT), one can further enhance CS performance by employing a basis to abstract the basic atoms that compose the signal ensemble. The process of learning such a basis is referred to as "sparsifying dictionary learning" and it has been widely investigated in the literature [13]–[17]. Considering that the sensing matrix and the dictionary jointly affect the performance of a CS system, applying only traditional dictionary learning in CS applications without considering the sensing matrix may lead to a non-optimal equivalent sensing matrix. In [18]–[20], joint optimization of the sensing matrix and the dictionary are investigated, demonstrating that it is beneficial to consider this joint approach in practical CS systems.

Multidimensional signals are mapped in vector format in the process of sensing and reconstruction of a CS system, because the conventional CS framework considers only vectorized signals. At the sensing node, such a vectorization requires the hardware to be capable of simultaneously multiplexing along all data dimensions, which is hard to achieve especially when one of the dimensions is along a timeline. Secondly, a real-world vectorized signal requires an enormous sensing matrix that has as many columns as the number of signal elements. Consequently such an approach imposes large demands on storage and processing power. In addition, the vectorization also results in a loss of structure along the various dimensions (e.g., rank deficient along one dimension while sparse along another [21]), the presence of which is beneficial for developing efficient reconstruction algorithms (see [21], [22]). For these reasons, applying conventional CS to applications that involve multidimensional signals is challenging.

Extending CS to multidimensional signals has attracted growing interest over the past few years. Most of the related work in the literature focuses on CS for 2D signals (i.e., matrices), e.g., matrix completion [23], [24], and the reconstruction of sparse and low rank matrices [25]–[27].

In [28], Kronecker product matrices are proposed for use in CS systems, which makes it possible to partition the sensing process along signal dimensions and paves the way to develop CS for tensors, i.e., signals with two or more dimensions. In [29], separable imaging has been studied, which demonstrated that the Tensor CS (TCS) model can ease the optical designs, data storage and computation complexity for 2D images. The optical hardware design using the TCS model for images has been investigated in [30]. There are many other applications that can benefit from TCS, e.g., hyperspectral imaging, video acquisition and distributed sensing, as detailed in [31]. TCS reconstruction has been studied in [32]–[35]. To the best of our knowledge, there is no prior work concerning the enhancement of TCS via optimizing the sensing matrices at various dimensions in a tensor. In addition, although dictionary learning techniques have been considered for tensors [36]–[38], it is still not clear how to conduct tensor dictionary learning to incorporate the influence of sensing matrices in TCS.

Following the train of thought in CS, in this paper, we aim to improve the average-case performance of TCS by optimizing both the sensing matrix and the dictionary. Furthermore, considering that the sensing matrix and the dictionary are also coupled in TCS systems, we will investigate their joint design. Unlike the optimization for a conventional CS system where a single sensing matrix and a sparsifying basis for vectorized signals are obtained, we estimate multiple sensing matrices and sparsifying bases, each acting along the various dimensions of the tensor, thereby maintaining the advantages of TCS. The Kronecker structure that is involved in a TCS system makes the optimization problem more challenging than the conventional design problem in a CS system, and it also leads to opportunities for us to develop novel design approaches. The contributions of this work are as follows:

- We are the first to consider the optimization of a multidimensional sensing matrix and dictionary for a TCS system, and we design a joint optimization of the two, which also includes particular cases of optimizing the sensing matrix for a given multilinear dictionary and learning the dictionary for a given multidimensional sensing matrix.
- We propose a separable approach for sensing matrix design by extending the existing work for conventional CS. We prove that the resulting optimization is separable, i.e., the sensing matrix along each dimension can be independently optimized, and the approach has a closed form solution.
- We put forth a non-separable method for sensing matrix design using a combination of the state-of-art measures for sensing matrix optimization. This approach leads to the best reconstruction performance in our comparison, but it is iterative and hence needs more computing power to implement.
- We propose a multidimensional dictionary learning approach that couples the optimization of the multidimensional sensing matrix. This approach extends KSVD [14] and coupled-KSVD [18] to take full advantage of the multidimensional structure in tensors with a reduced number of iterations required for the update of dictionary atoms.

The proposed approaches are demonstrated to enhance the performance of existing TCS systems via the use of extensive simulations using both synthetic data and real images. For real-world applications involving multidimensional signals, such as 2D imaging, hyperspectral imaging and video acquisition, the proposed approaches can be employed to design the optimized multidimensional sensing schemes and sparsifying basis, or approximations thereof, to subsample and sparsify these signals. Specifically, in hyperspectral imaging for instance, one can design the structure of Digital Micromirror Device (DMD) to achieve optimized sampling in the spatial dimensions, while designing the coded aperture for the optimized sensing across the spectra [39], and both spatial and spectral dictionaries can be learnt using the proposed method.

The remainder of this paper is organized as follows. Section II formulates CS and TCS, and introduces the related theory. Section III reviews the sensing matrix design approaches for CS and presents the proposed methods for TCS sensing matrix design. In Section IV, we review the related dictionary learning techniques, followed by the elaboration of the proposed multidimensional dictionary learning approach and we present the joint optimization algorithm. Experimental results are given in Section V and Section VI concludes the paper.

A. Multilinear Algebra and Notation

Boldface lower-case letters, boldface upper-case letters and non-boldface letters denote vectors, matrices and scalars, respectively. A mode-n tensor is an n-dimensional array $\underline{\mathbf{X}} \in \mathbb{R}^{N_1 \times \ldots \times N_n}$. The mode-$i$ vectors of a tensor are determined by fixing every index except the one in the mode i and the slices of a tensor are its two dimensional sections determined by fixing all but two indices. By arranging all the mode-i vectors as columns of a matrix, the mode-i unfolding matrix $\mathbf{X}_{(i)} \in \mathbb{R}^{N_i \times N_1 \ldots N_{i-1} N_{i+1} \ldots N_n}$ is obtained. The operator $fold_i(\cdot)$ is defined as the opposite operation of unfolding such that: $fold_i(\mathbf{X}_{(i)}) = \underline{\mathbf{X}}$, where the dimensions of the tensor are assumed to be known. The mode-k tensor by matrix product is defined as: $\underline{\mathbf{Z}} = \underline{\mathbf{X}} \times_k \mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{J \times N_k}$, $\underline{\mathbf{Z}} \in \mathbb{R}^{N_1 \times \ldots \times N_{k-1} \times J \times N_{k+1} \times \ldots \times N_n}$ and it is calculated by: $\underline{\mathbf{Z}} = fold_i(\mathbf{Z}_{(i)}) = fold_i(\mathbf{A}\mathbf{X}_{(i)})$, where $\mathbf{Z}_{(i)}$ is the mode i unfolding of $\underline{\mathbf{Z}}$ and $\mathbf{Z}_{(i)} = \mathbf{A}\mathbf{X}_{(i)}$. The matrix Kronecker product and vector outer product are denoted by $\mathbf{A} \otimes \mathbf{B}$ and $\mathbf{a} \circ \mathbf{b}$, respectively. The l_p norm of a vector is defined as: $||\mathbf{x}||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$. For vectors, matrices and tensors, the l_0 norm is given by the number of nonzero entries. $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. The operator $(\cdot)^{-1}$, $(\cdot)^T$ and $tr(\cdot)$ represent matrix inverse, matrix transpose and the trace of a matrix, respectively. The number of elements for a vector, matrix or tensor is denoted by $len(\cdot)$.

II. Compressive Sensing (CS) and Tensor Compressive Sensing (TCS)

A. Sensing Model

Consider a multidimensional signal $\underline{\mathbf{X}} \in \mathbb{R}^{N_1 \times ... \times N_n}$. Conventional CS takes measurements from its vectorized version via:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ $(N = \prod_i N_i)$ denotes the vectorized signal, $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ $(M < N)$ is the sensing matrix, $\mathbf{y} \in \mathbb{R}^M$ represents the measurement vector and $\mathbf{e} \in \mathbb{R}^M$ is a noise term. The vectorized signal is assumed to be sparse in some sparsifying basis $\mathbf{\Psi} \in \mathbb{R}^{N \times \hat{N}}$ $(N \leq \hat{N})$, i.e.,

$$\mathbf{x} = \mathbf{\Psi}\mathbf{s}, \tag{2}$$

where $\mathbf{s} \in \mathbb{R}^{\hat{N}}$ is the sparse representation of $\mathbf{x}$ and it has only K $(K \ll \hat{N})$ non-zero coefficients. Thus the sensing model can be rewritten as:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{\Psi}\mathbf{s} + \mathbf{e} = \mathbf{A}\mathbf{s} + \mathbf{e}, \tag{3}$$

where $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi} \in \mathbb{R}^{M \times \hat{N}}$ is the equivalent sensing matrix.

Even though CS has been successfully applied to practical sensing systems [40]–[42], the sensing model has a few drawbacks when it comes to tensor signals [28]. First of all, the multidimensional structure presented in the original signal $\underline{\mathbf{X}}$ is omitted due to the vectorization, which loses information that can lead to efficient reconstruction algorithms [21]. Besides, as stated by (1), the sensing system is required to operate along all dimensions of the signal simultaneously, which is difficult to achieve in practice [29], [30]. Furthermore, the size of $\mathbf{\Phi}$ associated with the vectorized signal becomes too large to be practical for applications involving multidimensional signals.

TCS tackles these problems by utilizing separable sensing operators along tensor modes and its sensing model is:

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_1 \mathbf{\Phi}_1 \times_2 \mathbf{\Phi}_2 ... \times_n \mathbf{\Phi}_n + \underline{\mathbf{E}}, \tag{4}$$

where $\underline{\mathbf{Y}} \in \mathbb{R}^{M_1 \times ... M_n}$ represents the measurement, $\underline{\mathbf{E}} \in \mathbb{R}^{M_1 \times ... M_n}$ denotes the noise term, $\mathbf{\Phi}_i \in \mathbb{R}^{M_i \times N_i}$ $(i = 1, ..., n)$ are sensing matrices and $M_i < N_i$. The multidimensional signal is assumed to be sparse in a separable sparsifying basis $\mathbf{\Psi}_i \in \mathbb{R}^{N_i \times \hat{N}_i}$ $(i = 1, ..., n)$, i.e.,

$$\underline{\mathbf{X}} = \underline{\mathbf{S}} \times_1 \mathbf{\Psi}_1 \times_2 \mathbf{\Psi}_2 ... \times_n \mathbf{\Psi}_n, \tag{5}$$

where $\underline{\mathbf{S}} \in \mathbb{R}^{\hat{N}_1 \times ... \hat{N}_n}$ is the sparse representation that has only K $(K \ll \prod_i \hat{N}_i)$ non-zero coefficients. The equivalent sensing model can then be written as:

$$\underline{\mathbf{Y}} = \underline{\mathbf{S}} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 ... \times_n \mathbf{A}_n + \underline{\mathbf{E}}, \tag{6}$$

where $\mathbf{A}_i = \mathbf{\Phi}_i\mathbf{\Psi}_i$ $(i = 1, ..., n)$ are the equivalent sensing matrices.

Using the TCS sensing model in (4), the sensing procedure in (1) is partitioned into a few processes having smaller sensing matrices $\mathbf{\Phi}_i \in \mathbb{R}^{M_i \times N_i}$ $(i = 1, ..., n)$ and yet it maintains the multidimensional structure of the original signal $\underline{\mathbf{X}}$. Regarding the hardware implementation of a TCS system, if the data along all dimensions exists at the same time, one may design a particular optical system to implement the sensing procedure, e.g., a 2D separable imaging scheme has been developed in [30]. Otherwise, if the data along some of the dimensions are not available at the same time, e.g., as for the time-dependent signals, the sensing procedure can still be achieved progressively (as detailed in [28]) due to the partitioned sensing nature of TCS.

It is also useful to mention that the TCS model in (6) is equivalent to:

$$\mathbf{y} = (\mathbf{A}_n \otimes \mathbf{A}_{n-1} \otimes ... \otimes \mathbf{A}_1)\mathbf{s} + \mathbf{e}, \tag{7}$$

as derived in [34]. By denoting $\overline{\mathbf{A}} = \mathbf{A}_n \otimes \mathbf{A}_{n-1} \otimes ... \otimes \mathbf{A}_1$, it becomes a conventional CS model akin to (3), except that the sensing matrix in (7) has a multilinear structure.

B. Signal Reconstruction

In conventional CS, the problem of reconstructing $\mathbf{s}$ from the measurement vector $\mathbf{y}$ captured using (3) is modeled as a l_0 minimization problem as follows:

$$\min_{\mathbf{s}} ||\mathbf{s}||_0, \ s.t. \ ||\mathbf{y} - \mathbf{A}\mathbf{s}|| \leq \varepsilon, \tag{8}$$

where ε is a tolerance parameter. Many algorithms have been developed to approximate the solution of this problem, including Basis Pursuit (BP) [1]–[3], [43], i.e., conducting convex optimization by relaxing the l_0 norm in (8) as the l_1 norm, and greedy algorithms such as Orthogonal Matching Pursuit (OMP) [44] and Iterative Hard Thresholding (IHT) [45]. The reconstruction performance of the l_1 minimization approach has been studied in [7], [43], where the well known Restricted Isometry Property (RIP) was introduced to provide a sufficient condition for successful signal recovery.

