Multi-way Compressed Sensing for Sparse Low-rank Tensors

Nicholas D. Sidiropoulos, *Fellow, IEEE*, and Anastasios Kyrillidis, *Student Member, IEEE*

***Abstract*—For linear models, compressed sensing theory and methods enable recovery of sparse signals of interest from few measurements - in the order of the number of nonzero entries as opposed to the length of the signal of interest. Results of similar flavor have more recently emerged for bilinear models, but no results are available for multilinear models of tensor data. In this contribution, we consider compressed sensing for sparse and low-rank tensors. More specifically, we consider low-rank tensors synthesized as sums of outer products of sparse loading vectors, and a special class of linear dimensionality-reducing transformations that reduce each mode individually. We prove interesting 'oracle' properties showing that it is possible to identify the uncompressed sparse loadings directly from the compressed tensor data. The proofs naturally suggest a two-step recovery process: fitting a low-rank model in compressed domain, followed by per-mode ℓ_0 / ℓ_1 de-compression. This two-step process is also appealing from a computational complexity and memory capacity point of view, especially for big tensor datasets.**

Keywords: Compressed sensing, tensor decomposition, multi-way analysis, CANDECOMP / PARAFAC

I. Introduction

For linear models, *compressed sensing* [1], [2] ideas have made headways in enabling compression down to levels proportional to the number of nonzero elements, well below equations-versus-unknowns considerations. These developments rely on latent sparsity and ℓ_1-relaxation of the ℓ_0 quasi-norm to recover the sparse unknown. Results of similar flavor have more recently emerged for bilinear models [3], [4], but, to the best of the author's knowledge, compressed sensing has not been generalized to higher-way multilinear models of *tensors*, also known as *multi-way arrays* [5]–[10].

In this contribution, we consider compressed sensing for sparse and low-rank tensors. A rank-one matrix is an outer product of two vectors; a rank-one tensor is an outer product of three or more (so-called *loading*) vectors. The rank of a tensor is the smallest number of rank-one tensors that sum up to the given tensor. A rank-one tensor is sparse if and only if one or more of the underlying loadings are sparse. For small enough rank, sparse loadings imply a sparse tensor. With F denoting tensor rank, n_a the number of nonzero elements per loading in one mode, and likewise n_b, n_c for the other modes, the synthesized tensor has at most $n_a n_b n_c F$ nonzero elements. The converse is not necessarily true: sparse tensor $\not\Rightarrow$ sparse loadings in general. On the other hand, the elements of a tensor are multivariate polynomials in the loadings, thus if the loadings are randomly drawn from a jointly continuous distribution, the tensor will not be sparse, almost surely. These considerations suggest that for low-enough rank it is reasonable to model sparse tensors as arising from sparse loadings. We therefore consider low-rank tensors synthesized as sums of outer products of sparse loading vectors, and a special class of linear dimensionality-reducing transformations that reduce each mode individually using a *random* compression matrix. We prove interesting 'oracle' properties showing that it is possible to identify the uncompressed sparse loadings directly from the compressed tensor data. The proofs naturally suggest a two-step recovery process: fitting a low-rank model in compressed domain, followed by per-mode ℓ_0 / ℓ_1 de-compression. This two-step process is also appealing from a computational complexity and memory capacity point of view, especially for big tensor datasets.

Our results appear to be the first to cross-leverage the identifiability properties of multilinear decomposition and compressive sensing. A few references have considered sparsity and incoherence properties of tensor decompositions, notably [11] and [12]. Latent sparsity is considered in [11] as a way to *select* subsets of elements in each mode to form co-clusters, without regard to identifiability properties though. Reference [12] considers identifiability conditions expressed in terms of restricted isometry / incoherence properties of the mode loading matrices; but it does not deal with tensor compression or compressive sensing for tensors.

Notation: A scalar is denoted by an italic letter, e.g. a. A column vector is denoted by a bold lowercase letter, e.g. $\mathbf{a}$ whose i-th entry is $\mathbf{a}(i)$. A matrix is denoted by a bold uppercase letter, e.g., $\mathbf{A}$ with (i,j)-th entry $\mathbf{A}(i,j)$; $\mathbf{A}(:,j)$ ($\mathbf{A}(i,:)$) denotes the j-th column (resp. i-th row) of $\mathbf{A}$. A three-way array is denoted by an underlined bold uppercase letter, e.g., $\underline{\mathbf{X}}$, with (i,j,k)-th entry $\underline{\mathbf{X}}(i,j,k)$. Vector, matrix and three-way array size parameters (mode lengths) are denoted by uppercase letters, e.g. I. $\circ$ stands for the vector outer product; i.e., for two vectors $\mathbf{a}$ $(I \times 1)$ and $\mathbf{b}$ $(J \times 1)$, $\mathbf{a} \circ \mathbf{b}$ is an $I \times J$ rank-one matrix with (i,j)-th element $\mathbf{a}(i)\mathbf{b}(j)$; i.e., $\mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^T$. For three vectors, $\mathbf{a}$ $(I \times 1)$, $\mathbf{b}$ $(J \times 1)$, $\mathbf{c}$ $(K \times 1)$, $\mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$ is an $I \times J \times K$ rank -one three-way array with (i,j,k)-th element $\mathbf{a}(i)\mathbf{b}(j)\mathbf{c}(k)$. The rank of a three-way array $\underline{\mathbf{X}}$ is the smallest number of outer products needed to synthesize $\underline{\mathbf{X}}$. The vec$(\cdot)$ operator stacks the columns of its matrix argument in one tall column; $\otimes$ stands for the Kronecker product; $\odot$ stands for the Khatri-Rao (column-wise Kronecker) product: given $\mathbf{A}$ $(I \times F)$ and $\mathbf{B}$ $(J \times F)$, $\mathbf{A} \odot \mathbf{B}$ is the $JI \times F$ matrix

