Joint Tensor Factorization and Outlying Slab Suppression With Applications

Xiao Fu, *Member, IEEE*, Kejun Huang, *Student Member, IEEE*, Wing-Kin Ma, *Senior Member, IEEE*, Nicholas D. Sidiropoulos, *Fellow, IEEE*, and Rasmus Bro

***Abstract*—We consider factoring low-rank tensors in the presence of outlying slabs. This problem is important in practice, because data collected in many real-world applications, such as speech, fluorescence, and some social network data, fit this paradigm. Prior work tackles this problem by iteratively selecting a fixed number of slabs and fitting, a procedure which may not converge. We formulate this problem from a group-sparsity promoting point of view, and propose an alternating optimization framework to handle the corresponding ℓ_p $(0 < p \leq 1)$ minimization-based low-rank tensor factorization problem. The proposed algorithm features a similar per-iteration complexity as the plain trilinear alternating least squares (TALS) algorithm. Convergence of the proposed algorithm is also easy to analyze under the framework of alternating optimization and its variants. In addition, regularization and constraints can be easily incorporated to make use of *a priori* information on the latent loading factors. Simulations and real data experiments on blind speech separation, fluorescence data analysis, and social network mining are used to showcase the effectiveness of the proposed algorithm.**

***Index Terms*—Canonical polyadic decomposition, group sparsity, iteratively reweighted, outliers, PARAFAC, robustness, tensor decomposition.**

I. Introduction

FACTORING a tensor (i.e., a data set indexed by three or more indices) into rank-one components is a decomposition problem which is frequently referred to as *parallel factor analysis* (PARAFAC) or *canonical decomposition* (CANDECOMP), or *canonical polyadic decomposition* (CPD). Unlike two-way factor analysis (i.e., matrix factorizations), three- or higher-way low-rank tensor factorization reveals essentially unique factors under quite mild conditions, which is desirable when dealing with latent parameter estimation problems. Since the late 1990s, PARAFAC has been successfully applied to wireless communications for blindly estimating the spatial channels or the users' code-division signatures [1], [2]; array processing for finding the directions-of-arrival of the emitters [3], [4]; chemometrics for resolving the spectra of chemical analytes [5]; blind speech and audio separation for estimating the mixing system [6], [7]; and, more recently, power spectra separation for cognitive radio [8], and big data mining for social group clustering [9].

A high-order tensor can also be considered as a set of lower-order tensors. For example, a data cube (i.e., a three-way tensor) can be considered as a set of matrices (two-way tensors), obtained by fixing one index to a particular value. Each such piece of the original data, whose order has been reduced by one, will be called a *slab*. Slabs are usually physically meaningful in various applications. For example, in blind speech and audio separation, the received signals' short-term covariance matrix, assumed constant within a short coherence interval and sometimes referred to as *local covariance* [10], [11], can be considered as a slab of a three-way tensor; in fluorescence data spectroscopy, a measurement matrix that consists of emissions and excitations of the stimulated analytes is a slab [5]; and in array processing, the received raw signals at a subarray can be considered as a slab [3]. Due to this physical correspondence, however, strong data contamination or corruption frequently happens at the slab level (rather than element-wise). A typical example is blind speech separation—it has been observed that locally correlated speech sources may create local covariances (slabs) that do not obey the low-rank tensor model [11]. Also, in chemometrics, e.g. in fluorescence spectroscopy, it is common that certain samples representing erratic measurements or samples of unusual constitution end up influencing the fitted model badly [12]–[14].

Factoring a low-rank tensor in the presence of outlying slabs has been considered before. In the literature, the most closely related work may be [12]. There, an algorithm that iteratively selects a fixed number of slabs to fit with a low-rank tensor model was proposed. A main drawback with this algorithm is that it may not converge. Also, it is not easy to determine how many slabs should be selected to fit in advance. Similar insights are also seen in the analytic chemistry context; see [13]–[15]. In [16], the authors considered a different yet related scenario. There, a PARAFAC approach was proposed by changing the least squares-based optimization criterion to the ℓ_1-norm based fitting criterion, to make the low-rank decomposition robust against outlying elements. The resulting algorithms are *alternating linear programming* or *alternating weighted median filtering* (WMF). The algorithms in [16] do not need to pre-define the number of slabs to select for fitting, but they can be inefficient even when the problem size is medium. In addition, the ℓ_1 criterion is optimal in the maximum likelihood sense, when the

Manuscript received April 04, 2015; revised July 15, 2015; accepted July 27, 2015. Date of publication August 18, 2015; date of current version October 29, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Andre Almeida. The work of X. Fu, K. Huang, and N. D. Sidiropoulos was supported in part by NSF IIS-1247632.

X. Fu, K. Huang, and N. D. Sidiropoulos are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: xfu@umn.edu; huang663@umn.edu; nikos@umn.edu).

W.-K. Ma is with the Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong (e-mail: wkma@ieee.org).

R. Bro is with the Department of Food Science, Faculty of Science, University of Copenhagen, DK-1958 Frederiksberg, Denmark (e-mail: rb@life.ku.dk).

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

Digital Object Identifier 10.1109/TSP.2015.2469642

noise follows the i.i.d. Laplacian distribution; but it is not specialized for (strong) slab-level outliers, as will be shown in the simulations.

Contributions: In this work, we consider modeling and tackling the low-rank tensor decomposition problem with outlying slabs from a different perspective. Specifically, we formulate the problem from a group-sparsity promoting viewpoint, and come up with an ℓ_p $(0 < p \leq 1)$ fitting criterion. We propose to tackle this hard optimization problem using an alternating optimization strategy: by judiciously recasting the original problem into a more convenient form, we show that it can be tackled using a simple algorithm whose block updates admit closed-form solutions. This algorithm tends to iteratively select some clean slabs to fit with a PARAFAC model and downweight the outlying slabs at the same time. Reminiscent of classical robust fitting, the proposed algorithm does not assume knowledge of the number of clean slabs. Plus, drawing from existing theoretical results on alternating optimization [17] and its variants such as *maximum block improvement* (MBI) [18], convergence of the proposed algorithm can be characterized. It is also worth noting that the proposed algorithm has almost the same per-iteration complexity as the *trilinear alternating least squares* (TALS) algorithm [1]–[3], which is computationally much cheaper than the algorithms in [16].