Definition 1: A matrix $\mathbf{A}$ satisfies the RIP of order K with Restricted Isometry Constant (RIC) δ_K being the smallest number such that

$$(1 - \delta_K)||\mathbf{s}||_2^2 \leq ||\mathbf{A}\mathbf{s}||_2^2 \leq (1 + \delta_K)||\mathbf{s}||_2^2 \tag{9}$$

holds for all $\mathbf{s}$ with $||\mathbf{s}||_0 \leq K$. ■

Theorem 1: Assume that $\delta_{2K} < \sqrt{2} - 1$ and $||\mathbf{e}||_2 \leq \varepsilon$. Then the solution $\hat{\mathbf{s}}$ to (8) obeys

$$||\hat{\mathbf{s}} - \mathbf{s}||_2 \leq C_0 K^{-1/2}||\mathbf{s} - \mathbf{s}_K||_1 + C_1\varepsilon \tag{10}$$

where $C_0 = \frac{2+(2\sqrt{2}-2)\delta_{2K}}{1-(\sqrt{2}+1)\delta_{2K}}$, $C_1 = \frac{4\sqrt{1+\delta_{2K}}}{1-(\sqrt{2}+1)\delta_{2K}}$, δ_{2K} is the RIC of matrix $\mathbf{A}$, $\mathbf{s}_K$ is an approximation of $\mathbf{s}$ with all but the K largest entries set to zero. ■

The previous theorem states that for the noiseless case, any sparse signal with fewer than K non-zero coefficients can be exactly recovered if the RIC of the equivalent sensing matrix satisfies $\delta_{2K} < \sqrt{2} - 1$; while for the noisy case and the not exactly sparse case, the reconstructed signal is still a good approximation of the original signal under the same condition. The theoretical guarantees of successful reconstruction for the greedy approaches have also been investigated in [44], [45].

The RIP essentially measures the quality of the equivalent sensing matrix $\mathbf{A}$, which closely relates to the design of $\mathbf{\Phi}$ and $\mathbf{\Psi}$. However, since the RIP condition is not computationally

tractable to verify for a given sensing matrix, another measure is often used for CS projection design, i.e., the mutual coherence of $\mathbf{A}$ [6] and it is defined by:

$$\mu(\mathbf{A}) = \max_{1\leq i,\ j\leq \hat{N},\ i\neq j} |\mathbf{a}_i^T \mathbf{a}_j|, \tag{11}$$

where $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$, and it is assumed to be normalized such that $\mathbf{a}_i^T \mathbf{a}_i = 1$. It has been shown that the reconstruction error of the l_1 minimization problem is bounded if $\mu(\mathbf{A}) < 1/(4K-1)$.

We can see that the mutual coherence reveals the degree of orthogonality between columns of $\mathbf{A}$. A smaller value of mutual coherence indicates that the equivalent sensing matrix $\mathbf{A}$ is closer to an orthogonal matrix, in which case the signal reconstruction becomes easier. In addition, we can see that the largest off-diagonal entries of the Gram matrix of $\mathbf{A}$, i.e., $\mathbf{G_A} = \mathbf{A}^T\mathbf{A}$, provides the value of the mutual coherence of $\mathbf{A}$. Therefore, based on this concept, optimal projection design approaches are derived, e.g., in [8], [9], [18].

When it comes to TCS, the reconstruction approaches for CS can still be utilized owing to the relationship in (7). However, for the algorithms where explicit usage of $\overline{\mathbf{A}}$ is required, e.g, OMP, the implementation is restricted by the large dimension of $\overline{\mathbf{A}}$. By extending the CS reconstruction approaches to utilize tensor-based operations, TCS reconstruction algorithms employing only small matrices $\mathbf{A}_i$ $(i = 1, ..., n)$ have been developed in [29], [30], [34], [35]. These methods maintain the theoretical guarantees of conventional CS when $\overline{\mathbf{A}}$ obeys the condition on the RIC or the mutual coherence, but reduce the computational complexity and relax the storage memory requirement.

Even so, the conditions on $\overline{\mathbf{A}}$ are not intuitive for a practical TCS system, which explicitly utilizes multiple separable sensing matrices $\mathbf{A}_i$ $(i = 1, ..., n)$ instead of a single matrix $\overline{\mathbf{A}}$. Fortunately, the authors of [28] have derived the following relationships to clarify the corresponding conditions on $\mathbf{A}_i$ $(i = 1, ..., n)$.

Theorem 2: Let $\mathbf{A}_i$ $(i = 1, ..., n)$ be matrices with RICs $\delta_K(\mathbf{A}_1)$, ..., $\delta_K(\mathbf{A}_n)$, respectively, and their mutual coherence are $\mu(\mathbf{A}_1)$, ..., $\mu(\mathbf{A}_n)$. Then for the matrix $\overline{\mathbf{A}} = \mathbf{A}_n \otimes \mathbf{A}_{n-1} \otimes ... \otimes \mathbf{A}_1$, we have

$$\mu(\overline{\mathbf{A}}) = \max[\mu(\mathbf{A}_1), ..., \mu(\mathbf{A}_n)], \tag{12}$$

$$\delta_K(\overline{\mathbf{A}}) \leq \prod_{i=1}^{n}(1 + \delta_K(\mathbf{A}_i)) - 1. \tag{13}$$

■

In [28], these relationships are then utilized to derive the reconstruction error bounds for a TCS system.

III. Optimized Multilinear Projections for TCS

In this section, we show how to optimize the multilinear sensing matrix when the dictionaries $\mathbf{\Psi}_i$ $(i = 1, ..., n)$ for each dimension are fixed. We first introduce the related design approaches for CS, then present the proposed methods for TCS, including a separable and a non-separable design approach.

A. Sensing Matrix Design for CS

We observe that the sufficient conditions on the RIC or the mutual coherence for successful CS reconstruction, as reviewed in Section II-B, only describe the worst case bound, which means that the average recovery performance is not reflected. The CS average-case performance has been analyzed in [46]–[50], that considers a distribution over the signals. In fact, the most challenging part of CS sensing matrix design lies in deriving a measure that can directly reveal the expected-case reconstruction accuracy when the signals are unknown.

In [8], Elad *et al.* proposed the notion of averaged mutual coherence, based on which an iterative algorithm is derived for optimal sensing matrix design. This approach aims to minimize the largest absolute values of the off-diagonal entries in the Gram matrix of $\mathbf{A}$, i.e., $\mathbf{G_A} = \mathbf{A}^T\mathbf{A}$. It has been shown to outperform a random Gaussian sensing matrix in terms of reconstruction accuracy, but is time-consuming to construct and can ruin the worst case guarantees by increasing off-diagonal values that are not the largest in the original Gram matrix.

In order to make any subset of columns in $\mathbf{A}$ as orthogonal as possible, Sapiro *et al.* proposed in [18] to make $\mathbf{G_A}$ as close as possible to an identity matrix, i.e., $\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi} \approx \mathbf{I}_{\hat{N}}$. It is then approximated by minimizing $||\mathbf{\Lambda} - \mathbf{\Lambda}\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{\Lambda}||_F^2$, where $\mathbf{\Lambda}$ comes from the eigen-decomposition of $\mathbf{\Psi}^T\mathbf{\Psi}$, i.e., $\mathbf{\Psi}^T\mathbf{\Psi} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, and $\mathbf{\Gamma} = \mathbf{\Phi}\mathbf{V}$. This approach is also iterative, but outperforms Elad's method. Considering the fact that $\mathbf{A}$ has minimum coherence when the magnitudes of all the off-diagonal entries of $\mathbf{G_A}$ are equal, Xu *et al.* proposed an Equiangular Tight Frame (ETF) [1] based method in [9]. The problem is modeled as: $\min_{\mathbf{G}_t\in\mathcal{H},\mathbf{\Phi}} ||\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi} - \mathbf{G}_t||_F^2$, where $\mathbf{G}_t$ is the target Gram matrix and $\mathcal{H}$ is the set of the ETF Gram matrices. The problem is solved following an alternating minimization procedure. Once the target Gram matrix is determined, the optimized projection matrix is constructed using a QR factorization with eigenvalue decomposition. Improved performance has been observed for the obtained sensing matrix.

More recently, based on the same idea as Sapiro, the problem of

$$\min_{\mathbf{\Phi}} ||\mathbf{I}_{\hat{N}} - \mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}||_F^2 \tag{14}$$

has been considered and an analytical solution has been derived in [11]. Meanwhile, in [10], [51], it has been shown that in order to achieve good expected-case Mean Squared Error (MSE) performance, the equivalent sensing matrix ought to be close to a Parseval tight frame, thus leading to the following design approach:

$$\min_{\mathbf{\Phi}} ||\mathbf{\Phi}||_F^2,\ s.t.\ \mathbf{\Phi}\mathbf{\Psi}\mathbf{\Psi}^T\mathbf{\Phi}^T = \mathbf{I}_M, \tag{15}$$

where $||\mathbf{\Phi}||_F^2$ is the sensing cost that also affects the reconstruction accuracy (as verified in [10], [51]). A closed form

[1] A matrix $\mathbf{X} \in \mathbb{R}^{N\times\hat{N}}$ is defined to be an Equiangular Tight Frame (ETF) if it is column normalized and for any $i \neq j$, $|\mathbf{X}^T(:,i)\mathbf{X}(:,j)| = \sqrt{(\hat{N}-N)/[N(\hat{N}-1)]}$. The Gram matrix of an ETF is called the ETF Gram.

solution to this problem was also obtained in [10], [51]. These approaches have further improved the average reconstruction performance for a CS system that is able to employ the optimized sensing matrix.

On the other hand, using the model of Xu's method [9], Cleju [12] proposed to take $\mathbf{G}_t = \mathbf{\Psi}^T\mathbf{\Psi}$ so that the equivalent sensing matrix has similar properties to those of $\mathbf{\Psi}$. Bai *et al.* [20] proposed combining the ETF Grams and that proposed by Cleju to solve:

$$\min_{\mathbf{G}_t\in\bar{\mathcal{H}},\mathbf{\Phi}} (1-\beta)||\mathbf{\Psi}^T\mathbf{\Psi}-\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}||_F^2+\beta||\mathbf{G}_t-\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}||_F^2, \tag{16}$$

where $\mathbf{G}_t$ is the target Gram matrix, β is a trade-off parameter and $\bar{\mathcal{H}}$ is relaxed from the set of the ETF Grams to:

$$\bar{\mathcal{H}} = \{\mathbf{G}_t : \mathbf{G}_t = \mathbf{G}_t^T, \mathbf{G}_t(k,k) = 1, \forall k, \max_{i\neq j}|\mathbf{G}_t(i,j)| \leq \underline{\mu}\}, \tag{17}$$

where $\underline{\mu} = \sqrt{(\hat{N}-N)/[N(\hat{N}-1)]}$. A closed form solution is then derived analytically to solve this problem [20]. The method is shown to be more general than other existing methods in the sense of the requirement on $\mathbf{\Psi}$, and promising results using this method are demonstrated.

B. Multidimensional Sensing Matrix Design for TCS

In contrast to the aforementioned methods, we consider optimization of the sensing matrix for TCS. Compared to the design process in conventional CS, the main distinction for the TCS is that we would like to optimize multiple separable sensing matrices $\mathbf{\Phi}_i$ $(i = 1, ..., n)$, rather than a single matrix $\mathbf{\Phi}$. In this section, we first extend the approaches in (14) and (15) to the TCS case, where our main contribution is to prove that the problem is separable along each of its dimensions and can be solved using the existing approaches. In addition, we propose a new approach for TCS sensing matrices design by incorporating the state-of-art metrics for the sensing matrix optimization as in [10], [12], [20]. To simplify our exposition, we elaborate our methods in the following sections for the case of $n = 2$, i.e., the tensor signal becomes a matrix, but note that the methods can be straightforwardly extended to an n mode tensor case ($n > 2$).

As reviewed in Section II-B, the performance of existing TCS reconstruction algorithms relies on the quality of $\overline{\mathbf{A}}$, where $\overline{\mathbf{A}} = \mathbf{A}_2 \otimes \mathbf{A}_1$ when $n = 2$. Therefore, when the multilinear dictionary $\overline{\mathbf{\Psi}} = \mathbf{\Psi}_2 \otimes \mathbf{\Psi}_1$ is given, one can optimize $\overline{\mathbf{\Phi}}$ (where $\overline{\mathbf{\Phi}} = \mathbf{\Phi}_2 \otimes \mathbf{\Phi}_1$) using the methods for CS as introduced in Section III-A.

However, when implementing a TCS system, it is still necessary to obtain the separable matrices, i.e., $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$. One intuitive solution is to design $\overline{\mathbf{\Phi}}$ using the aforementioned approaches for CS and then to decompose $\overline{\mathbf{\Phi}}$ by solving the following problem:

$$\min_{\mathbf{\Phi}_1,\mathbf{\Phi}_2} ||\overline{\mathbf{\Phi}} - \mathbf{\Phi}_2 \otimes \mathbf{\Phi}_1||_F^2, \tag{18}$$

which has been studied as a Nearest Kronecker Product (NKP) problem in [52]. But this is not a feasible solution for TCS sensing matrix design. First of all, $\overline{\mathbf{\Phi}}$ can only be exactly decomposed as $\mathbf{\Phi}_2 \otimes \mathbf{\Phi}_1$ when a certain permutation of $\overline{\mathbf{\Phi}}$ has rank 1 [52], which is not the case for most sensing strategies. When the term in (18) is minimized to a non-zero value, the solution $\hat{\mathbf{\Phi}}_1$, $\hat{\mathbf{\Phi}}_2$ leads to a sensing matrix $\hat{\mathbf{\Phi}}_2 \otimes \hat{\mathbf{\Phi}}_1$, which may not satisfy the condition of the sensing matrix $\overline{\mathbf{\Phi}}$ for good CS recovery (e.g., the requirement on the mutual coherence), thereby ruining the reconstruction guarantees. Secondly, although one may include the mutual coherence requirement as a constraint in the problem, to solve (18), explicit storage of $\overline{\mathbf{\Phi}}$ is necessary, which is restrictive for high dimensional problems. In addition, when the number of tensor modes increases, the problem becomes more complex to solve.