$$\mathbf{A} \odot \mathbf{B} = \left[\mathbf{A}(:,1) \otimes \mathbf{B}(:,1) \cdots \mathbf{A}(:,F) \otimes \mathbf{B}(:,F)\right]$$

II. Tensor decomposition preliminaries

There are two basic multi-way (tensor) models: Tucker3, and PARAFAC. Tucker3 is generally not identifiable, but it is useful for data compression / interpolation and as an exploratory tool. PARAFAC is identifiable under certain conditions, and is the model of choice when one is interested in unraveling *latent structure*. We refer the reader to [9], [10] for gentle introductions to tensor decompositions and applications. Here we briefly review Tucker3 and PARAFAC to lay the foundation for our main result.

Tucker3: Consider an $I \times J \times K$ three-way array $\underline{\mathbf{X}}$ comprising K matrix slabs $\{\mathbf{X}_k\}_{k=1}^K$, arranged into the tall matrix $\mathbf{X} := [\text{vec}(\mathbf{X}_1), \cdots, \text{vec}(\mathbf{X}_K)]$. The Tucker3 model (see also [13]) can be written as

$$\mathbf{X} \approx (\mathbf{B} \otimes \mathbf{A})\mathbf{G}\mathbf{C}^T,$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, are three *mode loading matrices*, assumed orthogonal without loss of generality, and $\mathbf{G}$ is the so-called *Tucker3 core tensor* $\underline{\mathbf{G}}$ recast in matrix form. The non-zero elements of the core

N.D. Sidiropoulos (contact author) is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA. E-mail nikos@umn.edu, phone: (612) 625-1242, Fax: (612) 625-4583

A. Kyrillidis is with the Laboratory for Information and Inference Systems, Institute of Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne, Switzerland. E-mail: anastasios.kyrillidis@epfl.ch

Submitted to *IEEE Signal Processing Letters*, June 7, 2012; revised July 17, 2012.

tensor determine the interactions between columns of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. The associated model-fitting problem is

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{G}} ||\mathbf{X} - (\mathbf{B} \otimes \mathbf{A})\mathbf{G}\mathbf{C}^T||_F^2,$$

which is usually solved using an alternating least squares procedure. The Tucker3 model can be fully vectorized as $\text{vec}(\mathbf{X}) \approx (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\, \text{vec}(\mathbf{G})$.

PARAFAC: When the core tensor $\underline{\mathbf{G}}$ is constrained to be diagonal (i.e., $\underline{\mathbf{G}}(\ell, m, n) = 0$ if $m \neq \ell$ or $n \neq \ell$), one obtains the parallel factor analysis (PARAFAC) [6], [7] model, sometimes also referred to as canonical decomposition (CANDECOMP) [5], or CP for CANDECOMP-PARAFAC. PARAFAC can be written as a system of matrix equations $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{B}^T$, where $\mathbf{D}_k(\mathbf{C})$ is a diagonal matrix holding the k-th row of $\mathbf{C}$ in its diagonal; or in compact matrix form as $\mathbf{X} \approx (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T$, using the Khatri-Rao product. PARAFAC is in a way the most basic tensor model, because of its direct relationship to tensor rank and the concept of low-rank decomposition or approximation. In particular, employing a property of the Khatri-Rao product,

$$\mathbf{X} \approx (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T \Longleftrightarrow \text{vec}(\mathbf{X}) \approx (\mathbf{C} \odot \mathbf{B} \odot \mathbf{A})\, \mathbf{1},$$

where $\mathbf{1}$ is a vector of all 1's. Equivalently,

$$\underline{\mathbf{X}} \approx \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f,$$

where $\mathbf{a}_f$ is the f-th column of $\mathbf{A}$, and analogously for $\mathbf{b}_f$ and $\mathbf{c}_f$.