Extensions to regularized and constrained cases are also considered in this work, since incorporating *a priori* information on the loading factors is important in applied data analysis. Following the same alternating optimization framework, we propose to handle the subproblems by employing an *alternating direction method of multipliers* [19] (ADMM)-based algorithm, which allows us to deal with different types of regularization and constraints of interest, under a unified update strategy.

Besides simulations using synthetic data for verifying the ideas, we use several simulations and experiments with real data to showcase the effectiveness of the proposed approaches. First, the basic robust algorithm is applied on blind speech separation simulations, where real speech segments are mixed under realistic room acoustic impulse response scenarios. The separation performance of the proposed algorithm is shown to be superior to the earlier state-of-the-art. Then, the proposed algorithm is applied to a fluorescence data set; and finally to the ENRON e-mail corpus. Interesting and nicely interpretable results are obtained in both cases.

Notation: We largely follow standard signal processing (and some Matlab) notational conventions, for convenience. Specifically, $\underline{\mathbf{X}} \in \mathbb{R}^{I\times J\times K}$ denotes a three-way tensor, and $\underline{\mathbf{X}}(i,j,k)$ denotes the element that is indexed by (i,j,k); $\underline{\mathbf{X}}(i,:,:)$, $\underline{\mathbf{X}}(:,j,:)$ and $\underline{\mathbf{X}}(:,:,k)$ denote the ith horizontal slab, the jth lateral slab, and the kth frontal slab, respectively; $\mathbf{X}(i,:)$ and $\mathbf{X}(:,j)$ denotes the ith row and the jth column of the matrix $\mathbf{X}$; ^{T} denotes the transpose operator; † denotes the Moore-Penrose pseudo-inverse operator; $\|\mathbf{x}\|_p = \left(\sum_{i=1}^m |x_i|^p\right)^{1/p}$ for $\mathbf{x}\in\mathbb{R}^m$ for $0 < p < \infty$; $\mathrm{vec}(\mathbf{X}) = [\mathbf{X}^T(:,1),\ldots,\mathbf{X}^T(:,J)]^T$ for $\mathbf{X}\in\mathbb{R}^{I\times J}$; $k_{\mathbf{X}}$ and $\mathrm{rank}(\mathbf{X})$ denote the Kruskal rank and the rank of $\mathbf{X}$, respectively; $\circ$, $\circledast$, $\otimes$ and $\odot$ denote the outer product, the Hadmard product, the Kronecker product, and the Khatri-Rao product, respectively; $\mathrm{Diag}(\mathbf{x})$ denotes a diagonal matrix that holds the $x_1,\ldots,x_m$ as the diagonal elements.

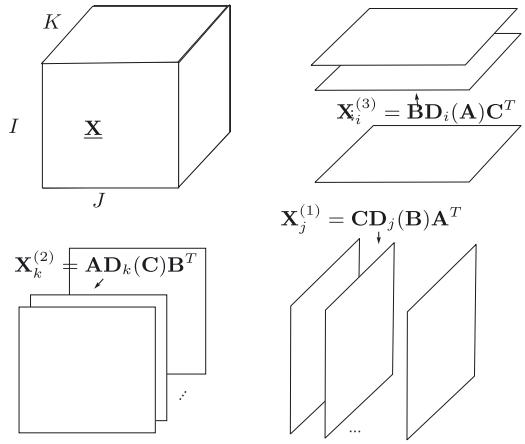


Fig. 1. Slabs of a three-way tensor.

II. Preliminaries on PARAFAC

A simple description of the PARAFAC model is as follows. PARAFAC aims to represent a three-way tensor $\underline{\mathbf{X}} \in \mathbb{C}^{I\times J\times K}$ using PARAFAC three latent factor matrices:

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \mathbf{A}(:,r)\circ\mathbf{B}(:,r)\circ\mathbf{C}(:,r), \tag{1}$$

where $\mathbf{A}\in\mathbb{C}^{I\times R}$, $\mathbf{B}\in\mathbb{C}^{J\times R}$, $\mathbf{C}\in\mathbb{C}^{K\times R}$, and R is called the rank of the PARAFAC model. Any tensor $\underline{\mathbf{X}}\in\mathbb{C}^{I\times J\times K}$ can be exactly represented this way if a large-enough $R \leq \min(IJ, JK, IK)$ is used; but we are usually interested in using relatively small R to capture the 'principal components' of $\underline{\mathbf{X}}$. Equivalently, each element of the tensor can be represented as $\underline{\mathbf{X}}(i,j,k) \approx \sum_{r=1}^R \mathbf{A}(i,r)\mathbf{B}(j,r)\mathbf{C}(k,r)$. A three-way tensor is also a set of matrices, or, slabs, which are obtained by fixing one index. There are three types of slabs of a three-way tensor, namely, the horizontal slabs $(\{\underline{\mathbf{X}}(i,:,:)\}_{i=1}^I)$, the lateral slabs $(\{\underline{\mathbf{X}}(:,j,:)\}_{j=1}^J)$, and the frontal slabs $(\{\underline{\mathbf{X}}(:,:,k)\}_{k=1}^K)$. If the PARAFAC model in (1) holds exactly, each type of slab has a compact representation, i.e.,

$$\begin{aligned}
\text{Lateral slabs} \quad & \left\{\mathbf{X}_j^{(1)} = \underline{\mathbf{X}}(:,j,:) = \mathbf{C}\mathbf{D}_j(\mathbf{B})\mathbf{A}^T\right\}_{j=1}^J,\\
\text{Frontal slabs} \quad & \left\{\mathbf{X}_k^{(2)} = \underline{\mathbf{X}}(:,:,k) = \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{B}^T\right\}_{k=1}^K,\\
\text{Horizontal slabs} \quad & \left\{\mathbf{X}_i^{(3)} = \underline{\mathbf{X}}(i,:,:) = \mathbf{B}\mathbf{D}_i(\mathbf{A})\mathbf{C}^T\right\}_{i=1}^I,
\end{aligned}$$

where $\mathbf{D}_r(\mathbf{X}) = \mathrm{Diag}(\mathbf{X}(r,:))$. Fig. 1 gives a visual illustration of a three-way tensor $\underline{\mathbf{X}}$ and its slabs.