Therefore, we aim to optimize $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ directly without knowing $\overline{\mathbf{\Phi}}$. Extending (14) and (15), we first propose a method that is shown to be separable as independent sub-design-problems. Then a non-separable design approach is presented and a gradient based algorithm is derived.

1) A Separable Design Approach: The proposed separable design approach (Approach I) is as follows:

$$\min_{\mathbf{\Phi}_1,\mathbf{\Phi}_2} ||\mathbf{I}_{\hat{N}_1\hat{N}_2}-(\mathbf{\Psi}_2^T\otimes\mathbf{\Psi}_1^T)(\mathbf{\Phi}_2^T\otimes\mathbf{\Phi}_1^T)(\mathbf{\Phi}_2\otimes\mathbf{\Phi}_1)(\mathbf{\Psi}_2\otimes\mathbf{\Psi}_1)||_F^2, \tag{19}$$

and it is an extension of (14) to the case when a multilinear sensing matrix is employed. The solution of (19) is presented in Theorem 3 and Approach I is also summarized in Algorithm 1.

Theorem 3: Assume for $i = 1,\ 2$, $\bar{N}_i = rank(\mathbf{\Psi}_i)$, $\mathbf{\Psi}_i = \mathbf{U}_{\mathbf{\Psi}_i}\begin{bmatrix}\mathbf{\Lambda}_{\mathbf{\Psi}_i} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{V}_{\mathbf{\Psi}_i}^T$ is an SVD of $\mathbf{\Psi}_i$ and $\mathbf{\Lambda}_{\mathbf{\Psi}_i} \in \mathbb{R}^{\bar{N}_i\times\bar{N}_i}$. Let $\hat{\mathbf{\Phi}}_i \in \mathbb{R}^{M_i\times\bar{N}_i}$ $(i = 1,\ 2)$ be matrices with $rank(\hat{\mathbf{\Phi}}_i) = M_i$ and $M_i \leq \bar{N}_i$ is assumed. Then

- the following equation is a solution to (19):

$$\hat{\mathbf{\Phi}}_i = \mathbf{U}\begin{bmatrix}\mathbf{I}_{M_i} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}^T\mathbf{\Lambda}_{\mathbf{\Psi}_i}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{U}_{\mathbf{\Psi}_i}^T, \tag{20}$$

  where $i = 1,\ 2$, $\mathbf{U} \in \mathbb{R}^{M_i\times M_i}$ and $\mathbf{V} \in \mathbb{R}^{\bar{N}_i\times\bar{N}_i}$ are arbitrary orthonormal matrices;
- the resulting equivalent sensing matrices $\hat{\mathbf{A}}_i = \hat{\mathbf{\Phi}}_i\mathbf{\Psi}_i$ $(i = 1,\ 2)$ are Parseval tight frames, i.e., $||\hat{\mathbf{A}}_i^T\mathbf{z}||_2 = ||\mathbf{z}||_2$, where $\mathbf{z} \in \mathbb{R}^{\hat{N}_i}$ is an arbitrary vector.
- the minimum of (19) is $\hat{N}_1\hat{N}_2 - M_1M_2$;
- separately solving the sub-problems

$$\min_{\mathbf{\Phi}_i} ||\mathbf{I}_{\hat{N}_i} - \mathbf{\Psi}_i^T\mathbf{\Phi}_i^T\mathbf{\Phi}_i\mathbf{\Psi}_i||_F^2 \tag{21}$$

  for $i = 1,\ 2$ leads to the same solutions as (20) and the resulting objective in (19) has the same minimum, i.e., $\hat{N}_1\hat{N}_2 - M_1M_2$. ■

Proof: The proof is given in Appendix A.

Clearly, Approach I is separable, which means that we can independently design each $\mathbf{\Phi}_i$ according to the corresponding sparsifying dictionary $\mathbf{\Psi}_i$ in mode i. This observation stays consistent when we consider the situation in an alternative way. Applying the method in (14) to acquire the optimal $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ independently, we are actually trying to make any subset of columns in $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively, as orthogonal as possible. As a result, the matrix $\overline{\mathbf{A}} = \mathbf{A}_2 \otimes \mathbf{A}_1$ that is obtained will

Algorithm 1 Design Approach I

Input: $\mathbf{\Psi}_i$ $(i = 1,\ 2)$.
Output: $\hat{\mathbf{\Phi}}_i$ $(i = 1,\ 2)$.
1: **for** $i = 1,\ 2$ **do**
2: Calculate optimized $\hat{\mathbf{\Phi}}_i$ using (20);
3: **end**
4: Normalization for $i = 1,\ 2$: $\hat{\mathbf{\Phi}}_i = \sqrt{N_i}\hat{\mathbf{\Phi}}_i/||\hat{\mathbf{\Phi}}_i||_F$.

also be as orthogonal as possible. This follows from the fact that for any two columns of $\overline{\mathbf{A}}$, we have

$$\begin{aligned}|\overline{\mathbf{a}}_p^T\overline{\mathbf{a}}_q| &= |[(\mathbf{a}_2)_l^T \otimes (\mathbf{a}_1)_s^T][(\mathbf{a}_2)_c \otimes (\mathbf{a}_1)_d]| \\ &= |[(\mathbf{a}_2)_l^T(\mathbf{a}_2)_c][(\mathbf{a}_1)_s^T(\mathbf{a}_1)_d]|, \end{aligned} \quad (22)$$

where $\overline{\mathbf{a}}$, $\mathbf{a}_1$ and $\mathbf{a}_2$ denote the column of $\overline{\mathbf{A}}$, $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively, and $p,\ q,\ l,\ s,\ c,\ d$ are the column indices.

Using the second statement of Theorem 3, we can derive the following corollary.

Corollary 1: The solution in (20) also solves the following problems for $i = 1,\ 2$:

$$\min_{\mathbf{\Phi}_i} ||\mathbf{\Phi}_i||_F^2,\ s.t.\ \mathbf{\Phi}_i\mathbf{\Psi}_i\mathbf{\Psi}_i^T\mathbf{\Phi}_i^T = \mathbf{I}_{M_i}, \quad (23)$$

which represent the separable sub-problems of the following design approach:

$$\min_{\mathbf{\Phi}_1,\mathbf{\Phi}_2} ||\mathbf{\Phi}_2 \otimes \mathbf{\Phi}_1||_F^2, \quad (24)$$

$$s.t.\ (\mathbf{\Phi}_2 \otimes \mathbf{\Phi}_1)(\mathbf{\Psi}_2 \otimes \mathbf{\Psi}_1)(\mathbf{\Psi}_2^T \otimes \mathbf{\Psi}_1^T)(\mathbf{\Phi}_2^T \otimes \mathbf{\Phi}_1^T) = \mathbf{I}_{M_1M_2},$$

and it is in fact a multidimensional extension of the CS sensing matrix design approach proposed in [10]. ■

Proof: Since the equivalent sensing matrices designed using Approach I are Parseval tight frames, it follows from the derivation in [10] that the sub-problems in (23) have the same solution as in (20). The problem in (24) can be proved separable simply by revealing the fact that $||\mathbf{\Phi}_2 \otimes \mathbf{\Phi}_1||_F^2 = ||\mathbf{\Phi}_2||_F^2||\mathbf{\Phi}_1||_F^2$, and when $\mathbf{\Phi}_i\mathbf{\Psi}_i\mathbf{\Psi}_i^T\mathbf{\Phi}_i^T = \mathbf{I}_{M_i}$ is satisfied for both $i = 1$ and 2, the constraint in (24) is also satisfied. ■

By decomposing the original problems into independent sub-problems, the sensing matrices can be designed in parallel and the problem becomes easier to solve. However, the CS sensing matrix design approaches are not always separable after being extended to the multidimensional case, because a variety of different criteria can be used for sensing matrix design as reviewed in Section III-A, and in many cases the decomposition is not provable. We will propose a non-separable approach in the following section.

2) A Non-separable Design Approach: Taking into account: i) the impact of sensing cost on reconstruction performance [10]; ii) the benefit of making the equivalent sensing matrix so that it has similar properties to those of the sparsifying dictionary [12]; and iii) the conventional requirement on the mutual coherence, we put forth the following Design Approach II:

$$\begin{aligned}\min_{\mathbf{\Phi}_1,\mathbf{\Phi}_2}\ & (1-\beta)||(\overline{\mathbf{\Psi}})^T\overline{\mathbf{\Psi}} - (\overline{\mathbf{\Psi}})^T(\overline{\mathbf{\Phi}})^T\overline{\mathbf{\Phi}}\overline{\mathbf{\Psi}}||_F^2 \\ &+ \alpha||\overline{\mathbf{\Phi}}||_F^2 + \beta||\mathbf{I}_{\hat{N}_1\hat{N}_2} - (\overline{\mathbf{\Psi}})^T(\overline{\mathbf{\Phi}})^T\overline{\mathbf{\Phi}}\overline{\mathbf{\Psi}}||_F^2, \end{aligned} \quad (25)$$

where $\overline{\mathbf{\Psi}} = \mathbf{\Psi}_2 \otimes \mathbf{\Psi}_1$, $\overline{\mathbf{\Phi}} = \mathbf{\Phi}_2 \otimes \mathbf{\Phi}_1$, α and β are tuning parameters. As investigated in [10] and [20], $\alpha \geq 0$ controls the sensing energy; while $\beta \in [0,\ 1]$ balances the impact of the first and third terms to achieve optimal performance under different conditions of the measurement noise. The choice of these parameters will be investigated in Section V-A. We note that this problem is non-separable owing to the multiple terms involved and the non-decomposable nature of the first term.

To solve (25), we adopt a coordinate descent method. Denoting the objective as $f(\mathbf{\Phi}_1, \mathbf{\Phi}_2)$, we first compute its gradient with respect to $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$, respectively, and the result is as follows:

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{\Phi}_i} &= 4||\mathbf{G}_{\mathbf{A}_j}||_F^2(\mathbf{A}_i\mathbf{G}_{\mathbf{A}_i}\mathbf{\Psi}_i^T) - 4\beta||\mathbf{A}_j||_F^2(\mathbf{A}_i\mathbf{\Psi}_i^T) \\ &+ 2\alpha||\mathbf{\Phi}_j||_F^2\mathbf{\Phi}_i + 4(\beta-1)||\mathbf{\Psi}_j\mathbf{A}_j^T||_F^2(\mathbf{A}_i\mathbf{G}_{\mathbf{\Psi}_i}\mathbf{\Psi}_i^T), \end{aligned} \quad (26)$$

where $i,\ j \in \{1,\ 2\}$ and $j \neq i$, $\mathbf{G}_{\mathbf{A}_i} = \mathbf{A}_i^T\mathbf{A}_i$ and $\mathbf{G}_{\mathbf{\Psi}_i} = \mathbf{\Psi}_i^T\mathbf{\Psi}_i$.

For generality, we also provide the result for the $n > 2$ case as follows:

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{\Phi}_i} &= 4\omega_i(\mathbf{A}_i\mathbf{G}_{\mathbf{A}_i}\mathbf{\Psi}_i^T) - 4\beta\theta_i(\mathbf{A}_i\mathbf{\Psi}_i^T) \\ &+ 2\alpha\tau_i\mathbf{\Phi}_i + (4\beta-4)\rho_i(\mathbf{A}_i\mathbf{G}_{\mathbf{\Psi}_i}\mathbf{\Psi}_i^T), \end{aligned} \quad (27)$$

where $i,\ j \in \{1, ..., n\}$ and $j \neq i$, $\omega_i = \prod_j ||\mathbf{G}_{\mathbf{A}_j}||_F^2$, $\theta_i = \prod_j ||\mathbf{A}_j||_F^2$, $\tau_i = \prod_j ||\mathbf{\Phi}_j||_F^2$, $\rho_i = \prod_j ||\mathbf{\Psi}_j\mathbf{A}_j^T||_F^2$.

With the gradient obtained, we can solve (25) by alternatively updating $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ as follows:

$$\mathbf{\Phi}_i^{(t+1)} = \mathbf{\Phi}_i^{(t)} - \eta\frac{\partial f}{\partial \mathbf{\Phi}_i}, \quad (28)$$

where $\eta > 0$ is a step size parameter. The algorithm for solving (25) is summarized in Algorithm 2.

Algorithm 2 Design Approach II

Input: $\mathbf{\Psi}_i$ $(i = 1,\ 2)$, $\mathbf{\Phi}_i^{(0)}$ $(i = 1,\ 2)$, α, β, η, $t = 0$.
Output: $\hat{\mathbf{\Phi}}_i$ $(i = 1,\ 2)$.
1: **Repeat**
2: **for** $i = 1,\ 2$ **do**
3: $\mathbf{\Phi}_i^{(t+1)} = \mathbf{\Phi}_i^{(t)} - \eta\frac{\partial f}{\partial \mathbf{\Phi}_i}$, where $\frac{\partial f}{\partial \mathbf{\Phi}_i}$ is given by (26);
4: **end**
5: $t = t + 1$;
6: **Until** a stopping criteria is met.
7: Normalization for $i = 1,\ 2$: $\hat{\mathbf{\Phi}}_i = \sqrt{N_i}\mathbf{\Phi}_i/||\mathbf{\Phi}_i||_F$.

So far, we have considered optimizing the multidimensional sensing matrix when the sparsifying dictionaries for each tensor mode are given. For the purpose of joint optimization, we will proceed to optimize the dictionaries by coupling fixed sensing matrices. The joint optimization will eventually be achieved by alternately optimizing the sensing matrices and the sparsifying dictionaries.