The distinguishing feature of the PARAFAC model is its essential uniqueness: under certain conditions, $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ can be identified from $\mathbf{X}$, i.e., they are unique up to permutation and scaling of columns [5]–[8], [14]–[16]. Consider an $I \times J \times K$ tensor $\underline{\mathbf{X}}$ of rank F. In vectorized form, it can be written as the $IJK \times 1$ vector $\mathbf{x} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\, \mathbf{1}$, for some $\mathbf{A}$ $(I \times F)$, $\mathbf{B}$ $(J \times F)$, and $\mathbf{C}$ $(K \times F)$ - a PARAFAC model of size $I \times J \times K$ and order F parameterized by $(\mathbf{A}, \mathbf{B}, \mathbf{C})$. The *Kruskal-rank* of $\mathbf{A}$, denoted $k_{\mathbf{A}}$, is the maximum k such that *any* k columns of $\mathbf{A}$ are linearly independent ($k_{\mathbf{A}} \leq r_{\mathbf{A}} := \text{rank}(\mathbf{A})$). Given $\underline{\mathbf{X}}$ ($\Leftrightarrow \mathbf{x}$), if $k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2F + 2$, then $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are unique up to a common column permutation and scaling, i.e., $\mathbf{x} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\, \mathbf{1} = (\bar{\mathbf{A}} \odot \bar{\mathbf{B}} \odot \bar{\mathbf{C}})\, \mathbf{1} \Longrightarrow \bar{\mathbf{A}} = \mathbf{A}\mathbf{\Pi}\mathbf{\Delta}_a$, $\bar{\mathbf{B}} = \mathbf{B}\mathbf{\Pi}\mathbf{\Delta}_b$, $\bar{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Delta}_c$, where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{\Delta}_a$, $\mathbf{\Delta}_b$, $\mathbf{\Delta}_c$ non-singular diagonal matrices such that $\mathbf{\Delta}_a\mathbf{\Delta}_b\mathbf{\Delta}_c = \mathbf{I}$, see [8], [14]–[16].

When dealing with big tensors $\underline{\mathbf{X}}$ that do not fit in main memory, a reasonable idea is to try to compress $\underline{\mathbf{X}}$ to a much smaller tensor that somehow captures most of the systematic variation in $\underline{\mathbf{X}}$. The commonly used compression method is to fit a low-dimensional orthogonal Tucker3 model (with low mode-ranks) [9], [10], then regress the data onto the fitted mode-bases. This idea [17], [18] has been exploited in existing PARAFAC model-fitting software, such as COMFAC [19], as a useful quick-and-dirty way to initialize alternating least squares computations in the uncompressed domain, thus accelerating convergence. Tucker3 compression requires a separate preprocessing stage that can be cumbersome for big tensors, and fitting a PARAFAC model to the compressed data only yields an approximate model for the uncompressed data. Eventually, decompression and iterations with the full data are needed to obtain fine estimates.

Memory-efficient implementation avoiding intermediate data explosion for algebraic computations with sparse tensors (including Tucker3 and PARAFAC factorization of sparse tensors) has been considered in [20], [21] for cases where the sparse tensor can fit in main memory.

III. Results

Consider compressing $\mathbf{x}$ into $\mathbf{y} = \mathbf{S}\mathbf{x}$, where $\mathbf{S}$ is $d \times IJK$, $d \ll IJK$. In particular, we propose to consider a specially structured compression matrix $\mathbf{S} = \mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T$, which corresponds to multiplying (every slab of) $\underline{\mathbf{X}}$ from the I-mode with $\mathbf{U}^T$, from the J-mode with $\mathbf{V}^T$, and from the K-mode with $\mathbf{W}^T$, where $\mathbf{U}$ is $I \times L$, $\mathbf{V}$ is $J \times M$, and $\mathbf{W}$ is $K \times N$, with $L \leq I$, $M \leq J$, $N \leq K$ and $LMN \ll IJK$; see Fig. III. Such an $\mathbf{S}$ corresponds to compressing each mode individually, which is often natural, and the associated multiplications can be efficiently implemented when the tensor is sparse. Due to a property of the Kronecker product [22],

$$\left(\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T\right)(\mathbf{A} \odot \mathbf{B} \odot \mathbf{C}) =$$

$$\left((\mathbf{U}^T\mathbf{A}) \odot (\mathbf{V}^T\mathbf{B}) \odot (\mathbf{W}^T\mathbf{C})\right),$$

from which it follows that

$$\mathbf{y} = \left((\mathbf{U}^T\mathbf{A}) \odot (\mathbf{V}^T\mathbf{B}) \odot (\mathbf{W}^T\mathbf{C})\right)\mathbf{1} = \left(\tilde{\mathbf{A}} \odot \tilde{\mathbf{B}} \odot \tilde{\mathbf{C}}\right)\mathbf{1}.$$

i.e., the compressed data follow a PARAFAC model of size $L \times M \times N$ and order F parameterized by $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$, with $\tilde{\mathbf{A}} := \mathbf{U}^T\mathbf{A}$, $\tilde{\mathbf{B}} := \mathbf{V}^T\mathbf{B}$, $\tilde{\mathbf{C}} := \mathbf{W}^T\mathbf{C}$. We have the following result.