Unlike matrix factorizations, which are in general non-unique, the PARAFAC decomposition has (essentially) unique solution under quite mild conditions. For example, Kruskal proved the following result for a real-valued low-rank tensor [20]: If

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2R + 2, \tag{2}$$

then $(\mathbf{A},\mathbf{B},\mathbf{C})$ are unique up to a common column permutation and scaling, i.e., $\underline{\mathbf{X}} = \sum_{r=1}^R \mathbf{A}(:,r)\circ\mathbf{B}(:,r)\circ\mathbf{C}(:,r) = \sum_{r=1}^R \bar{\mathbf{A}}(:,r)\circ\bar{\mathbf{B}}(:,r)\circ\bar{\mathbf{C}}(:,r) \Rightarrow \bar{\mathbf{A}} = \mathbf{A}\mathbf{\Pi}\mathbf{\Delta}_a$, $\bar{\mathbf{B}} = \mathbf{B}\mathbf{\Pi}\mathbf{\Delta}_b$, $\bar{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Delta}_c$, where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{\Delta}_a$, $\mathbf{\Delta}_b$, $\mathbf{\Delta}_c$, are full-rank diagonal matrices such that $\mathbf{\Delta}_a\mathbf{\Delta}_b\mathbf{\Delta}_c = \mathbf{I}$. If $\mathbf{A}$ is drawn from an absolutely continuous distribution over $\mathbb{R}^{I\times R}$,

then $k_{\mathbf{A}} = \text{rank}(\mathbf{A}) = \min(I, R)$ with probability one. It follows that if $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are drawn this way, then the condition in (2) can be simplified: if

$$\min\{I, R\} + \min\{J, R\} + \min\{K, R\} \geq 2R + 2, \qquad (3)$$

then $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are unique up to a common column permutation and scaling, with probability one. Notice that under the condition in (2) or (3), the loading matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ need not be tall. This is advantageous in challenging application scenarios, e.g., mixing system identification when the system is under-determined [2], [10].

In practice, when modeling error and noise exist, it makes more sense to seek the best rank-R approximation of a tensor rather than computing its exact rank factorization. To find such an approximation, the least squares criterion is commonly adopted:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \underline{\mathbf{X}} - \sum_{r=1}^{R} \mathbf{A}(:,r) \circ \mathbf{B}(:,r) \circ \mathbf{C}(:,r) \right\|_F^2. \qquad (4)$$

The above problem is nonconvex, and thus could be very difficult to solve. In fact, recent research [21] showed that Problem (4) may even be 'ill-posed', meaning that the best rank-R approximation of a tensor may not even exist. In practice, nevertheless, the formulation in (4) allows one to devise computationally affordable (albeit generally suboptimal) algorithms, and some of these algorithms have proven successful in various applications. To deal with the optimization problem in (4), a popular way is to make use of the *matrix unfoldings* of the tensor. Specifically, by vectorizing each type of slabs and treating them as columns of a matrix, we obtain the three matrix unfoldings, namely, $\underline{\mathbf{X}}^{(1)} = (\mathbf{A} \odot \mathbf{C})\mathbf{B}^T$, $\underline{\mathbf{X}}^{(2)} = (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T$, and $\underline{\mathbf{X}}^{(3)} = (\mathbf{C} \odot \mathbf{B})\mathbf{A}^T$, where we have used the vectorization property of the Khatri-Rao product $\text{vec}(\mathbf{X}\text{Diag}(\mathbf{z})\mathbf{Y}^T) = (\mathbf{Y} \odot \mathbf{X})\text{vec}(\mathbf{z})$. Using the unfoldings, Problem (4) can be tackled by cyclically solving the following three least squares problems:

$$\mathbf{B} := \arg\min_{\mathbf{B}} \left\| \underline{\mathbf{X}}^{(1)} - (\mathbf{A} \odot \mathbf{C})\mathbf{B}^T \right\|_F^2 \qquad (5a)$$

$$\mathbf{C} := \arg\min_{\mathbf{C}} \left\| \underline{\mathbf{X}}^{(2)} - (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T \right\|_F^2 \qquad (5b)$$

$$\mathbf{A} := \arg\min_{\mathbf{A}} \left\| \underline{\mathbf{X}}^{(3)} - (\mathbf{C} \odot \mathbf{B})\mathbf{A}^T \right\|_F^2. \qquad (5c)$$

The above updates yield the popular *trilinear alternating least squares* (TALS) algorithm [1], [2].

Although quite a lot of different PARAFAC algorithms exist, e.g., [10], [22]–[25], TALS (and its close relatives) has been the workhorse of low-rank tensor decomposition for decades for several reasons: First, TALS can be easily implemented, since each iteration only involves relatively simple linear least squares subproblems. Second, it features monotone convergence of the cost function, without the need to tune (e.g., step-size) parameters to ensure this. Third, it has the flexibility to incorporate constraints and regularization on the loading factors under its alternating optimization framework, with a reasonable complexity increase.

III. A Closer Look at Motivating Examples

In many applications, some slabs of the collected tensor data are highly corrupted, for various reasons. In this section, we take a closer look at some pertinent examples that we have encountered in rather different fields. In all of them, corrupted slabs can throw off the analysis, producing inconsistent and hard to interpret PARAFAC models.