IV. Jointly Learning The Multidimensional Dictionary and Sensing Matrix

In this section, we first propose a sensing-matrix-coupled method for multidimensional sparsifying dictionary learning. Then it is combined with the previously introduced optimization approach for a multilinear sensing matrix to yield a joint optimization algorithm. In the spirit of the coupled KSVD method [18], our approach for dictionary learning can be viewed as a sensing-matrix-coupled version of a tensor KSVD algorithm. We start by briefly introducing the coupled KSVD method.

A. Coupled KSVD and Related Work

The Coupled KSVD (cKSVD) [18] is a dictionary learning approach for vectorized signals. Let $\mathbf{X} = [\ \mathbf{x}_1 \quad ... \quad \mathbf{x}_T\ ]$ be a $N \times T$ matrix containing a training sequence of T signals $\mathbf{x}_1, ..., \mathbf{x}_T$. The cKSVD aims to solve the following problem, i.e., to learn a dictionary $\mathbf{\Psi} \in \mathbb{R}^{N \times \hat{N}}$ from $\mathbf{X}$:

$$\min_{\mathbf{\Psi},\mathbf{S}} \gamma||\mathbf{X} - \mathbf{\Psi}\mathbf{S}||_F^2 + ||\mathbf{Y} - \mathbf{\Phi}\mathbf{\Psi}\mathbf{S}||_F^2,\ s.t.\ \forall i,\ ||\mathbf{s}_i||_0 \leq K, \tag{29}$$

where $\mathbf{S} = [\ \mathbf{s}_1 \quad ... \quad \mathbf{s}_T\ ]$ is the sparse representation with size $\hat{N} \times T$, $\gamma > 0$ is a tuning parameter and $\mathbf{Y} \in \mathbb{R}^{M \times T}$ contains the measurement vectors taken by the sensing matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$, i.e., $\mathbf{Y} = [\ \mathbf{y}_1 \quad ... \quad \mathbf{y}_T\ ]$ and $\mathbf{Y} = \mathbf{\Phi}\mathbf{X} + \mathbf{E}$ with $\mathbf{E} \in \mathbb{R}^{M \times T}$ representing the noise. Then the problem in (29) is reformatted as:

$$\min_{\mathbf{\Psi},\mathbf{S}} ||\mathbf{Z} - \mathbf{D}\mathbf{S}||_F^2,\ s.t.\ \forall i,\ ||\mathbf{s}_i||_0 \leq K, \tag{30}$$

where $\mathbf{Z} = \left[\ \gamma\mathbf{X}^T \quad \mathbf{Y}^T\ \right]^T$, $\mathbf{D} = \left[\ \gamma\mathbf{I}_N \quad \mathbf{\Phi}^T\ \right]^T \mathbf{\Psi}$. The problem can then be solved following the conventional KSVD algorithm [14] and conducting proper normalization.

Specifically, with an initial arbitrary $\mathbf{\Psi}$, it first recovers $\mathbf{S}$ using some available algorithms, e.g., OMP. Then the objective in (30) is rewritten as:

$$\min_{\mathbf{\Psi},\mathbf{S}} ||\tilde{\mathbf{R}}_p - \mathbf{d}_p\tilde{\mathbf{s}}_p^T||_F^2, \tag{31}$$

where p is the index of the current atom we aim to update, $\tilde{\mathbf{s}}_p^T$ is the row of $\mathbf{S}$ where the zeros have been removed, $\mathbf{R}_p = \mathbf{Z} - \sum_{q \neq p} \mathbf{d}_q\mathbf{s}_q^T$ and $\tilde{\mathbf{R}}_p$ denotes the columns of $\mathbf{R}_p$ corresponding to the nonzero entries of $\mathbf{s}_p^T$. Let $\tilde{\mathbf{R}}_p = \mathbf{U}_\mathbf{R}\mathbf{\Lambda}_\mathbf{R}\mathbf{V}_\mathbf{R}^T$ be a SVD of $\tilde{\mathbf{R}}_p$, then the highest component of the coupled error $\tilde{\mathbf{R}}_p$ can be eliminated by defining:

$$\hat{\boldsymbol{\psi}}_p = (\gamma^2\mathbf{I}_N + \mathbf{\Phi}^T\mathbf{\Phi})^{-1}\left[\ \gamma\mathbf{I}_N \quad \mathbf{\Phi}^T\ \right]\mathbf{u}_\mathbf{R}^1, \tag{32}$$

$$\tilde{\mathbf{s}}_p = ||\hat{\boldsymbol{\psi}}_p||_2\lambda_\mathbf{R}^1\mathbf{v}_\mathbf{R}^1, \tag{33}$$

where $\lambda_\mathbf{R}^1$ is the largest singular value of $\tilde{\mathbf{R}}_p$ and $\mathbf{u}_\mathbf{R}^1$, $\mathbf{v}_\mathbf{R}^1$ are the corresponding left and right singular vectors. The update column p of $\mathbf{\Psi}$ is obtained after normalization: $\hat{\boldsymbol{\psi}}_p = \hat{\boldsymbol{\psi}}_p/||\hat{\boldsymbol{\psi}}_p||_2$. The above process is then iterated to update every atom of $\mathbf{\Psi}$.

Clearly the sensing matrix has been taken into account during the dictionary learning process, which has been shown to be beneficial for CS reconstruction performance [18].

When it comes to the tensor case, KHOSVD [37], that is a tensor-based dictionary learning approach, is obtained by extending KSVD. It models the 2D dictionary learning problem as:

$$\min_{\mathbf{\Psi}_1,\mathbf{\Psi}_2,\underline{\mathbf{S}}} ||\underline{\mathbf{X}} - \underline{\mathbf{S}} \times_1 \mathbf{\Psi}_1 \times_2 \mathbf{\Psi}_2||_F^2,\ s.t.,\ \forall i,\ ||\mathbf{S}_i||_0 \leq K, \tag{34}$$

where $\mathbf{\Psi}_1 \in \mathbb{R}^{N_1 \times \hat{N}_1}$ and $\mathbf{\Psi}_2 \in \mathbb{R}^{N_2 \times \hat{N}_2}$ are the dictionaries to be learnt, $\underline{\mathbf{X}}$ is the tensor containing the training data, $\underline{\mathbf{S}}$ is the sparse representation and $\mathbf{S}_i$ is its mode-i unfolding. Its learning process follows the same train of thought as with the conventional KSVD method, except that to eliminate the largest error in each iteration, a Higher Order SVD (HOSVD) [53], i.e., SVD for tensors, is employed. First, the sparse representation is estimated using OMP. Then, an atom of $\mathbf{\Psi}_1$ and an atom of $\mathbf{\Psi}_2$, that correspond to an element of $\underline{\mathbf{S}}$ are updated, and $\hat{N}_1\hat{N}_2$ iterations are performed to complete the update of all the atoms. This process is then repeated until the algorithm converges. Clearly, the inner iterations involve duplicated updating of the dictionary atoms because there are only $\hat{N}_1 + \hat{N}_2$ atoms in total. In addition, the effect owing to the sensing matrix is not considered in this approach.

B. The cTKSVD Approach

In order to learn multidimensional separable dictionaries for high dimensional signals, and eventually to achieve joint optimization of the multidimensional dictionary and sensing matrix, we will derive a coupled-KSVD algorithm for a tensor, i.e., cTKSVD, in this section. The proposed cTKSVD algorithm follows the same train of thought as the cKSVD method, however, the specific steps are different due to the involvement of extra dimensions and dictionaries. We will first provide the problem formulation for cTKSVD, which will then be solved by alternately updating the sparse coefficients and the atoms of various dictionaries. Note that each of the updating steps is also different from that of cKSVD in terms of either the employed solver or the elements that are updated. Again for simplicity we will still describe the main flow for 2-D signals, i.e., $n = 2$.

Consider a training sequence of 2-D signals $\mathbf{X}_1, ..., \mathbf{X}_T$, we obtain a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{N_1 \times N_2 \times T}$ by stacking them along the third dimension. Denoting the stack of the sparse representations $\mathbf{S}_i \in \mathbb{R}^{\hat{N}_1 \times \hat{N}_2},\ (i = 1, ..., T)$ by $\underline{\mathbf{S}} \in \mathbb{R}^{\hat{N}_1 \times \hat{N}_2 \times T}$, we propose the following optimization problem to learn the multidimensional dictionary:

$$\min_{\mathbf{\Psi}_1,\mathbf{\Psi}_2,\underline{\mathbf{S}}} ||\underline{\mathbf{Z}} - \underline{\mathbf{S}} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2||_F^2,\ s.t.,\ \forall i,\ ||\mathbf{S}_i||_0 \leq K, \tag{35}$$

in which

$$\underline{\mathbf{Z}} = \begin{bmatrix} \gamma^2\underline{\mathbf{X}} & \gamma\underline{\mathbf{Y}}_2 \\ \gamma\underline{\mathbf{Y}}_1 & \underline{\mathbf{Y}} \end{bmatrix},\ \underline{\mathbf{Y}}_i = \underline{\mathbf{X}} \times_i \mathbf{\Phi}_i + \underline{\mathbf{E}}_i, \tag{36}$$

$$\mathbf{D}_1 = \begin{bmatrix} \gamma\mathbf{I}_{\hat{N}_1} \\ \mathbf{\Phi}_1 \end{bmatrix}\mathbf{\Psi}_1,\ \mathbf{D}_2 = \begin{bmatrix} \gamma\mathbf{I}_{\hat{N}_2} \\ \mathbf{\Phi}_2 \end{bmatrix}\mathbf{\Psi}_2, \tag{37}$$

and $\gamma > 0$ is a tuning parameter.

The problem in (35) aims to minimize the representation error $||\underline{\mathbf{X}} - \underline{\mathbf{S}} \times_1 \mathbf{\Psi}_1 \times_2 \mathbf{\Psi}_2||_F^2$ and the overall projection error

$||\underline{\mathbf{Y}} - \underline{\mathbf{S}} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2||_F^2$ with constraints on the sparsity of each slice of the tensor. In addition, it also takes into account the projection errors induced by $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ individually.

Using an available sparse reconstruction algorithm for the TCS, e.g., Tensor OMP (TOMP) [35], and initial dictionaries $\mathbf{\Psi}_1,\ \mathbf{\Psi}_2$, the sparse representation $\underline{\mathbf{S}}$ can be estimated first. Then we update the multilinear dictionary alternately. We first update the atoms of $\mathbf{\Psi}_1$ with $\mathbf{\Psi}_2$ fixed. The objective in (35) is rewritten as:

$$||\underline{\mathbf{R}}_{p_1} - \sum_{q_2} (\mathbf{d}_1)_{p_1} \circ (\mathbf{d}_2)_{q_2} \circ \mathbf{s}_{(p_1-1)\hat{N}_2+q_2}||_F^2, \tag{38}$$

where $\underline{\mathbf{R}}_{p_1} = \underline{\mathbf{Z}} - \sum_{q_1 \neq p_1} \sum_{q_2} (\mathbf{d}_1)_{q_1} \circ (\mathbf{d}_2)_{q_2} \circ \mathbf{s}_{(q_1-1)\hat{N}_2+q_2}$; p_1 is the index of the atom for the current update and $q_1,\ q_2$ denote the indices of the remaining atoms of $\mathbf{\Psi}_1$ and all the atoms of $\mathbf{\Psi}_2$, respectively; $\mathbf{d}_1, \mathbf{d}_2$ are columns of $\mathbf{D}_1, \mathbf{D}_2$; $\mathbf{s}$ is the mode-3 vector of $\underline{\mathbf{S}}$. Then to satisfy the sparsity constraint in (35), we only keep the non-zero entries of $\mathbf{s}_{(p_1-1)\hat{N}_2+q_2}$ and the corresponding subset of $\underline{\mathbf{R}}_{p_1}$ to obtain:

$$||\tilde{\underline{\mathbf{R}}}_{p_1} - \sum_{q_2} (\mathbf{d}_1)_{p_1} \circ (\mathbf{d}_2)_{q_2} \circ \tilde{\mathbf{s}}_{(p_1-1)\hat{N}_2+q_2}||_F^2. \tag{39}$$

Assuming that after carrying out a HOSVD [53] for $\tilde{\underline{\mathbf{R}}}_{p_1}$, the largest singular value is $\lambda_{\mathbf{R}}^1$ and the corresponding singular vectors are $\mathbf{u}_{\mathbf{R}}^1$, $\mathbf{v}_{\mathbf{R}}^1$ and $\boldsymbol{\omega}_{\mathbf{R}}^1$, we eliminate the largest error by:

$$(\hat{\mathbf{d}}_1)_{p_1} = \mathbf{u}_{\mathbf{R}}^1,\ \mathbf{D}_2 \tilde{\mathbf{S}}_{p_1,:,:} = \mathbf{v}_{\mathbf{R}}^1 \circ (\lambda_{\mathbf{R}}^1 \boldsymbol{\omega}_{\mathbf{R}}^1), \tag{40}$$

where $\tilde{\mathbf{S}}_{p_1,:,:}$ denotes the horizontal slice of $\underline{\mathbf{S}}$ at index p_1 that contains only non-zero mode-2 vectors. The atom of $\mathbf{\Psi}_1$ is then calculated using the pseudo-inverse as:

$$(\hat{\boldsymbol{\psi}}_1)_{p_1} = (\gamma^2 \mathbf{I}_{N_1} + \mathbf{\Phi}_1^T \mathbf{\Phi}_1)^{-1} \begin{bmatrix} \gamma \mathbf{I}_{N_1} & \mathbf{\Phi}_1^T \end{bmatrix} \mathbf{u}_{\mathbf{R}}^1. \tag{41}$$

The current update is then obtained after normalization:

$$(\hat{\boldsymbol{\psi}}_1)_{p_1} = \frac{(\hat{\boldsymbol{\psi}}_1)_{p_1}}{||(\hat{\boldsymbol{\psi}}_1)_{p_1}||_2}, \tag{42}$$

$$\mathbf{D}_2 \tilde{\mathbf{S}}_{p_1,:,:} = ||(\hat{\boldsymbol{\psi}}_1)_{p_1}||_2 \mathbf{v}_{\mathbf{R}}^1 \circ (\lambda_{\mathbf{R}}^1 \boldsymbol{\omega}_{\mathbf{R}}^1). \tag{43}$$

Since $\mathbf{D}_2$ and the support indices of each mode-2 vector in $\tilde{\mathbf{S}}_{p_1,:,:}$ are known, the updated coefficients $\tilde{\mathbf{S}}_{p_1,:,:}$ can be easily calculated by the Least Square (LS) solution. The above process is repeated for all the atoms to update the dictionary $\mathbf{\Psi}_1$.