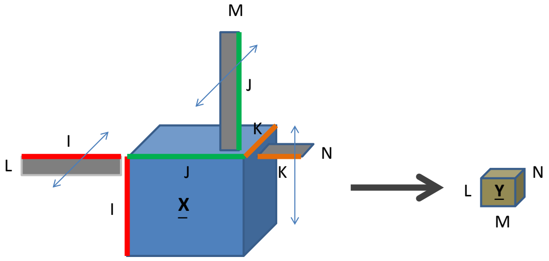

Fig. 1. Schematic illustration of tensor compression: going from an $I \times J \times K$ tensor $\underline{\mathbf{X}}$ to a much smaller $L \times M \times N$ tensor $\underline{\mathbf{Y}}$ via multiplying (every slab of) $\underline{\mathbf{X}}$ from the I-mode with $\mathbf{U}^T$, from the J-mode with $\mathbf{V}^T$, and from the K-mode with $\mathbf{W}^T$, where $\mathbf{U}$ is $I \times L$, $\mathbf{V}$ is $J \times M$, and $\mathbf{W}$ is $K \times N$.

Theorem 1: Let $\mathbf{x} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\, \mathbf{1} \in \mathbb{R}^{IJK}$, where $\mathbf{A}$ is $I \times F$, $\mathbf{B}$ is $J \times F$, $\mathbf{C}$ is $K \times F$, and consider compressing it to $\mathbf{y} = \left(\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T\right)\mathbf{x} = \left((\mathbf{U}^T\mathbf{A}) \odot (\mathbf{V}^T\mathbf{B}) \odot (\mathbf{W}^T\mathbf{C})\right)\mathbf{1} = \left(\tilde{\mathbf{A}} \odot \tilde{\mathbf{B}} \odot \tilde{\mathbf{C}}\right)\mathbf{1} \in \mathbb{R}^{LMN}$, where the mode-compression matrices $\mathbf{U}$ $(I \times L, L \leq I)$, $\mathbf{V}$ $(J \times M, M \leq J)$, and $\mathbf{W}$ $(K \times N, N \leq K)$ are randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{IL}$, $\mathbb{R}^{JM}$, and $\mathbb{R}^{KN}$, respectively. Assume that the columns of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are sparse, and let n_a (n_b, n_c) be an upper bound on the number of nonzero elements per column of $\mathbf{A}$ (respectively $\mathbf{B}$, $\mathbf{C}$). If

$$\min(L, k_{\mathbf{A}}) + \min(M, k_{\mathbf{B}}) + \min(N, k_{\mathbf{C}}) \geq 2F + 2, \quad \text{and}$$

$$L \geq 2n_a, \quad M \geq 2n_b, \quad N \geq 2n_c,$$

then the original factor loadings $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are almost surely identifiable from the compressed data $\mathbf{y}$, i.e., if $\left((\mathbf{U}^T\bar{\mathbf{A}}) \odot (\mathbf{V}^T\bar{\mathbf{B}}) \odot (\mathbf{W}^T\bar{\mathbf{C}})\right)\mathbf{1} = \mathbf{y}$, then, with probability 1, $\bar{\mathbf{A}} = \mathbf{A}\mathbf{\Pi}\mathbf{\Delta}_a$, $\bar{\mathbf{B}} = \mathbf{B}\mathbf{\Pi}\mathbf{\Delta}_b$, $\bar{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Delta}_c$, where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{\Delta}_a$, $\mathbf{\Delta}_b$, $\mathbf{\Delta}_c$ non-singular diagonal matrices such that $\mathbf{\Delta}_a\mathbf{\Delta}_b\mathbf{\Delta}_c = \mathbf{I}$.

For the proof, we will need two Lemmas. The first is the following.

Lemma 1: Consider $\tilde{\mathbf{A}} := \mathbf{U}^T\mathbf{A}$, where $\mathbf{A}$ is $I \times F$, and let the $I \times L$ matrix $\mathbf{U}$ be randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{IL}$ (e.g., multivariate Gaussian with a non-singular covariance matrix). Then $k_{\tilde{\mathbf{A}}} = \min(L, k_{\mathbf{A}})$ almost surely (with probability 1).