A. Blind Speech Separation

It has been shown that PARAFAC can be applied to blind speech separation (BSS) to identify the mixing system [6], [7]. As a quick review, the BSS signal model is

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t),\ t = 1, 2, \ldots \qquad (6)$$

where $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T \in \mathbb{R}^I$ denotes the received signals by the I sensors at time t, $\mathbf{A} \in \mathbb{R}^{I \times R}$ denotes the mixing system, $\mathbf{s}(t) = [s_1(t), \ldots, s_R(t)]^T \in \mathbb{R}^R$ denotes the R speech sources (presumed to be uncorrelated), and $\mathbf{n}(t) = [n_1(t), \ldots, n_I(t)]^T \in \mathbb{R}^I$ denotes zero-mean i.i.d. Gaussian noise with variance σ^2. To connect this model to the PARAFAC model, we calculate the local covariance of the received signals within time frame k by

$$\begin{aligned} \underline{\mathbf{X}}(:,:,k) &= \mathbb{E}\left\{\mathbf{x}(t)\mathbf{x}^T(t)\right\} - \hat{\sigma}^2\mathbf{I} \\ &\approx \mathbf{A}\mathbb{E}\left\{\mathbf{s}(t)\mathbf{s}^T(t)\right\}\mathbf{A}^T, \quad t \in [(k-1)L + 1, kL], \end{aligned}$$

where $\hat{\sigma}^2$ represents the estimated noise variance and L denotes the time frame length. By assuming that the sources are uncorrelated, we see that the local covariance of the sources in frame k, i.e., for $t \in [(k-1)L + 1, kL]$,

$$\mathbb{E}\left\{\mathbf{s}(t)\mathbf{s}^T(t)\right\} = \text{Diag}\left(\left[\mathbb{E}\left|s_1(t)\right|^2, \ldots, \mathbb{E}\left|s_R(t)\right|^2\right]\right),$$

is a diagonal matrix. Hence, if we let $\mathbf{C}(k,:) = [\mathbb{E}|s_1(t)|^2, \ldots, \mathbb{E}|s_R(t)|^2]$ for $t \in [(k-1)L + 1, kL]$, we see that $\underline{\mathbf{X}}(:,:,k) = \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{A}^T$ is a frontal slab of a three-way tensor (with $\mathbf{B} = \mathbf{A}$), and thus PARAFAC can be applied to $\underline{\mathbf{X}}$ to estimate the mixing system $\mathbf{A}$. Using the estimated $\hat{\mathbf{A}}$, the individual source signals can be estimated. In the presence of reverberation, the mixing system model becomes convolutive (i.e., frequency-selective) instead of instantaneous. This is a more challenging scenario, which can again be tackled using PARAFAC in the frequency domain, see [6], [7], [11] and references therein.

A more subtle difficulty is that some speech sources exhibit (strong) short-term cross correlations, even though they are approximately uncorrelated over the long run. Consequently, the local covariances of the sources in some frames have significant off-diagonal elements, and the corresponding slabs deviate from the nominal model $\underline{\mathbf{X}}(:,:,k) = \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{A}^T$. In such cases, directly applying standard PARAFAC algorithms may not yield satisfactory speech separation performance [11].

B. Fluorescence Spectroscopy

Fluorescence excitation-emission measurements (EEMs) are used in many different fields such as skin analysis, fermentation monitoring, environmental, food, and clinical analysis [14]. A fluorescence sample is obtained by using a beam of light that excites the electrons in molecules of certain compounds and causes them to emit light; the emission spectra are then measured at several excitation wavelengths. A fluorescence EEM sample can be represented by

$$\underline{\mathbf{X}}(i,:,:) = \mathbf{B}\mathbf{D}_i(\mathbf{A})\mathbf{C}^T,$$

where $\mathbf{B}(:,r)$ for $r = 1,\ldots,R$ corresponds to the spectral emission r, $\mathbf{C}(:,r)$ denotes the corresponding excitation values, and $\mathbf{A}(i,r)$ denotes the corresponding concentration (scaling) at sample i. By measuring multiple samples, a PARAFAC model can be formed, and each sample is a slab.

Fluorescence data analysis has been recognized as a very successful example of applying PARAFAC algorithms to real-world data. At the same time, it has also been noticed that anomalous EEM samples occur frequently due to various reasons [12]–[15].

C. Social Network Mining

For some three-way social network data sets, every slab $\underline{\mathbf{X}}(:,:,k)$ is a connected graph measured within time period k. For example, in the ENRON e-mail data set [26], $\underline{\mathbf{X}}(i,j,k)$ denotes the 'connection intensity' of person i and person j at time period k (i.e., the number of e-mails sent by person i to person j within month k). Another example is the Amazon purchase data. There, $\underline{\mathbf{X}}(i,j,k)$ represents the amount of product j bought by person i in week k. For such data, each rank-one component of the PARAFAC model can be interpreted as the interaction pattern of a social group over time [27]. To be specific, consider

$$\underline{\mathbf{X}}(:,:,k) \approx \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{B}^T = \sum_{r=1}^{R} \mathbf{C}(k,r)\mathbf{A}(:,r)(\mathbf{B}(:,r))^T.$$

Here, the nonzero elements in $\mathbf{A}(:,r)$ and $\mathbf{B}(:,r)$ create a clique (a subgraph) $\mathbf{A}(:,r)\mathbf{B}^T(:,r)$, which can be interpreted as a social group, and R corresponds to the number of social groups. Taking the ENRON e-mail data as an example, $\mathbf{A}(:,r)\mathbf{B}^T(:,r)$ is a group, where the people corresponding to the non-zero elements of $\mathbf{A}(:,r)$ have similar e-mail sending patterns to those corresponding to the non-zero elements of $\mathbf{B}(:,r)$. $\mathbf{C}(k,r)$ is a time-varying parameter of this group, which means that the e-mail sending pattern of this group is a rank-one matrix factor whose intensity (e-mail volume) varies with time.