The next step is to update $\mathbf{\Psi}_2$ with the obtained $\mathbf{\Psi}_1$ fixed. It follows a similar procedure to that described previously. Specifically, the objective in (35) is rewritten as:

$$||\tilde{\underline{\mathbf{R}}}_{p_2} - \sum_{q_1} (\mathbf{d}_1)_{q_1} \circ (\mathbf{d}_2)_{p_2} \circ \tilde{\mathbf{s}}_{(q_1-1)\hat{N}_2+p_2}||_F^2, \tag{44}$$

where $\tilde{\mathbf{s}}$ is the mode-3 vector with only non-zero entries, $\tilde{\underline{\mathbf{R}}}_{p_2}$ is the corresponding subset of $\underline{\mathbf{R}}_{p_2}$, $\underline{\mathbf{R}}_{p_2} = \underline{\mathbf{Z}} - \sum_{q_1} \sum_{q_2 \neq p_2} (\mathbf{d}_1)_{q_1} \circ (\mathbf{d}_2)_{q_2} \circ \mathbf{s}_{(q_1-1)\hat{N}_2+q_2}$ and p_2 is the index of the atom for current update. A HOSVD is carried out for $\tilde{\underline{\mathbf{R}}}_{p_2}$ and the update steps corresponding to (40) - (43) now become:

$$(\hat{\mathbf{d}}_2)_{p_2} = \mathbf{v}_{\mathbf{R}}^1,\ \mathbf{D}_1 \tilde{\mathbf{S}}_{:,p_2,:} = \mathbf{u}_{\mathbf{R}}^1 \circ (\lambda_{\mathbf{R}}^1 \boldsymbol{\omega}_{\mathbf{R}}^1), \tag{45}$$

$$(\hat{\boldsymbol{\psi}}_2)_{p_2} = (\gamma^2 \mathbf{I}_{N_2} + \mathbf{\Phi}_2^T \mathbf{\Phi}_2)^{-1} \begin{bmatrix} \gamma \mathbf{I}_{N_2} & \mathbf{\Phi}_2^T \end{bmatrix} \mathbf{v}_{\mathbf{R}}^1, \tag{46}$$

$$(\hat{\boldsymbol{\psi}}_2)_{p_2} = (\hat{\boldsymbol{\psi}}_2)_{p_2} / ||(\hat{\boldsymbol{\psi}}_2)_{p_2}||_2, \tag{47}$$

$$\mathbf{D}_1 \tilde{\mathbf{S}}_{:,p_2,:} = ||(\hat{\boldsymbol{\psi}}_2)_{p_2}||_2 \mathbf{u}_{\mathbf{R}}^1 \circ (\lambda_{\mathbf{R}}^1 \boldsymbol{\omega}_{\mathbf{R}}^1), \tag{48}$$

in which $\tilde{\mathbf{S}}_{:,p_2,:}$ represents the lateral slice at index p_2 and its updated elements can also be calculated using LS. The dictionary $\mathbf{\Psi}_2$ is then updated iteratively. The whole process of updating $\underline{\mathbf{S}},\ \mathbf{\Psi}_1,\ \mathbf{\Psi}_2$ is repeated to obtain the final solution of (35).

The uncoupled version of the proposed cTKSVD method (denoted by TKSVD) can be easily obtained by modifying the problem in (35) to:

$$\min_{\mathbf{\Psi}_1, \mathbf{\Psi}_2, \underline{\mathbf{S}}} ||\underline{\mathbf{X}} - \underline{\mathbf{S}} \times_1 \mathbf{\Psi}_1 \times_2 \mathbf{\Psi}_2||_F^2,\ s.t.\ \forall i,\ ||\mathbf{S}_i||_0 \leq K, \tag{49}$$

and it can be solved following the same procedures as described previously for cTKSVD except that the steps of pseudo-inverse and normalization are no longer needed.

Compared to the KHOSVD method [37], the proposed cTKSVD for multidimensional dictionary learning has lower complexity. As mentioned in Section IV-A, in each iteration of the outer loop of KHOSVD, $\hat{N}_1 \hat{N}_2$ inner iterations are performed - this involves duplicated updating of the dictionary atoms because there are only $\hat{N}_1 + \hat{N}_2$ atoms in total. However, for the proposed cTKSVD approach, the duplications are removed by fully considering the multidimensional structure. Since only $\hat{N}_1 + \hat{N}_2$ inner iterations are required for each iteration of the outer loop, cTKSVD requires HOSVD to be executed $\hat{N}_1 \hat{N}_2 - \hat{N}_1 - \hat{N}_2$ fewer times than for the KHOSVD method and hence reduces the complexity. In addition, KHOSVD does not take into account the influence of the sensing matrix. The benefit of coupling of the sensing matrices in cTKSVD will be shown by simulations in Section V-B.

For the cases where $n > 2$, cTKSVD can be formulated following a similar strategy as that employed in (35), and the problem can then be solved following similar steps to those introduced earlier in this section.

We have now derived the method of learning the sparsifying dictionaries when the multilinear sensing matrix is fixed. Using an alternative optimization method, we can then jointly optimize $\mathbf{\Phi}_1,\ \mathbf{\Phi}_2$ and $\mathbf{\Psi}_1,\ \mathbf{\Psi}_2$. The overall procedure is summarized in Algorithm 3. This algorithm works as follows. First, the dictionaries $\mathbf{\Psi}_1,\ \mathbf{\Psi}_2$ are fixed, and the sensing matrices $\mathbf{\Phi}_1,\ \mathbf{\Phi}_2$ are optimized using the methods elaborated in Section III-B. Then the sensing matrices are fixed and the dictionaries are learnt using cTKSVD detailed previously in this section. This process is repeated until convergence.

V. Experimental Results

In this section, we evaluate the proposed approaches via simulations using both synthetic data and real images. We first test the sensing matrix design approaches proposed in Section III-B with the sparsifying dictionaries being given. Then the

Algorithm 3 Joint Optimization

Input: $\mathbf{\Psi}_i^{(0)}$ $(i = 1,\ 2)$, $\mathbf{\Phi}_i^{(0)}$ $(i = 1,\ 2)$, $\underline{\mathbf{X}}$, α, β, η, γ, $iter = 0$.
Output: $\hat{\mathbf{\Phi}}_i$ $(i = 1,\ 2)$, $\hat{\mathbf{\Psi}}_i$ $(i = 1,\ 2)$.
1: **Repeat until convergence:**
2: For $\hat{\mathbf{\Psi}}_i^{(iter)}$ $(i = 1,\ 2)$ fixed, optimize $\hat{\mathbf{\Phi}}_i^{(iter+1)}$ $(i = 1,\ 2)$ using one of the approaches given in Section III-B;
3: For $\hat{\mathbf{\Psi}}_i^{(iter)}$, $\hat{\mathbf{\Phi}}_i^{(iter+1)}$ $(i = 1,\ 2)$ fixed, solve (35) using TOMP to obtain $\hat{\underline{\mathbf{S}}}$;
4: **For** $p_1 = 1$ to $\hat{N}_1$
5: Compute $\underline{\tilde{\mathbf{R}}}_{p_1}$ using (35) - (38);
6: Do HOSVD to $\underline{\tilde{\mathbf{R}}}_{p_1}$ to obtain $\lambda_{\mathbf{R}}^1$, $\mathbf{u}_{\mathbf{R}}^1$, $\mathbf{v}_{\mathbf{R}}^1$ and $\boldsymbol{\omega}_{\mathbf{R}}^1$;
7: Update $(\hat{\boldsymbol{\psi}}_1^{(iter+1)})_{p_1}$, $\mathbf{D}_2\tilde{\mathbf{S}}_{p_{1,:,:}}$ using (41) - (43) and calculate $\tilde{\mathbf{S}}_{p_{1,:,:}}$ by LS;
8: **end**
9: **For** $p_2 = 1$ to $\hat{N}_2$
10: Compute $\underline{\tilde{\mathbf{R}}}_{p_2}$ using (35) and (44);
11: Do HOSVD to $\underline{\tilde{\mathbf{R}}}_{p_2}$ to obtain $\lambda_{\mathbf{R}}^1$, $\mathbf{u}_{\mathbf{R}}^1$, $\mathbf{v}_{\mathbf{R}}^1$ and $\boldsymbol{\omega}_{\mathbf{R}}^1$;
12: Update $(\hat{\boldsymbol{\psi}}_2^{(iter+1)})_{p_2}$, $\mathbf{D}_1\tilde{\mathbf{S}}_{:,p_{2,:}}$ using (46) - (48) and calculate $\tilde{\mathbf{S}}_{:,p_{2,:}}$ by LS;
13: **end**
14: $iter = iter + 1$;

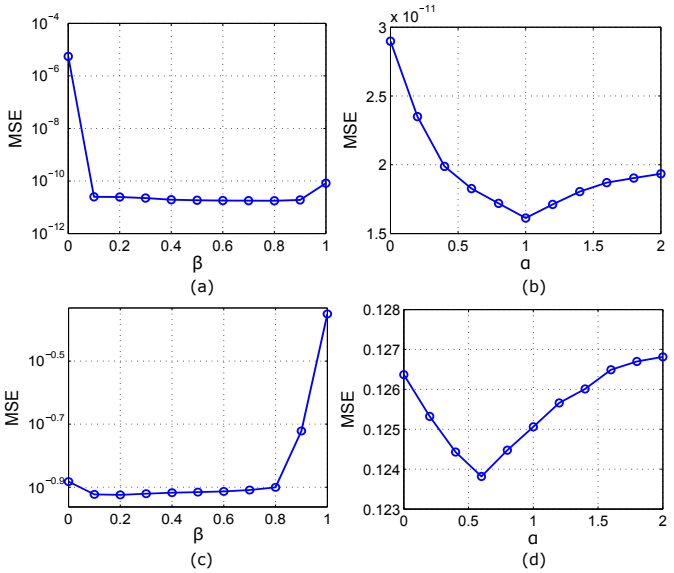


Fig. 1: MSE performance of sensing matrices generated by Approach II with different values of α and β. (a) $\sigma^2 = 0,\ \alpha = 1$; (b) $\sigma^2 = 0,\ \beta = 0.8$; (c) $\sigma^2 = 10^{-2},\ \alpha = 1$; (d) $\sigma^2 = 10^{-2},\ \beta = 0.2$.

cTKSVD approach is evaluated when the sensing matrices are fixed. Finally the experiments for the joint optimization of the two are presented.

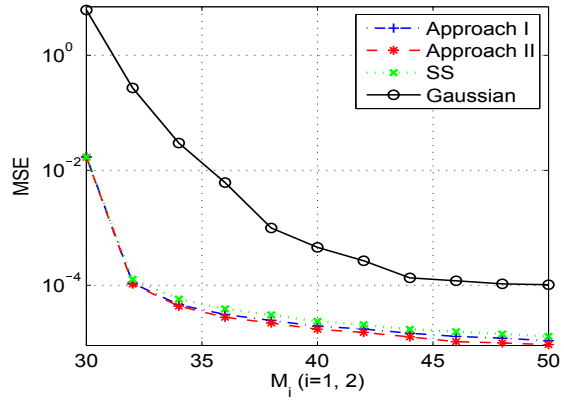
A. Optimal Multidimensional Sensing Matrix

This section is intended to examine the proposed separable Approach I and non-separable Approach II for multidimensional sensing matrix design. Before doing so, we first test the tuning parameters for Approach II, i.e., the non-separable design approach presented in Section III-B-2. As detailed in Section III-B-1, Approach I has a closed form solution and so there are no tuning parameters involved.

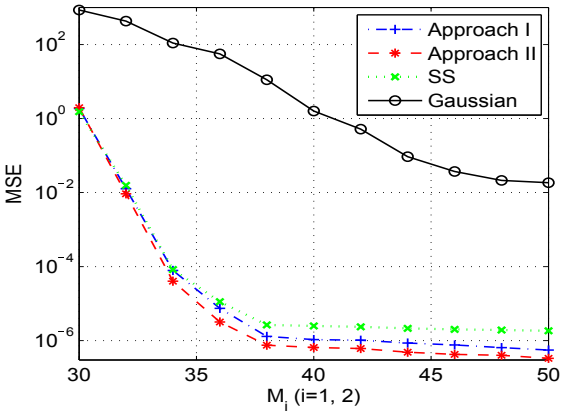
We evaluate the Mean Squared Error (MSE) performance of different sensing matrices generated using Approach II with various parameters and the results are reported by averaging over 500 trials. A random 2D signal $\mathbf{S} \in \mathbb{R}^{64\times 64}$ with sparsity $K = 80$ is generated, where the randomly placed non-zero elements follow an i.i.d zero-mean unit-variance Gaussian distribution. Both the dictionaries $\mathbf{\Psi}_i \in \mathbb{R}^{64\times 256}$ $(i = 1,\ 2)$ and the initial sensing matrices $\mathbf{\Phi}_i \in \mathbb{R}^{40\times 64}$ $(i = 1,\ 2)$ are generated randomly with i.i.d zero-mean unit-variance Gaussian distributions, and the dictionaries are then column normalized while the sensing matrices are normalized by: $\mathbf{\Phi}_i = \sqrt{64}\mathbf{\Phi}_i/||\mathbf{\Phi}_i||_F$. When taking measurements, random additive Gaussian noise with variance σ^2 is induced. A constant step size $\eta = 1e-7$ is selected empirically for Approach II and the BP solver SPGL1 [54] is employed for the reconstructions.