Proof: From Sylvester's inequality it follows that $k_{\tilde{\mathbf{A}}}$ cannot exceed $\min(L, k_{\mathbf{A}})$. Let $k := \min(L, k_{\mathbf{A}})$. It suffices to show that any k columns of $\tilde{\mathbf{A}}$ are linearly independent, for all $\mathbf{U}^T$ except for a set of measure zero. Any selection of k columns of $\tilde{\mathbf{A}}$ can be written as $\tilde{\mathbf{A}}_s = \mathbf{U}^T\mathbf{A}_s$, where $\mathbf{A}_s$ holds the respective columns of $\mathbf{A}$. Consider the square $k \times k$ top sub-matrix $\tilde{\mathbf{A}}_{s,t} = \mathbf{U}_t^T\mathbf{A}_s$, where $\mathbf{U}_t^T$ holds the top k rows of $\mathbf{U}^T$. Note that $\det(\tilde{\mathbf{A}}_{s,t})$ is an analytic function of the elements of $\mathbf{U}_t^T$ (a multivariate polynomial, in fact). An analytic function that is not zero everywhere is nonzero almost everywhere; see e.g., [23] and references therein. To prove that $\det(\tilde{\mathbf{A}}_{s,t}) \neq 0$ for almost every $\mathbf{U}_t^T$, it suffices to find one $\mathbf{U}_t^T$ for which $\det(\tilde{\mathbf{A}}_{s,t}) \neq 0$. Towards this end, note that since $k \leq k_{\mathbf{A}}$, $\mathbf{A}_s$ is full column rank, k. It therefore has a subset of k linearly independent rows. Let the corresponding k columns of $\mathbf{U}_t^T$ form a $k \times k$ identity matrix, and set the rest of the entries of $\mathbf{U}_t^T$ to zero. Then $\det(\tilde{\mathbf{A}}_{s,t}) \neq 0$ for this particular $\mathbf{U}_t^T$. This shows that the k selected columns of $\tilde{\mathbf{A}}$ (in $\tilde{\mathbf{A}}_s$) are linearly independent for all $\mathbf{U}^T$ except for a set of measure zero. There are $\binom{F}{k}$ ways to select k columns out of F, and each excludes a set of measure zero. The union of a finite number of measure zero sets has measure zero, thus all possible subsets of k columns of $\tilde{\mathbf{A}}$ are linearly independent almost surely. ■

We will also need the following Lemma, which is well-known in the compressed sensing literature [1], albeit usually not stated in Kruskal-rank terms:

Lemma 2: Consider $\tilde{\mathbf{A}} := \mathbf{U}^T\mathbf{A}$, where $\tilde{\mathbf{A}}$ and $\mathbf{U}$ are given and $\mathbf{A}$ is sought. Suppose that every column of $\mathbf{A}$ has at most n_a nonzero elements, and that $k_{\mathbf{U}^T} \geq 2n_a$. (The latter holds with probability 1 if the $I \times L$ matrix $\mathbf{U}$ is randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{IL}$, and $\min(I, L) \geq 2n_a$.) Then $\mathbf{A}$ is the unique solution with at most n_a nonzero elements per column.

Proof: Consider $\mathbf{U}^T\mathbf{a}_1 = \mathbf{U}^T\mathbf{a}_2 \Rightarrow \mathbf{U}^T(\mathbf{a}_1 - \mathbf{a}_2) = \mathbf{0}$, but $\mathbf{a}_1 - \mathbf{a}_2$ has at most $2n_a$ nonzero elements, hence the only way for this to happen is if $(\mathbf{a}_1 - \mathbf{a}_2) = \mathbf{0}$, since *any* $2n_a$ columns of $\mathbf{U}^T$ are linearly independent, by definition of k-rank. ■

We can now prove Theorem 1.

Proof: Using Lemma 1 and Kruskal's condition applied to the compressed tensor $\mathbf{y} = \left(\tilde{\mathbf{A}} \odot \tilde{\mathbf{B}} \odot \tilde{\mathbf{C}}\right)\mathbf{1}$ establishes uniqueness of $\tilde{\mathbf{A}} := \mathbf{U}^T\mathbf{A}$, $\tilde{\mathbf{B}} := \mathbf{V}^T\mathbf{B}$, $\tilde{\mathbf{C}} := \mathbf{W}^T\mathbf{C}$, up to common permutation and scaling / counter-scaling of columns, i.e., $\tilde{\mathbf{A}}\mathbf{\Pi}\mathbf{\Lambda}_a$, $\tilde{\mathbf{B}}\mathbf{\Pi}\mathbf{\Lambda}_b$, $\tilde{\mathbf{C}}\mathbf{\Pi}\mathbf{\Lambda}_c$ will be identified, where $\mathbf{\Pi}$ is a permutation matrix, and $\mathbf{\Lambda}_a$, $\mathbf{\Lambda}_b$, $\mathbf{\Lambda}_c$ are diagonal matrices such that $\mathbf{\Lambda}_a\mathbf{\Lambda}_b\mathbf{\Lambda}_c = \mathbf{I}$. Then Lemma 2 finishes the job, as it ensures that, e.g., $\mathbf{A}$ will be recovered from $\tilde{\mathbf{A}}\mathbf{\Pi}\mathbf{\Lambda}_a$ up to column permutation and scaling, and likewise for $\mathbf{B}$ and $\mathbf{C}$. ■