With this model, factoring the data box into its latent factors is equivalent to mining the underlying social groups, which finds applications in designing recommendation systems, analyzing ethic and cultural groups, and even detecting criminal organizations. However, the social network data sets are in general not following a generative signal model, which means that several slabs may have large modeling errors. As we will see later, some unexpected events (such as the ENRON crisis) might make the group e-mail patterns quite irregular during some period. The slabs measured in these irregular time intervals might need to be identified and somehow down-weighted when the objective is to analyze the normal interaction patterns, or to detect those anomalies.

IV. Problem Formulation

Motivated by the examples in the previous section, we will focus on modeling, formulating, and solving the low-rank tensor decomposition problem in the presence of outlying slabs. Our main goal is an easily implemented optimization framework; practical considerations such as regularization, constraints, initialization and complexity will also be discussed. For presentation simplicity, we will assume that corruption happens in some horizontal slabs throughout the development of the algorithm; see Fig. 2. Algorithms dealing with corrupt lateral or frontal slabs can be obtained by simply permuting the modes of the tensor, by virtue of symmetry.

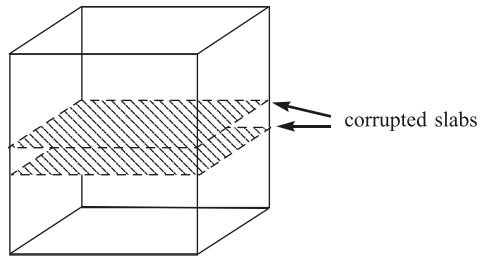


Fig. 2. The corruption model: some horizontal slabs are outliers.

To begin with, let us assume that some horizontal slabs have been corrupted by gross errors; i.e., we have

$$\mathbf{X}_i^{(3)} = \begin{cases} \mathbf{B}\mathbf{D}_i(\mathbf{A})\mathbf{C}^T + \mathbf{O}_i, & i \in \mathcal{N}, \\ \mathbf{B}\mathbf{D}_i(\mathbf{A})\mathbf{C}^T, & i \in \mathcal{N}_c, \end{cases} \tag{7}$$

where $\mathcal{N} \subset \{1,\ldots,I\}$ is the index set of the *outlying slabs* and $\mathcal{N}_c = \{1,\ldots,I\} - \mathcal{N}$. The gross error component $\mathbf{O}_i$ could be strong so that $\mathbf{X}_i^{(3)}$ is far from the nominal 'clean signal model', i.e., $\mathbf{X}_i^{(3)} = \mathbf{B}\mathbf{D}_i(\mathbf{A})\mathbf{C}^T$. Under the corruption model in (7), our first observation here is that there may still be enough clean data to enable us to recover $\mathbf{B}$ and $\mathbf{C}$ intact. Thus, our idea begins with a formulation that guarantees the identifiability of $\mathbf{B}$ and $\mathbf{C}$ under some conditions.

We wish to fit the clean data slabs with a PARAFAC model. In practice, $\mathcal{N}$ is usually unknown, but its cardinality may be small relative to I. Hence, we address this problem from a group-sparsity promoting viewpoint. We formulate the problem as

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{i=1}^{I} \mathcal{I}\left(\left\| \underline{\mathbf{X}}^{(3)}(:,i) - (\mathbf{C}\odot\mathbf{B})\left(\mathbf{A}(i,:)\right)^T \right\|_2 \right), \tag{8}$$

where $\mathcal{I}(x)$ is defined as

$$\mathcal{I}(x) = \begin{cases} 1, & x \neq 0 \\ 0, & x = 0. \end{cases}.$$

The criterion tends to make $\underline{\mathbf{X}}^{(3)}(:,i) - (\mathbf{C}\odot\mathbf{B})(\mathbf{A}(i,:))^T = \mathbf{0}$ for as many i's as possible. Intuitively, if there are enough clean slabs to identify the underlying nominal PARAFAC model, solving the above optimization problem should identify $\mathbf{B}$ and $\mathbf{C}$. The following result confirms this intuition.

Claim 1: Assume that the elements of $\mathbf{A}$ are drawn from an absolutely continuous distribution over $\mathbb{R}^{I\times R}$, and likewise $\mathbf{B}$ and $\mathbf{C}$ are drawn from absolutely continuous distributions over $\mathbb{R}^{J\times R}$ and $\mathbb{R}^{K\times R}$, respectively. Define

$$c := 2R + 2 - \min\{J,R\} - \min\{K,R\},$$

and suppose that $c \leq \min\{|\mathcal{N}_c|, R\}$, and

$$|\mathcal{N}_c| \geq \frac{I+c}{2}. \tag{9}$$

Then, with probability one, the optimal $\mathbf{B}^\star$, $\mathbf{C}^\star$, and $\mathbf{A}^\star(\mathcal{N}_c,:)$ that solve Problem (8) are $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{A}(\mathcal{N}_c,:)$ with a common column permutation and scaling; i.e., $\mathbf{A}^\star(\mathcal{N}_c,:) = \mathbf{A}(\mathcal{N}_c,:)\mathbf{\Pi}\mathbf{\Delta}_a$, $\mathbf{B}^\star = \mathbf{B}\mathbf{\Pi}\mathbf{\Delta}_b$, $\mathbf{C}^\star = \mathbf{C}\mathbf{\Pi}\mathbf{\Delta}_c$, where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{\Delta}_a$, $\mathbf{\Delta}_b$, $\mathbf{\Delta}_c$, are full-rank diagonal matrices such that $\mathbf{\Delta}_a\mathbf{\Delta}_b\mathbf{\Delta}_c = \mathbf{I}$.