Fig. 1 illustrates the results for the parameter tests. In Fig. 1 (a) and (c), the parameter β is evaluated for the noiseless ($\sigma^2 = 0$) and high noise ($\sigma^2 = 10^{-2}$) cases, respectively, when $\alpha = 1$. From both (a) and (c), we can see that when $\beta = 0$ or 1, the MSE is larger than that for the other values, which means that both terms of Approach II that are controlled by β are essential for obtaining optimal sensing matrices. In addition, we can see that when β becomes larger in the range of $[0.1,\ 0.9]$, the MSE decreases slightly in (a), but increases slightly in (c). This indicates the choice of β under different conditions of sensing noise, which is consistent with that observed in [20]. Even so, owing to the extra term concerning the sensing energy involved in our approach, the MSE performance is much less sensitive to the choice of β than that in [20]. In the remaining experiments, we take $\beta = 0.8$ when sensing noise is low and $\beta = 0.2$ when the noise is high. Fig. 1 (b) and (d) demonstrate the MSE results for the tests of parameter α. It is observed that $\alpha = 1$ is optimal for the noiseless case while $\alpha = 0.6$ when high noise exists. Therefore a larger α is preferred when low noise is involved, which needs to be reduced accordingly when the noise becomes higher.

We then proceed to examine the performance of both the proposed approaches. As this is the first work to optimize the multidimensional sensing matrix, we take the i.i.d Gaussian sensing matrices that are commonly used in CS problems for comparison. Besides, since Sapiro's approach [18] has the same spirit to that of Approach I (as reviewed in Section III-A), it can be easily extended to the multidimensional case, i.e., individually generating $\mathbf{\Phi}_i$ $(i = 1,\ 2)$ using the approach in [18]. We hence also include it in the comparisons and denote it by Separable Sapiro's approach (SS). The previously described synthetic data is generated for the experiments and akin to [8]–[10], [12], BP and OMP are selected as representative of the

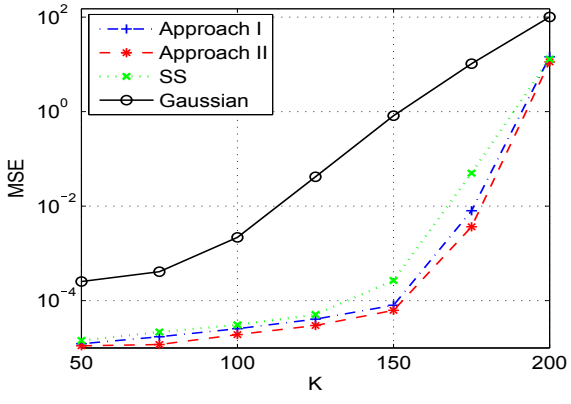


(a)

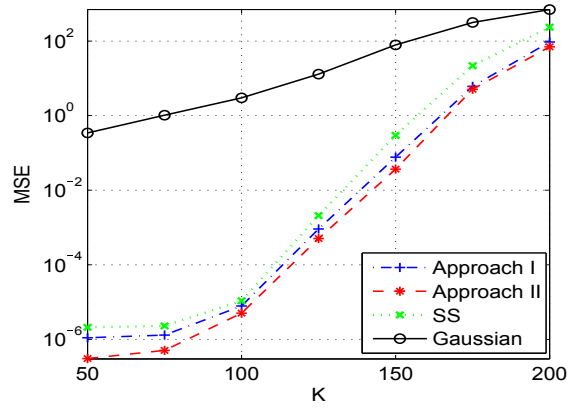


(b)

Fig. 2: MSE performance of different sensing matrices for (a) BP, (b) OMP when M_i $(i = 1,\ 2)$ varies. ($K = 80,\ N_1 = N_2 = 64,\ \hat{N}_1 = \hat{N}_2 = 256$ and $\sigma^2 = 10^{-4}$)



(a)



(b)

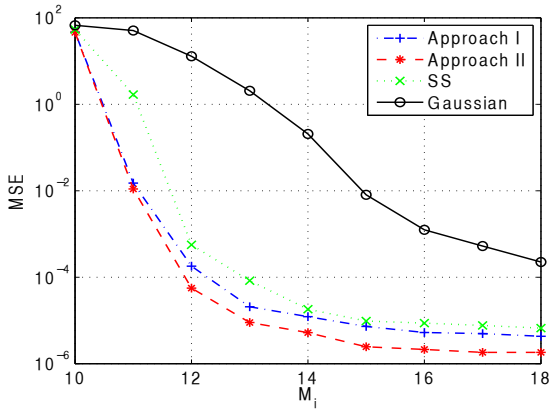
Fig. 3: MSE performance of different sensing matrices for (a) BP, (b) OMP when K varies. ($M_1 = M_2 = 40,\ N_1 = N_2 = 64,\ \hat{N}_1 = \hat{N}_2 = 256$ and $\sigma^2 = 10^{-4}$)

main CS solvers to investigate the reconstruction performance.

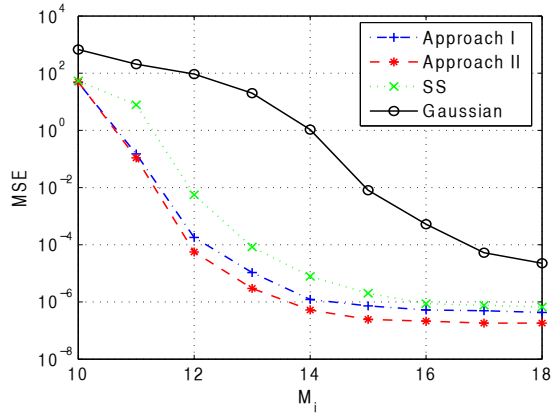
Different sensing matrices are first evaluated using BP and OMP when the number of measurements varies. A small amount of noise ($\sigma^2 = 10^{-4}$) is added when taking measurements and the parameters are chosen as: $\alpha = 1,\ \beta = 0.8$. From Fig. 2, it can be observed that both the proposed approaches perform much better than the Gaussian sensing matrices, among which Approach II has better performance. In general, the SS method performs worse than Approach I, although the difference is not always easily observable. Note that SS is an iterative method while Approach I is non-iterative.

The proposed approaches are again observed to be superior to the other methods when the number of measurements is fixed but the signal sparsity K is varied, as shown in Fig. 3. Compared to Approach I, Approach II exhibits better performance, but at the cost of higher computational complexity and the proper choice of the parameters. If we assume that $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$, $\mathbf{\Psi} \in \mathbb{R}^{N\times \hat{N}}$ and the rank of $\mathbf{\Psi}$ is $\bar{N}$, the complexity of Approach I is then $\mathcal{O}(N^2\bar{N})$, while Approach II has the complexity of $\mathcal{O}(M\hat{N}^2)$ in each of its iterations and it is consequently more complex than Approach I. The complexity of the SS approach is $\mathcal{O}(N^3)$.

We also examine the proposed approaches for 3D TCS problems, i.e., 3D tensor signals are measured using three optimally designed sensing matrices following the TCS model. The 3D signals $\underline{\mathbf{S}} \in \mathbb{R}^{20\times 20\times 20}$ are randomly generated with sparsity $K = 80$ and the non-zero elements follow an i.i.d zero-mean unit-variance Gaussian distribution. We generate the dictionaries $\mathbf{\Psi}_i \in \mathbb{R}^{20\times 80}$ $(i = 1,\ 2,\ 3)$ randomly with i.i.d zero-mean unit-variance Gaussian distributions, and the other experimental settings remain the same as used previously for the 2D case. The MSE performance when the number of measurements varies is shown in Fig. 4, where the advantage of the proposed approaches is still evident.

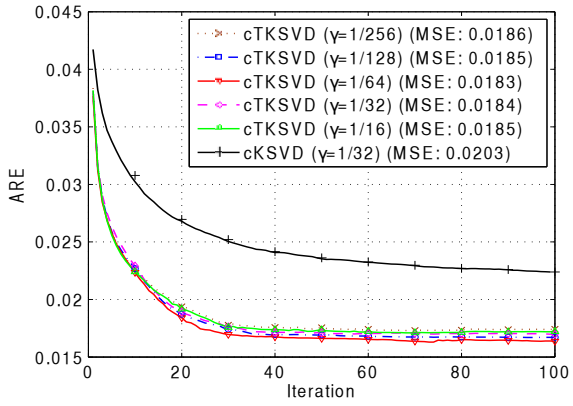


(a)

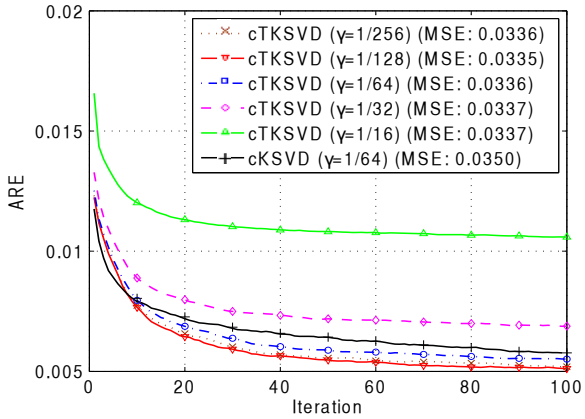


(b)

Fig. 4: MSE performance of different sensing matrices (3D) for (a) BP, (b) OMP when M_i $(i = 1,\ 2,\ 3)$ varies. ($K = 80,\ N_1 = N_2 = N_3 = 20,\ \hat{N}_1 = \hat{N}_2 = \hat{N}_3 = 80$ and $\sigma^2 = 10^{-4}$)



(a)



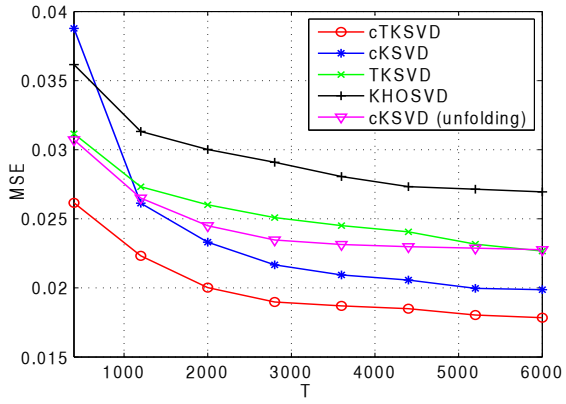
(b)

Fig. 5: Convergence behavior of cTKSVD with different values of γ compared to that of cKSVD with its optimal parameter setting when (a) $M_1 = M_2 = 7$; (b) $M_1 = M_2 = 3$.

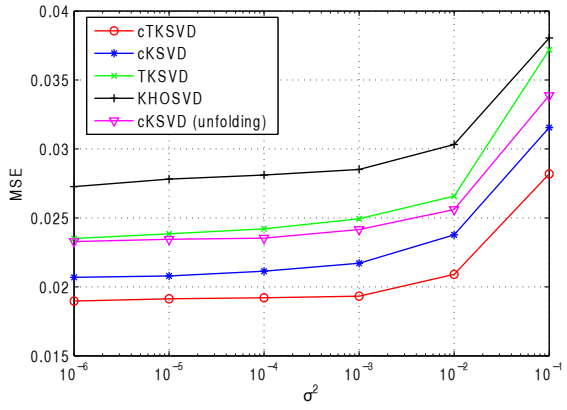
B. Optimal Multidimensional Dictionary with the Sensing Matrices Coupled

In this section, we evaluate the proposed cTKSVD method with a given multidimensional sensing matrix. A training sequence of 5000 2D signals ($T = 5000$) is generated, i.e., $\underline{\mathbf{S}} \in \mathbb{R}^{18\times18\times5000}$, where each signal has $K = 4$ (2×2) randomly placed non-zero elements that follow an i.i.d zero-mean unit-variance Gaussian distribution. The dictionaries $\mathbf{\Psi}_i \in \mathbb{R}^{10\times18}$ $(i = 1,\ 2)$ are also drawn from i.i.d Gaussian distributions, followed by normalization such that they have unit-norm columns. The time-domain training signals $\underline{\mathbf{X}} \in \mathbb{R}^{10\times10\times5000}$ are then formed by: $\underline{\mathbf{X}} = \underline{\mathbf{S}} \times_1 \mathbf{\Psi}_1 \times_2 \mathbf{\Psi}_2$. The test data of size $10 \times 10 \times 5000$ is generated following the same procedure. Random Gaussian noise with variance σ^2 is added to both the training and test data. Two i.i.d random Gaussian matrices are employed as the sensing matrices $\mathbf{\Phi}_i \in \mathbb{R}^{M_i\times10}$ $(i = 1,\ 2)$, normalized by: $\mathbf{\Phi}_i = \sqrt{10}\mathbf{\Phi}_i/||\mathbf{\Phi}_i||_F$. TOMP [34] is utilized in both the training stage and the reconstructions of the test stage for tensor-based approaches and OMP is employed for the vector-based approaches.