Remark 1: Note that the condition in Theorem 1 does not require L, M, or N to be $\geq F$; if $L \geq k_{\mathbf{A}}$, $M \geq k_{\mathbf{B}}$, $N \geq k_{\mathbf{C}}$ (which is true *a fortiori* if $L \geq \max(F, 2n_a)$, $M \geq \max(F, 2n_b)$, $N \geq \max(F, 2n_c)$), however, then Theorem 1 asserts that it is possible to identify $\mathbf{A}, \mathbf{B}, \mathbf{C}$ from the compressed data $\mathbf{y}$ under the same k-rank condition as if the uncompressed data $\mathbf{x}$ were available. If one ignores the underlying low-rank (multi-linear / Khatri-Rao) structure in $\mathbf{x}$ and attempts to recover it as a sparse but otherwise unstructured vector comprising up to $Fn_an_bn_c$ non-zero elements, then $LMN \geq 2Fn_an_bn_c$ is required. Consider a symmetric situation wherein $n_a = n_b = n_c = n$, $L = M = N$, and $F \leq 2n$. Then unstructured sparse recovery requires a total compressed sample size of $2Fn^3$, whereas the theorem calls for a total compressed sample size of $8n^3$.

Remark 2: Optimal PARAFAC fitting (and even optimal rank-one tensor approximation) is NP-hard [24]; in practice though, alternating least squares (ALS)-based fitting algorithms offer satisfactory approximation accuracy at complexity $O(IJKF)$ in raw space, and $O(LMNF)$ in compressed space (assuming that a hard limit on the number of iterations is enforced). Computing the minimum ℓ_1 norm solution of a system of under-determined equations in Q unknowns entails worst-case complexity $O(Q^{3.5})$ [25], [26]. If one ignores the underlying multi-linear structure and tries to recover the sparse $IJK \times 1$ vectorized tensor $\mathbf{x}$ directly from the compressed data $\mathbf{y}$, that has complexity $O((IJK)^{3.5})$. If one i) first fits a PARAFAC model to the compressed data at complexity $O(LMNF)$, and ii) then solves an under-determined ℓ_1 minimization subproblem for each column of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, at complexity $O(I^{3.5})$, $O(J^{3.5})$, and $O(K^{3.5})$, respectively, the overall complexity of part ii) amounts to $O((I^{3.5} + J^{3.5} + K^{3.5})F)$. Summarizing, first fitting PARAFAC in compressed space and then recovering the sparse $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ from the fitted compressed factors entails complexity $O(LMNF + (I^{3.5} + J^{3.5} + K^{3.5})F)$. Using sparsity first and then fitting PARAFAC in raw space entails complexity $O(IJKF + (IJK)^{3.5})$ - the difference is huge. Also note that the proposed approach does not require computations in the uncompressed data domain, which is important for big data that do not fit in memory for processing.

If one mode is not compressed under F, say $N \geq F$, then it is possible to guarantee identifiability with higher compression factors (smaller L, M) in the other two modes, as shown next. In what follows, we consider i.i.d. Gaussian compression matrices for simplicity.

Theorem 2: Let $\mathbf{x} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\,\mathbf{1} \in \mathbb{R}^{IJK}$, where $\mathbf{A}$ is $I \times F$, $\mathbf{B}$ is $J \times F$, $\mathbf{C}$ is $K \times F$, and consider compressing it to $\mathbf{y} = \left(\mathbf{U}^T \otimes \mathbf{V}^T \otimes \mathbf{W}^T\right)\mathbf{x} = \left((\mathbf{U}^T\mathbf{A}) \odot (\mathbf{V}^T\mathbf{B}) \odot (\mathbf{W}^T\mathbf{C})\right)\mathbf{1} = \left(\tilde{\mathbf{A}} \odot \tilde{\mathbf{B}} \odot \tilde{\mathbf{C}}\right)\mathbf{1} \in \mathbb{R}^{LMN}$, where the mode-compression matrices $\mathbf{U}$ $(I \times L, L \leq I)$, $\mathbf{V}$ $(J \times M, M \leq J)$, and $\mathbf{W}$ $(K \times N, N \leq K)$ have i.i.d. Gaussian zero mean, unit variance entries. Assume that the columns of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are sparse, and let n_a (n_b, n_c) be an upper bound on the number of nonzero elements per column of $\mathbf{A}$ (respectively $\mathbf{B}$, $\mathbf{C}$). If

$$r_{\mathbf{A}} = r_{\mathbf{B}} = r_{\mathbf{C}} = F$$

$$L(L-1)M(M-1) \geq 2F(F-1), \;\; N \geq F, \quad \text{and}$$

$$L \geq 2n_a, \quad M \geq 2n_b, \quad N \geq 2n_c,$$

then the original factor loadings $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are almost surely identifiable from the compressed data $\mathbf{y}$ up to a common column permutation and scaling.

Notice that this second theorem allows compression down to order of $\sqrt{F}$ in two out of three modes. For the proof, we will need the following Lemma:

Lemma 3: Consider $\tilde{\mathbf{A}} = \mathbf{U}^T\mathbf{A}$, where $\mathbf{A}$ $(I \times F)$ is deterministic, tall/square $(I \geq F)$ and full column rank $r_{\mathbf{A}} = F$, and the elements of $\mathbf{U}$ $(I \times L)$ are i.i.d. Gaussian zero mean, unit variance random variables. Then the distribution of $\tilde{\mathbf{A}}$ is absolutely continuous (nonsingular multivariate Gaussian) with respect to the Lebesgue measure in $\mathbb{R}^{LF}$.