The proof of Claim 1 can be found in Appendix A. Claim 1 helps us understand the fundamental limitation of the proposed criterion in (8): Under the signal model in (7), if about one half of the horizontal slabs follow the clean signal model, we can still correctly identify the two loading factors $\mathbf{B}$ and $\mathbf{C}$ (and at least part of $\mathbf{A}$). Solving Problem (8) is very challenging though—both PARAFAC decomposition and group-sparsity maximization (cardinality minimization) are nonconvex problems on their own, so (8) is compounding two already challenging problems. In the next section, a more practical optimization surrogate will be employed to approximate Problem (8), and a simple alternating optimization algorithm will be presented to tackle this surrogate optimization problem.

V. Basic Algorithmic Framework

To approximate Problem (8), we propose to employ the smoothed ℓ_p quasi-norm as our working objective; i.e., by replacing $\sum_{i=1}^{I}\mathcal{I}(x_i)$ by $\sum_{i=1}^{I}(x_i^2+\epsilon)^{p/2}$, we deal with the following surrogate:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{i=1}^{I}\left(\left\|\underline{\mathbf{X}}^{(3)}(:,i)-(\mathbf{C}\odot\mathbf{B})\left(\mathbf{A}(i,:)\right)^T\right\|_2^2+\epsilon\right)^{p/2}, \tag{10}$$

where $0<p\leq 1$ and $\epsilon>0$. The idea comes from compressive sensing, where the quasi ℓ_0 norm is often approximated by the quasi ℓ_p norm or ℓ_1 norm, since the latter two are computationally tractable and often yield practically good results [28]–[31]. Here, ϵ is a small smoothing parameter to keep the cost function in its continuously differentiable domain.

The cost function in (10) can be manipulated according to the following lemma:

Lemma 1: Assume $0<p<2$, $\epsilon\geq 0$, and $\phi_p(w):=\frac{2-p}{2}\left(\frac{2}{p}w\right)^{\frac{p}{p-2}}+\epsilon w$. Then, we have

$$(x^2+\epsilon)^{p/2}=\min_{w\geq 0}\ wx^2+\phi_p(w),$$

and the unique minimizer is

$$w_{\rm opt}=\frac{p}{2}(x^2+\epsilon)^{\frac{p-2}{2}}. \tag{11}$$

Proof: First, it can be seen that $\phi_p(w)$ is strictly convex on its domain (i.e., the interior of $w\geq 0$), since its second order derivative is positive when w is positive, i.e.,

$$\nabla^2\phi_p(w)=-\frac{4}{p(p-2)}\left(\frac{2}{p}w\right)^{\frac{4-p}{p-2}}>0.$$

Therefore,

$$\min_{w\geq 0}\ wx^2+\phi_p(w) \tag{12}$$

admits a unique optimal solution $w_{\rm opt}=(p/2)(x^2+\epsilon)^{(p-2)/2}$, which can be obtained by simply checking the first order optimality condition. Substituting $w_{\rm opt}$ back into the cost of (12), the minimum cost is $(x^2+\epsilon)^{p/2}$.

By Lemma 1, Problem (10) can be re-expressed as the following problem:

$$\min_{\substack{\mathbf{A},\mathbf{B},\mathbf{C},\\ \{w_i\geq 0\}}} \sum_{i=1}^{I} w_i\left\|\underline{\mathbf{X}}^{(3)}(:,i)-(\mathbf{C}\odot\mathbf{B})\left(\mathbf{A}(i,:)\right)^T\right\|_2^2 + \sum_{i=1}^{I}\phi_p(w_i). \tag{13}$$

The structure of Problem (13) is nice: it allows us to optimize its cost with respect to (w.r.t.) the four blocks $\mathbf{A},\mathbf{B},\mathbf{C}$, and $\{w_i\}_{i=1}^I$ in an alternating optimization fashion, fixing three blocks and updating one each time.[1] As we will show next, each conditional optimization problem has a closed-form solution.

First, the problem w.r.t. $\mathbf{A}$ is separable w.r.t. i. For each i, the problem w.r.t. $\mathbf{A}(i,:)$ is a simple least squares problem. Hence, the subproblem w.r.t. $\mathbf{A}$ admits the following closed-form solution:

$$\mathbf{A}=\left((\mathbf{C}\odot\mathbf{B})^{\dagger}\underline{\mathbf{X}}^{(3)}\right)^T,$$

which is the same as that in the plain TALS [1], [2]. Notice that in practice, we compute $\mathbf{A}$ by the following expression:

$$\mathbf{A}^T=\left(\mathbf{C}^T\mathbf{C}\circledast\mathbf{B}^T\mathbf{B}\right)^{-1}(\mathbf{C}\odot\mathbf{B})^T\underline{\mathbf{X}}^{(3)}.$$

In practice, the matrix inversion part and $(\mathbf{C}\odot\mathbf{B})^T\underline{\mathbf{X}}^{(3)}$ should be computed separately. The reasons are as follows. First, the inversion part, i.e., $\left(\mathbf{C}^T\mathbf{C}\circledast\mathbf{B}^T\mathbf{B}\right)^{-1}$, is usually the inverse of a small (R-by-R) matrix. Second, the multiplication of a Khatri-Rao structured matrix and an unfolded tensor is a computationally expensive operation if I,J,K are large (specifically, this single step costs $2RIJK$ flops), but fast algorithms are available when $\underline{\mathbf{X}}$ is sparse [9], [34]–[37].