We first investigate the convergence behavior of the c-TKSVD approach and examine the choice of the parameter γ. We define the Average Representation Error (ARE) [14], [19] of cTKSVD as: $\sqrt{||\underline{\mathbf{Z}} - \underline{\mathbf{S}} \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2||_F^2/len(\underline{\mathbf{Z}})}$, where $\underline{\mathbf{Z}}$, $\mathbf{D}_1$ and $\mathbf{D}_2$ have the same definitions as in (35). Fig. 5 shows the ARE of cTKSVD at different numbers of iterations for different values of γ (these values are selected based on the values used in [18]). The cKSVD method [18] (reviewed in Section III-A) is also tested and only the results of the optimal γ are displayed in Fig. 5. Note that cKSVD learns a single dictionary $\mathbf{\Psi} \in \mathbb{R}^{100\times324}$, rather than the separable multilinear dictionaries $\mathbf{\Psi}_i \in \mathbb{R}^{10\times18}$ $(i = 1,\ 2)$. The ARE of cKSVD is thus modified accordingly as: $\sqrt{||\mathbf{Z} - \mathbf{DS}||_F^2/len(\mathbf{Z})}$, in which the symbols follow the definitions in (30). From Fig. 5, it can be seen that cTKSVD exhibits stable convergence



(a)



(b)

Fig. 6: MSE performance of different dictionaries when (a) T varies ($\sigma^2 = 0$), (b) σ^2 varies ($T = 5000$). ($K = 4,\ M_1 = M_2 = 7,\ N_1 = N_2 = 10,\ \hat{N}_1 = \hat{N}_2 = 18$)

behavior with different parameters. It converges to a lowest ARE with $\gamma = 1/64$ when $M_i = 7$ and the optimal γ is $1/128$ when $M_i = 3$. The reconstruction MSE values are also shown in the legend, which are similar to each other but reveal the same optimal choice of γ. Thus the optimal γ is lower when the number of measurements decreases, which is consistent with the observation in [18]. Even so, noting the similarity of the MSE values, we observe that the algorithm is not very sensitive to the choice of γ. In both experiments, cTKSVD with optimal γ outperforms cKSVD in terms of ARE and MSE.

Then the MSE performance of dictionaries learned by cTKSVD is compared with that of cKSVD [18] and KHOSVD [37] when the number of training sequences T and the noise variance σ^2 vary. We use $\gamma = 1/64$ for cTKSVD and $\gamma = 1/32$ for cKSVD. To see the benefit of coupling sensing matrices, we also evaluate the uncoupled version of the proposed approach, i.e., TKSVD, in the experiments. In addition, since cKSVD is for vectorized signals that form

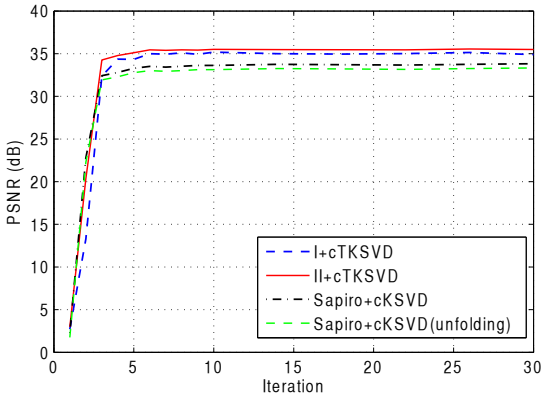


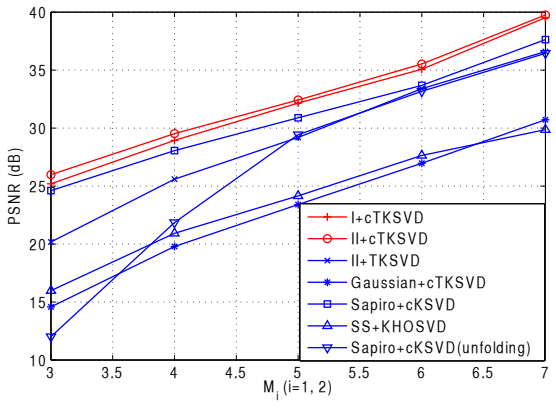
Fig. 7: Convergence behavior of various joint optimization methods. ($T = 5000,\ K = 4,\ M_1 = M_2 = 6,\ N_1 = N_2 = 8,\ \hat{N}_1 = \hat{N}_2 = 16,\ \sigma^2 = 0$)

a training matrix, we also separately implement the cKSVD algorithm for unfoldings of the data tensor, and then build the Kronecker dictionary using the obtained dictionaries (denoted by "cKSVD (unfolding)"). The results can be found in Fig. 6. It is observable that cTKSVD outperforms all the other methods in terms of the reconstruction MSE. The sensing-matrix-coupled approaches (cKSVD, cKSVD (unfolding) and cTKSVD) are superior to the uncoupled approaches (TKSVD and KHOSVD). The TKSVD method leads to smaller MSEs compared to KHOSVD, as it fully exploits the multidimensional structure. In addition, since cKSVD is not an approach that explicitly considers a multidimensional dictionary, it requires longer training sequences to learn the multilinear structure from the vectorized data. As seen in Fig. 6 (a), to achieve a MSE of 0.02, cTKSVD only needs 2000 items of training data; while approximately 6000 is required for the cKSVD approach. For the same reason, the performance of cKSVD degrades dramatically when the training data is less than 1000. The cKSVD (unfolding) approach does not suffer from this problem, because it is explicitly implemented for the tensor unfoldings.

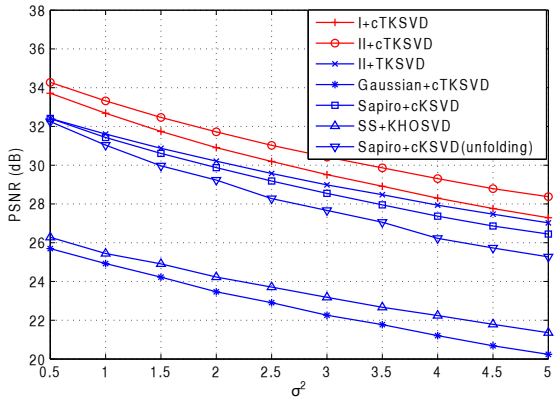
Regarding the computational complexities of these approaches, we can compare them in the various stages. First, in the sparse reconstruction stage, cKSVD and KHOSVD employ the OMP algorithm, while TKSVD and cTSKVD utilize the TOMP approach, which is less complex than the former (as computed in [35]). Then, in the atom update steps, cKSVD and KHOSVD both require $N_1 N_2$ iterations, while only $N_1 + N_2$ is needed for TKSVD and cTSKVD. However, it is noticed that the higher order SVD approach, i.e., HOSVD, employed in KHOSVD, TKSVD and cTKSVD is more complex than the conventional SVD (since HOSVD is performed by multiple SVDs).

C. TCS with Jointly Optimized Sensing Matrix and Dictionary

Now we examine the performance of the proposed joint optimization approach in Algorithm 3. The training data consists



(a)



(b)

Fig. 8: PSNR performance of different methods when (a) M_i $(i = 1,\ 2)$ varies $(\sigma^2 = 0)$, (b) σ^2 varies $(M_1 = M_2 = 6)$. $(T = 5000,\ K = 4,\ M_1 = M_2 = 6,\ N_1 = N_2 = 8,\ \hat{N}_1 = \hat{N}_2 = 16)$

of 5000 8×8 patches obtained by randomly extracting 25 patches from each of the 200 images in a training set from the Berkeley segmentation dataset [55]. The test data is obtained by extracting non-overlapping 8×8 patches from the other 100 images (different from the images for training) in the dataset. A 2D Discrete Cosine Transform (DCT) is employed to initialize the dictionaries $\mathbf{\Psi}_i \in \mathbb{R}^{8\times 16}$ $(i = 1,\ 2)$ and i.i.d Gaussian matrices are used as the initial sensing matrices $\mathbf{\Phi}_i \in \mathbb{R}^{M_i \times 8}$ $(i = 1,\ 2)$. We employ TOMP for reconstruction and the Peak Signal to Noise Ratio (PSNR) is used as the evaluation criteria.

In the first experiment, we examine the convergence behavior of Algorithm 3 when the proposed Approach I and II are utilized for the sensing matrix optimization step (respectively denote by I + cTKSVD and II + cTKSVD). We take $M_1 = M_2 = 6$ and no noise is added to the measurements at the test stage, i.e., $\sigma^2 = 0$. By conducting the simulations performed previously to obtain the results in Fig. 1 and 5, the parameters are chosen as: $\alpha = 3,\ \beta = 0.8,\ \gamma = 1/8$. The step size for II + cTKSVD is set empirically as: $\eta = 1e-5$. The PSNR performance for different numbers of iterations is illustrated in Fig. 7. Since Sapiro's approach in [18] also jointly optimizes the sensing matrix and dictionary, we include it in this figure (denoted by Sapiro's + cKSVD). The parameter γ is optimal at $1/2$ for cKSVD under our settings. However, note that Sapiro's approach is only for vectorized signals in the conventional CS problem, i.e., a single sensing matrix $\mathbf{\Phi} \in \mathbb{R}^{36\times 64}$ and a dictionary $\mathbf{\Psi} \in \mathbb{R}^{64\times 256}$ are obtained. We also include the Sapiro + cKSVD (unfolding) approach, i.e., the implementation of Sapiro + cKSVD for the unfoldings of the data tensor, in the comparison. From Fig. 7, we can see that the proposed approaches outperform Sapiro's approach in both forms in the sense that their PSNR values are higher at convergence while all the methods converge in less than 10 iterations.

Then the proposed approaches are compared with various other approaches when the number of measurements (M_i $(i = 1,\ 2)$) and the noise variance (σ^2) vary. Specifically, using the notation employed previously and by denoting the method of combining sensing matrix design with that of the dictionary learning using a "+", the methods for comparison are: II + TKSVD, Gaussian + cTKSVD, Sapiro + cKSVD, SS + KHOSVD and Sapiro + cKSVD (unfolding), where the Sapiro + cKSVD (unfolding) approach refers to the implementation of Sapiro + cKSVD for the unfoldings of the data tensor. In these approaches, II + TKSVD and SS + KHOSVD are uncoupled methods; Gaussian + cTKSVD does not involve sensing matrix optimization; Sapiro + cKSVD is for conventional CS system only.

The results are shown in Fig. 8. We can see that the proposed approaches obtain higher PSNR values than all of the other methods and II + cTKSVD performs best. To see the gain of coupling sensing matrices during dictionary learning and optimizing the sensing matrices, respectively, we compare II + cTKSVD with II + TKSVD and Gaussian + cTKSVD. For instance, when $M_i = 5,\ \sigma^2 = 0$, II + cTKSVD has a gain of about 3dB over II + TKSVD and nearly 9dB over Gaussian + cTKSVD. Although Sapiro + cKSVD has a similar performance to ours at some specific settings, it is not for a TCS system that requires multiple separable sensing matrices. Even though we can implement it for the tensor unfoldings, i.e., the Sapiro + cKSVD (unfolding) approach, its performance is not superior to that of Sapiro + cKSVD. Examples of reconstructed images using these methods are demonstrated in Fig. 9 and 10 with the corresponding PSNR values listed. All of the conducted simulations verify that the proposed methods of multidimensional sensing matrix and dictionary optimization improve the performance of a TCS system.

VI. Conclusions

In this paper, we propose to jointly optimize the multidimensional sensing matrix and dictionary for TCS systems. To obtain the optimized sensing matrices, a separable approach with

Fig. 9: Reconstruction example when $M_1 = M_2 = 6$. The images from left to right, top to bottom and their PSNR (dB) values are: Original, II+cTKSVD (35.41), I+cTKSVD (34.97), Sapiro+cKSVD (33.64), II+TKSVD (33.57), S-apiro+cKSVD (unfolding) (33.44), SS+KHOSVD (28.62), Gaussian+cTKSVD (28.05).

Fig. 10: Reconstruction example when $M_1 = M_2 = 4$. The images from left to right, top to bottom and their PSNR (dB) values are: Original, II+cTKSVD (29.91), I+cTKSVD (29.45), Sapiro+cKSVD (28.72), II+TKSVD (26.60), S-apiro+cKSVD (unfolding) (23.51), SS+KHOSVD (22.62), Gaussian+cTKSVD (21.94).

closed form solutions has been presented and a joint iterative approach with novel design measures has also been proposed. The iterative approach certainly has higher complexity, but also exhibits better performance. An approach to learning the multidimensional dictionary has been designed, which explicitly takes the multidimensional structure into account and removes the redundant updates in the existing multilinear approaches in the literature. Further improvement is obtained by coupling the multidimensional sensing matrix while learning the dictionary. The performance advantage of the proposed approaches has been demonstrated by experiments using both synthetic data and real images.