Proof: Define $\mathbf{z} := \text{vec}(\tilde{\mathbf{A}}^T)$, and $\mathbf{u} := \text{vec}(\mathbf{U})$. Then $\mathbf{z} = \text{vec}(\mathbf{A}^T\mathbf{U}\mathbf{I}) = \left(\mathbf{I} \otimes \mathbf{A}^T\right)\text{vec}(\mathbf{U}) = \left(\mathbf{I} \otimes \mathbf{A}^T\right)\mathbf{u}$, and therefore $\mathbf{R}_z := E[\mathbf{z}\mathbf{z}^T] = \left(\mathbf{I} \otimes \mathbf{A}^T\right)E[\mathbf{u}\mathbf{u}^T](\mathbf{I} \otimes \mathbf{A}) = \left(\mathbf{I} \otimes \mathbf{A}^T\right)(\mathbf{I} \otimes \mathbf{A}) = \mathbf{I} \otimes \left(\mathbf{A}^T\mathbf{A}\right)$, where we have used the vectorization and mixed product rules for the Kronecker product [22]. The rank of the Kronecker product is the product of the ranks, hence $r_{\mathbf{R}_z} = LF$. ■

We can now prove Theorem 2.

Proof: From [27] (see also [15] for a deterministic counterpart), we know that PARAFAC is almost surely identifiable if the loading matrices $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$ are randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{(L+M)F}$, $\tilde{\mathbf{C}}$ is full column rank, and $L(L-1)M(M-1) \geq 2F(F-1)$. Full rank of $\tilde{\mathbf{C}}$ is ensured almost surely by Lemma 1. Lemma 3 and independence of $\mathbf{U}$ and $\mathbf{V}$ imply that the joint distribution of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^{(L+M)F}$. ■

Theorems 1 and 2 readily generalize to four-and higher-way tensors (having any number of modes). As an example, using the generalization of Kruskal's condition in [14]:

Theorem 3: Let $\mathbf{x} = (\mathbf{A}_1 \odot \cdots \odot \mathbf{A}_\delta)\,\mathbf{1} \in \mathbb{R}^{\prod_{d=1}^{\delta} I_d}$, where $\mathbf{A}_d$ is $I_d \times F$, and consider compressing it to $\mathbf{y} = \left(\mathbf{U}_1^T \otimes \cdots \otimes \mathbf{U}_\delta^T\right)\mathbf{x} = \left((\mathbf{U}_1^T\mathbf{A}_1) \odot \cdots \odot (\mathbf{U}_\delta^T\mathbf{A}_\delta)\right)\mathbf{1} = \left(\tilde{\mathbf{A}}_1 \odot \cdots \odot \tilde{\mathbf{A}}_\delta\right)\mathbf{1} \in \mathbb{R}^{\prod_{d=1}^{\delta} L_d}$, where the mode-compression matrices $\mathbf{U}_d$ ($I_d \times L_d, L_d \leq I_d$) are randomly drawn from an absolutely continuous distribution with respect to the Lebesgue measure in $\mathbb{R}^{I_d L_d}$. Assume that the columns of $\mathbf{A}_d$ are sparse, and let n_d be an upper bound on the number of nonzero elements per column of $\mathbf{A}_d$, for each $d \in \{1, \cdots, \delta\}$. If

$$\sum_{d=1}^{\delta} \min(L_d, k_{\mathbf{A}_d}) \geq 2F+\delta-1, \quad \text{and} \quad L_d \geq 2n_d, \quad \forall d \in \{1, \cdots, \delta\},$$

then the original factor loadings $\{\mathbf{A}_d\}_{d=1}^{\delta}$ are almost surely identifiable from the compressed data $\mathbf{y}$ up to a common column permutation and scaling.

IV. Discussion

Identifiability guarantees are of course nice, at the end of the day, however, practitioners are interested in actually computing the underlying loading matrices $\{\mathbf{A}_d\}_{d=1}^{\delta}$. Our results naturally suggest a two-step recovery process: fitting a PARAFAC model to the compressed data using any of the available algorithms, such as [19] or those in [9]; then recovering each $\mathbf{A}_d$ from the recovered $\tilde{\mathbf{A}}_d = \mathbf{U}_d^T\mathbf{A}_d$ using any estimation algorithm from the compressed sensing literature. We have written code to corroborate our identifiability claims, using [19] for the first step and enumeration-based (ℓ_0) de-compression for the second step. This code is made available as proof-of-concept, and will be posted at www.ece.umn.edu/~nikos. Recall that optimal PARAFAC fitting is NP-hard, hence any computational procedure cannot be fail-safe, but in our tests the results were consistent. Also note that, while identifiability considerations and ℓ_0 recovery only demand that $L_d \geq 2n_d$, ℓ_1-based recovery algorithms typically need $L_d \geq (3 \div 5)n_d$ to produce acceptable results. In the same vain, while PARAFAC identifiability only requires $\sum_{d=1}^{\delta} \min(L_d, k_{\mathbf{A}_d}) \geq 2F+\delta-1$, good estimation performance often calls for higher L_d's, which however can still afford very significant compression ratios.