To update $\mathbf{B}$, we consider using the lateral slabs $\{\mathbf{C}\mathbf{D}_j(\mathbf{B})\mathbf{A}^T\}_{j=1}^J$. From Problem (13), it can be readily seen that the ith column of $\{\mathbf{C}\mathbf{D}_j(\mathbf{B})\mathbf{A}^T\}$ is scaled by $\sqrt{w_i}$. Thus, the subproblem w.r.t. $\mathbf{B}$ can be written as

$$\min_{\mathbf{B}}\sum_{j=1}^{J}\left\|\mathbf{X}_j^{(1)}\mathbf{W}-\mathbf{C}\mathbf{D}_j(\mathbf{B})\mathbf{A}^T\mathbf{W}\right\|_F^2,$$

where $\mathbf{W}=\mathrm{Diag}(\sqrt{w_1},\ldots,\sqrt{w_I})$, or, in the following more compact form,

$$\min_{\mathbf{B}}\left\|(\mathbf{W}\otimes\mathbf{I})\underline{\mathbf{X}}^{(1)}-\left((\mathbf{W}\mathbf{A})\odot\mathbf{C}\right)\mathbf{B}^T\right\|_F^2.$$

The above is still a least squares problem. Therefore, the solution is simply

$$\mathbf{B}=\left(\left((\mathbf{W}\mathbf{A})\odot\mathbf{C}\right)^{\dagger}(\mathbf{W}\otimes\mathbf{I})\underline{\mathbf{X}}^{(1)}\right)^T.$$

In practice, the above solution can be written as follows:

$$\begin{aligned}\mathbf{B}^T&=(\mathbf{W}\mathbf{A}\odot\mathbf{C})^{\dagger}(\mathbf{W}\otimes\mathbf{I})\underline{\mathbf{X}}^{(1)}, &\text{(14a)}\\ &=\left((\mathbf{W}\mathbf{A}\odot\mathbf{C})^T(\mathbf{W}\mathbf{A}\odot\mathbf{C})\right)^{-1}\\ &\quad\times(\mathbf{W}\mathbf{A}\odot\mathbf{C})^T(\mathbf{W}\otimes\mathbf{I})\underline{\mathbf{X}}^{(1)} &\text{(14b)}\\ &=\left(\mathbf{A}^T\mathbf{W}^2\mathbf{A}\circledast\mathbf{C}^T\mathbf{C}\right)^{-1}\left(\mathbf{W}^2\mathbf{A}\odot\mathbf{C}\right)^T\underline{\mathbf{X}}^{(1)}, &\text{(14c)}\end{aligned}$$

where we have used the property

$$(\mathbf{U}_1\odot\mathbf{V}_1)^T(\mathbf{U}_2\odot\mathbf{V}_2)=\mathbf{U}_1^T\mathbf{U}_2\circledast\mathbf{V}_1^T\mathbf{V}_2$$

to obtain (14b), and

$$(\mathbf{U}_1\otimes\mathbf{V}_1)^T(\mathbf{U}_2\odot\mathbf{V}_2)=\mathbf{U}_1^T\mathbf{U}_2\odot\mathbf{V}_1^T\mathbf{V}_2$$

[1] A similar auxiliary variable-based technique for splitting *convex* ℓ_p norms ($1\leq p<2$) has appeared in [32], [33]. Lemma 1 can be considered as a nonconvex extension of the prior works in [32], [33].

to reach (14c). Putting $\mathbf{B}$ in the form of (14c) is important. The reason is twofold: First, one does not have to actually compute and save $(\mathbf{W}\otimes\mathbf{I})\underline{\mathbf{X}}^{(1)}$ since saving IJK elements after each iteration is cumbersome when I, J, K are large (e.g., for $I=J=K=100$, a million variables have to be saved in each iteration). Second, the efficient solvers for computing the product of a Khatri-Rao structured matrix and an unfolded tensor can be directly applied to $(\mathbf{W}^2\mathbf{A}\odot\mathbf{C})^T\underline{\mathbf{X}}^{(1)}$.

To update $\mathbf{C}$, the rationale follows that of updating $\mathbf{B}$. Specifically, as the ith row of each frontal slab is scaled by $\sqrt{w_i}$, we have can express the conditional problem w.r.t. $\mathbf{C}$ as

$$\min_{\mathbf{C}} \sum_{k=1}^{K} \left\| \mathbf{W}\mathbf{X}_k^{(2)} - \mathbf{W}\mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{B}^T \right\|_F^2,$$

and the solution is also in closed form:

$$\mathbf{C} = \left((\mathbf{B}\odot\mathbf{W}\mathbf{A})^{\dagger}(\mathbf{I}\otimes\mathbf{W})\underline{\mathbf{X}}^{(2)} \right)^T.$$

Similar to the $\mathbf{B}$ case, we can express $\mathbf{C}^T$ as

$$\mathbf{C}^T = (\mathbf{B}^T\mathbf{B}\circledast\mathbf{A}^T\mathbf{W}^2\mathbf{A})^{-1}(\mathbf{B}\odot\mathbf{W}^2\mathbf{A})^T\underline{\mathbf{X}}^{(2)},$$

and thus $(\mathbf{B}^T\mathbf{B}\circledast\mathbf{A}^T\mathbf{W}^2\mathbf{A})^{-1}$ and $(\mathbf{B}\odot\mathbf{W}^2\mathbf{A})^T\underline{\mathbf{X}}^{(2)}$ can be computed separately, if necessary in practice.

The update w.r.t. $\{w_i\}_{i=1}^I$ follows Lemma 1, i.e.,

$$w_i := \frac{p}{2}\left(\left\| \underline{\mathbf{X}}^{(3)} - (\mathbf{C}\odot\mathbf{B})\mathbf{A}^T(i,:) \right\|_2^2 + \epsilon \right)^{\frac{p-2}{2}}, \quad \forall i.$$

Given these conditional updates, a simple strategy is to cyclically update $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\{w_1,\ldots,w_I\}$. The algorithm is summarized in Algorithm 1; we will henceforth refer to it as *Iteratively Reweighted Alternating Least Squares* (IRALS), since $w_1,\ldots,w_I$ can be interpreted as weights applied to the frontal slabs. From an algorithmic structure viewpoint, IRALS can be considered as an extension of the iteratively reweighted least squares (IRLS) algorithm [30] to tensor factorization. Since each partial minimization does not increase the value of the cost function and the function is lower bounded by zero, IRALS *guarantees the convergence of the cost function of Problem* (13).