Appendix A
Proof of Theorem 3

Assume $\mathbf{A}_i = \mathbf{\Phi}_i\mathbf{\Psi}_i = \mathbf{U}_{\mathbf{A}_i}[\ \mathbf{\Lambda}_{\mathbf{A}_i}\ \ \mathbf{0}\ ]\mathbf{V}_{\mathbf{A}_i}^T$ is an SVD of $\mathbf{A}_i$ for $i = 1,\ 2$ and $rank(\mathbf{A}_i) = M_i$. Then the objective we want to minimize in (19) can be rewritten as:

$$\left|\left|\mathbf{I}_{\hat{N}_1\hat{N}_2} - (\mathbf{V}_{\mathbf{A}_2}\begin{bmatrix}\mathbf{\Lambda}_{\mathbf{A}_2}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{V}_{\mathbf{A}_2}^T) \otimes (\mathbf{V}_{\mathbf{A}_1}\begin{bmatrix}\mathbf{\Lambda}_{\mathbf{A}_1}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{V}_{\mathbf{A}_1}^T)\right|\right|_F^2.$$

Denote $\mathbf{\Sigma} = \begin{bmatrix}\mathbf{\Lambda}_{\mathbf{A}_2}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} \otimes \begin{bmatrix}\mathbf{\Lambda}_{\mathbf{A}_1}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} = diag(\boldsymbol{\nu}_{\mathbf{A}_2} \otimes \boldsymbol{\nu}_{\mathbf{A}_1})$, $\boldsymbol{\nu}_{\mathbf{A}_i} = diag(\begin{bmatrix}\mathbf{\Lambda}_{\mathbf{A}_i}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix})$, then we have

$$||\mathbf{I}_{\hat{N}_1\hat{N}_2} - (\mathbf{V}_{\mathbf{A}_2} \otimes \mathbf{V}_{\mathbf{A}_1})\mathbf{\Sigma}(\mathbf{V}_{\mathbf{A}_2}^T \otimes \mathbf{V}_{\mathbf{A}_1}^T)||_F^2. \quad (50)$$

Let $\boldsymbol{\nu}_{\mathbf{A}_i} = [(v_i)_1, ..., (v_i)_{M_i}, \mathbf{0}]^T$, then the sub-vector of the diagonal of $\mathbf{\Sigma}$ containing its non-zero values is: $\hat{\boldsymbol{\nu}} = [(v_2)_1(v_1)_1, ..., (v_2)_1(v_1)_{M_1}, ..., (v_2)_{M_2}(v_1)_1, ..., (v_2)_{M_2}(v_1)_{M_1}]^T$. Thus (50) becomes:

$$||\mathbf{I}_{\hat{N}_1\hat{N}_2} - \mathbf{\Sigma}||_F^2 = \hat{N}_1\hat{N}_2 - M_1M_2 + \sum_{p=1}^{M_2}\sum_{q=1}^{M_1}(1-(v_2)_p(v_1)_q)^2. \quad (51)$$

Therefore we can obtain that the minimum value of (19) is $\hat{N}_1\hat{N}_2 - M_1M_2$, and that it is achieved when the entries of $\hat{\boldsymbol{\nu}}$ are all unity.

Clearly $\mathbf{\Lambda}_{\mathbf{A}_i}$=$\mathbf{I}_{M_i}$ for $i = 1,\ 2$ is a solution, i.e., $\mathbf{A}_i = \mathbf{U}_{\mathbf{A}_i}[\ \mathbf{I}_{M_i}\ \ \mathbf{0}\ ]\mathbf{V}_{\mathbf{A}_i}^T$ with $\mathbf{U}_{\mathbf{A}_i} \in \mathbb{R}^{M_i\times M_i}$ and $\mathbf{V}_{\mathbf{A}_i} \in \mathbb{R}^{\hat{N}_i\times\hat{N}_i}$ being arbitrary orthonormal matrices. Then we would like to find $\mathbf{\Phi}_i$ $(i = 1,\ 2)$ such that $\mathbf{\Phi}_i\mathbf{\Psi}_i = \mathbf{U}_{\mathbf{A}_i}[\ \mathbf{I}_{M_i}\ \ \mathbf{0}\ ]\mathbf{V}_{\mathbf{A}_i}^T$. Following the derivation of Theorem 2 in [11], the solution in (20) can be found.

With this solution, for an arbitrary vector $\mathbf{z} \in \mathbb{R}^{\hat{N}_i}$, we have $||\mathbf{A}_i^T\mathbf{z}||_2^2 = tr(\mathbf{z}^T\mathbf{A}_i\mathbf{A}_i^T\mathbf{z}) = tr(\mathbf{z}^T\mathbf{z}) = ||\mathbf{z}||_2^2$, which indicates that the resulting equivalent sensing matrices $\mathbf{A}_i$ $(i = 1,\ 2)$ are Parseval tight frames. In addition, we observe that the solution in (20) can be obtained by separately solving the sub-problems in (21), of which the solutions have been derived in [11]. By substituting the solutions of the sub-problems into (19), we can conclude the minimum remains as $\hat{N}_1\hat{N}_2 - M_1M_2$.

REFERENCES

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[3] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Magazine*, vol. 25, no. 2, pp. 21 –30, March 2008.

[4] M. Lustig, D. J. Donoho, and J. Pauly., "Sparse MRI: the application of compressed sensing for rapid MR imaging," *Magn Reson Med*, vol. 58, pp. 1182–1195, 2007.

[5] W. Chen and I.J. Wassell, "Energy-efficient signal acquisition in wireless sensor networks: a compressive sensing framework," *IET Wireless Sensor Systems*, vol. 2, no. 1, pp. 1–8, March 2012.

[6] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Information Theory*, vol. 52, no. 1, pp. 6–18, 2006.

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

[8] M. Elad, "Optimized projections for compressed sensing," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5695–5702, 2007.

[9] J. Xu, Y. Pi, and Z. Cao, "Optimized projection matrix for compressive sensing," *EURASIP J. on Advances in Signal Process.*, vol. 2010, pp. 43, 2010.

[10] W. Chen, M. R. D Rodrigues, and I. Wassell, "Projection design for statistical compressive sensing: A tight frame based approach," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 2016–2029, 2013.

[11] G. Li, Z. Zhu, D. Yang, L. Chang, and H. Bai, "On projection matrix optimization for compressive sensing systems," *IEEE Trans. Signal Process.*, vol. 61, no. 11, pp. 2887–2898, 2013.

[12] N. Cleju, "Optimized projections for compressed sensing via rank-constrained nearest correlation matrix," *Applied. Comput. Harmonic Analysis*, vol. 36, no. 3, pp. 495–507, 2014.

[13] K. Engan, S. O. Aase, and J. H. Husøy, "Multi-frame compression: Theory and design," *EURASIP J. Signal Process.*, vol. 80, no. 10, pp. 2121–2140, 2000.

[14] M. Aharon, M. Elad, and A. Bruckstein, "The KSVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[15] I. Tošić and P. Frossard, "Dictionary learning: What is the right representation for my signals," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, 2011.

[16] S. K. Sahoo and A. Makur, "Dictionary training for sparse representation as generalization of K-means clustering," *IEEE Signal Process. Letters*, vol. 20, no. 6, pp. 587–590, 2013.

[17] W. Dai, T. Xu, and W. Wang, "Simultaneous codeword optimization (simco) for dictionary update and learning," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6340–6353, 2012.

[18] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1395–1408, 2009.

[19] W. Chen and M. R. D. Rodrigues, "Dictionary learning with optimized projection design for compressive sensing applications," *IEEE Signal Process. Letters*, vol. 20, no. 10, pp. 992–995, 2013.

[20] H. Bai, G. Li, S. Li, Q. Li, Q. Jiang, and L. Chang, "Alternating optimization of sensing matrix and sparsifying dictionary for compressed sensing," *IEEE Trans. Signal Process.*, vol. 63, no. 6, pp. 1581–1594, 2015.

[21] X. Ding, W. Chen, and I. J. Wassell, "Nonconvex compressive sensing reconstruction for tensor using structures in modes," *IEEE ICASSP*, May 2016.

[22] S. Oymak, A. Jalali, M. Fazel, Y. C Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Trans. Information Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

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

[24] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, pp. 111–119, June 2012.

[25] M. Golbabaee and P. Vandergheynst, "Compressed sensing of simultaneous low-rank and joint-sparse matrices," *arXiv*, 2012.

[26] R. Chartrand, "Nonconvex splitting for regularized low-rank+ sparse decomposition," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5810–5819, 2012.

[27] R. Otazo, E. Candès, and D. K. Sodickson, "Low-rank plus sparse matrix decomposition for accelerated dynamic mri with separation of background and dynamic components," *Mag. Res. in Medicine*, vol. 73, no. 3, pp. 1125–1136, 2015.

[28] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 494–504, Feb 2012.

[29] Y. Rivenson and A. Stern, "Compressed imaging with a separable sensing operator," *IEEE Signal Process. Letters*, vol. 16, no. 6, pp. 449–452, June 2009.

[30] Y. Rivenson and A. Stern, "Practical compressive sensing of large images," in *proc. DSP*, July 2009, pp. 1–8.

[31] M. F. Duarte and R. G. Baraniuk, "Kronecker product matrices for compressive sensing," *Technical Report, Rice University*, 2011.

[32] N. D. Sidiropoulos and A. Kyrillidis, "Multi-way compressed sensing for sparse low-rank tensors," *IEEE Signal Process. Letters*, vol. 19, no. 11, pp. 757–760, 2012.

[33] S. Friedland, Q. Li, and D. Schonfeld, "Compressive sensing of sparse tensors," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4438–4447, Oct 2014.

[34] C. F. Caiafa and A. Cichocki, "Multidimensional compressed sensing and their applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 6, pp. 355–380, 2013.

[35] C. F. Caiafa and A. Cichocki, "Computing sparse representations of multidimensional signals using kronecker bases," *Neural Comput.*, vol. 25, no. 1, pp. 186–220, Jan. 2013.

[36] M. Seibert, J. Wormann, R. Gribonval, and M. Kleinsteuber, "Separable cosparse analysis operator learning," in *Proc. EUSIPCO*, 2014, pp. 770–774.

[37] F. Roemer, G. D. Galdo, and M. Haardt, "Tensor-based algorithms for learning multidimensional separable dictionaries," in *Proc. IEEE ICASSP*, 2014, pp. 3963–3967.

[38] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, and B. Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *proc. IEEE CVPR*, 2014, pp. 2949–2956.

[39] Y August, C Vachman, Y Rivenson, and A Stern, "Compressive hyperspectral imaging by random separable projections in both the spatial and the spectral domains," *Applied optics*, vol. 52, no. 10, pp. 46–54, 2013.

[40] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, Ting S., K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.,*, vol. 25, no. 2, pp. 83 –91, March 2008.

[41] R. F. Marcia, Z. T. Harmany, and R. M. Willett, "Compressive coded aperture imaging," *SPIE 7246, Comput. Imag. VII*, p. 72460G, 2009.

[42] V. Majidzadeh, L. Jacques, A. Schmid, P. Vandergheynst, and Y. Leblebici, "A (256x256) pixel 76.7mW CMOS imager/ compressor

based on real-time in-pixel compressive sensing," in *in Proc. IEEE ISCAS*, June 2010, pp. 2956 –2959.

[43] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Information Theory*, vol. 51, no. 12, pp. 4203 – 4215, Dec. 2005.

[44] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Information Theory*, vol. 53, no. 12, pp. 4655 –4666, Dec. 2007.

[45] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[46] J. A. Tropp, "On the conditioning of random subdicitionaries," *Appl. Comput. Harmon. Anal.*, vol. 25, pp. 1–24, 2008.

[47] J. A. Tropp, "Norms of random submatrices and sparse approximation," *C. R. Acad. Sci. Paris*, vol. 346, no. 23-24, pp. 1271–1274, 2008.

[48] M. A. Herman and T. Strohmer, "High resolution radar via compressed sensing," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2275–2284, 2009.

[49] G. E. Pfander and H. Rauhut, "Sparsity in time-frequency representations," *J. Fourier Anal. Appl.*, vol. 16, pp. 233–260, 2010.

[50] E. J. Candès and Y. Plan, "Near-ideal model selection by l1 minimization," *Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2177, 2009.

[51] W. Chen, M. R. D. Rodrigues, and I. J. Wassell, "On the use of unit-norm tight frames to improve the average mse performance in compressive sensing applications," *IEEE Signal Process. Letters*, vol. 19, no. 1, pp. 8–11, 2012.

[52] C. F. Van Loan, "The ubiquitous Kronecker product," *J. comput. applied mathematics*, vol. 123, no. 1, pp. 85–100, 2000.

[53] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. on Matrix Analysis. Applications.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[54] E. V. D. Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[55] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE ICCV*, 2001, vol. 2, pp. 416–423.

Xin Ding (S'14) received the B.Eng. degree in electronic engineering from China Agricultural University, Beijing, China, in 2010, the M.S. degree in wireless communications from the University of Bristol, Bristol, UK, in 2011, and the Ph.D. degree in Computer Science from the University of Cambridge, UK, in 2016. Her research interests are in compressive sensing, digital filtering, video processing, and multidimensional signal processing.

Wei Chen (M'13) received the B.Eng. degree and M.Eng. degree in Communications Engineering from Beijing University of Posts and Telecommunications, China, in 2006 and 2009, respectively, and the Ph.D. degree in Computer Science from the University of Cambridge, UK, in 2012. He had been an Research Associate with the Computer Laboratory, University of Cambridge from 2013 to 2016. He is currently a Professor with Beijing Jiaotong University, Beijing, China. His current research interests include sparse representation, Bayesian inference, image processing and wireless sensor networks.

Ian Wassell received the B.Sc. and B.Eng. degrees from the University of Loughborough in 1983, and the Ph.D. degree from the University of Southampton in 1990. He is a Senior Lecturer at the University of Cambridge Computer Laboratory and has in excess of 15 years experience in the simulation and design of radio communication systems gained via a number of positions in industry and higher education. He has published more than 180 papers concerning wireless communication systems and his current research interests include: fixed wireless access, sensor networks, cooperative networks, propagation modelling, compressive sensing and cognitive radio. He is a member of the IET and a Chartered Engineer.