Tomioka *et al* [28] considered low mode-rank tensor recovery from compressed measurements, and derived approximation error bounds without requiring sparsity. Instead, we focused on *exact* recovery of the latent loadings / rank-one factors for the uncompressed tensor from the compressed measurements, assuming low tensor rank (note tensor rank $\neq$ mode-ranks) and latent sparsity. Tomioka's approach could be used to reconstruct the full tensor from the compressed one, and then apply CP decomposition to the full tensor. This however would give up the memory / storage / complexity benefits of our approach.

References

[1] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.

[2] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[3] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[4] E. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[5] J. Carroll and J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[6] R. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[7] ——, "Determination and proof of minimum uniqueness conditions for PARAFAC-1," *UCLA Working Papers in Phonetics*, vol. 22, pp. 111–117, 1972.

[8] J. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra and its Applications*, vol. 18, no. 2, pp. 95–138, 1977.

[9] A. Smilde, R. Bro, P. Geladi, and J. Wiley, *Multi-way analysis with applications in the chemical sciences*. Wiley, 2004.

[10] P. Kroonenberg, *Applied multiway data analysis*. Wiley, 2008.

[11] E. Papalexakis and N. Sidiropoulos, "Co-clustering as multilinear decomposition with sparse latent factors," in *IEEE ICASSP 2011, May 22-27, Prague, Czech Republic*, pp. 2064–2067.

[12] L.-H. Lim and P. Comon, "Multiarray signal processing: Tensor decomposition meets compressed sensing," *Comptes Rendus Mécanique*, vol. 338, no. 6, pp. 311–320, 2010.

[13] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A Multilinear Singular Value Decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[14] N. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," *Journal of chemometrics*, vol. 14, no. 3, pp. 229–239, 2000.

[15] T. Jiang and N. Sidiropoulos, "Kruskal's permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear models with constant modulus constraints," *IEEE Transactions on Signal Processing*, vol. 52, no. 9, pp. 2625–2636, 2004.

[16] A. Stegeman and N. Sidiropoulos, "On Kruskal's uniqueness condition for the CANDECOMP/PARAFAC decomposition," *Linear Algebra and its Applications*, vol. 420, no. 2-3, pp. 540–552, 2007.

[17] R. Bro and C. Anderson, "Improving the speed of multiway algorithms Part II: Compression," *Chemometrics and Intelligent Laboratory Systems*, vol. 42, pp. 105–113, 1998.

[18] J. Carroll, S. Pruzansky, and J. Kruskal, "CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters," *Psychometrika*, vol. 45, no. 1, pp. 3–24, 1980.

[19] R. Bro, N. Sidiropoulos, and G. Giannakis, "A fast least squares algorithm for separating trilinear mixtures," in *Proc. ICA99 Int. Workshop on Independent Component Analysis and Blind Signal Separation*, 1999, pp. 289–294. [Online]. Available: http://www.ece.umn.edu/~nikos/comfac.m

[20] B. Bader and T. Kolda, "Efficient MATLAB computations with sparse and factored tensors," *SIAM Journal on Scientific Computing*, vol. 30, no. 1, pp. 205–231, 2007.

[21] T. Kolda and J. Sun, "Scalable tensor decompositions for multi-aspect data mining," in *ICDM 2008: Proceedings of the 8th IEEE International Conference on Data Mining*, pp. 363–372.

[22] J. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. on Circuits and Systems*, vol. 25, no. 9, pp. 772–781, 1978.

[23] T. Jiang, N. Sidiropoulos, and J. ten Berge, "Almost sure identifiability of multidimensional harmonic retrieval," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1849–1859, 2001.

[24] C. Hillar and L.-H. Lim, "Most Tensor Problems are NP-hard," 2009. [Online]. Available: http://arxiv.org/abs/0911.1393

[25] E. Candès and J. Romberg, "ℓ_1-Magic: Recovery of Sparse Signals via Convex Programming," 2005. [Online]. Available: http://www-stat.stanford.edu/~candes/l1magic/downloads/l1magic.pdf

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[27] A. Stegeman, J. ten Berge, and L. De Lathauwer, "Sufficient conditions for uniqueness in CANDECOMP/PARAFAC and INDSCAL with random component matrices," *Psychometrika*, vol. 71, no. 2, pp. 219–229, 2006.

[28] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima, "Statistical performance of convex tensor decomposition," in *J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, eds, Advances in Neural Information Processing Systems 24*, 2011, pp. 972–980.