Remark 1: One may notice that we have not characterized the convergence of the solution sequence produced by IRALS yet. By some existing theories of alternating optimization, a stationary point for TALS and IRALS may be attained if the conditional objective function of every block is strictly convex and is continuously differentiable on the interior of the feasible set *throughout all iterations* [17, Proposition 2.7.1]. In our context, this requires $\mathrm{rank}(\mathbf{B}\odot\mathbf{A}) = \mathrm{rank}(\mathbf{C}\odot\mathbf{B}) = \mathrm{rank}(\mathbf{A}\odot\mathbf{C}) = R$ *throughout all iterations*, which is hard to check [38]. Nevertheless, convergence to a stationary point of Problem (10) can be shown by employing some variants of alternating optimization, e.g., *maximum block improvement* (MBI) [18]. In this work, we adopt cyclic alternating optimization instead of MBI, for implementation simplicity and speed. Also, in practice, we are often interested in PARAFAC with regularization on the loading factors; in such cases, convergence to a stationary point of alternating optimization is usually not a problem any more [38]—see the next section for details.

Remark 2: Until now, we have been dealing with the problem of interest (i.e., Problem (10)) *indirectly*. It is interesting to consider the relationship between the solutions of our working problem, i.e., Problem (13), and Problem (10). It can be shown that

Algorithm 1: IRALS

input : $\underline{\mathbf{X}}$; , $\mathbf{B}_0$, $\mathbf{C}_0$ (initialization); and $p\in(0,1]$.
$\mathbf{B}=\mathbf{B}_0$;
$\mathbf{C}=\mathbf{C}_0$;
$\mathbf{W}=\mathbf{I}$;
repeat
  $\mathbf{A} := \left((\mathbf{C}^T\mathbf{C}\circledast\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{C}\odot\mathbf{B})^T\underline{\mathbf{X}}^{(3)} \right)^T$
  $\mathbf{B} := \left((\mathbf{A}^T\mathbf{W}^2\mathbf{A}\circledast\mathbf{C}^T\mathbf{C})^{-1}(\mathbf{W}^2\mathbf{A}\odot\mathbf{C})^T\underline{\mathbf{X}}^{(1)} \right)^T$;
  $\mathbf{C} := \left((\mathbf{B}^T\mathbf{B}\circledast\mathbf{A}^T\mathbf{W}^2\mathbf{A})^{-1}(\mathbf{B}\odot\mathbf{W}^2\mathbf{A})^T\underline{\mathbf{X}}^{(2)} \right)^T$;
  $w_i := \frac{p}{2}\left(\left\| \underline{\mathbf{X}}^{(3)}(:,i) - (\mathbf{C}\odot\mathbf{B})\mathbf{A}^T(i,:) \right\|_2^2 + \epsilon \right)^{\frac{p-2}{2}}, \quad \forall i$;
  $\mathbf{W}^2 := \mathrm{Diag}(w_1,\ldots,w_I)$;
until *some stopping criterion is satisfied*;
output: $\mathbf{B}$, $\mathbf{C}$.

Claim 2: Assume that $(\mathbf{A}^\star,\mathbf{B}^\star,\mathbf{C}^\star,\{w_i^\star\}_{i=1}^I)$ is a stationary point of Problem (13). Then, $(\mathbf{A}^\star,\mathbf{B}^\star,\mathbf{C}^\star)$ is also a stationary point of Problem (10).

The proof of Claim 2 can be found in Appendix B. The key step is to invoke the uniqueness of the subproblem w.r.t. $\{w_i\}$ following Lemma 1 and marginalize it. By this claim, we see that dealing with Problem (13) can yield a stationary point of Problem (10), whenever a limit point is reached.

VI. Extension: Constrained and Regularized Robust Tensor Factorization

In this section, we consider practical extensions of IRALS, namely, constrained and regularized optimization.

A. Adding Constraints and Regularization

In data analytics, constrained and regularized low-rank tensor factorization often makes a lot of sense, since combining different types of *a priori* information may help find interpretable factors when modeling error and noise exist. Hence, there are many cases in which we are interested in solving the following problem:

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \quad & \frac{1}{2}\sum_{i=1}^{I}\left(\left\| \underline{\mathbf{X}}^{(3)}(:,i) - (\mathbf{C}\odot\mathbf{B})\mathbf{A}^T(i,:) \right\|_2^2 + \epsilon \right)^{p/2} \\ & \qquad + \lambda_a f(\mathbf{A}) + \lambda_b g(\mathbf{B}) + \lambda_c h(\mathbf{C}), \\ \text{s.t.} \quad & \mathbf{A}\in\mathcal{A},\ \mathbf{B}\in\mathcal{B},\ \mathbf{C}\in\mathcal{C}, \end{aligned} \tag{15}$$

where λ_a, λ_b and λ_c are nonnegative regularization parameters, $f(\mathbf{A})$, $g(\mathbf{B})$ and $g(\mathbf{C})$ are appropriate regularization functions, and $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ represent (hard) constraints on the loading factors.

In many cases, the constraints of interest include nonnegativity of the loadings, stemming from physical, chemical, or modeling considerations—e.g., concentrations, spectra, and e-mail counts are all nonnegative, and nonnegativity of the latent factors is important in social network mining [27] and in fluorescence spectroscopy [14]. More general 'box' constraints of type $a_l \le \mathbf{A}(i,r) \le a_h$ may also be appropriate, e.g.,

when we also have prior knowledge on the maximum possible concentration.

Soft constraints may also be of interest, and these can be represented using appropriate regularization terms. If we know that the columns of $\mathbf{B}$ should be smooth, for example, we can employ the regularization $g(\mathbf{B}) = \|\mathbf{T}\mathbf{B}\|_F^2$, where [39]

$$\mathbf{T} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & \cdots & \cdots \\ 0 & 1 & -2 & 1 & 0 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & 0 & 1 & -2 & 1 \end{bmatrix}. \quad (16)$$

If we know that the loading factors are sparse, we can use $\|\cdot\|_1$ or any other sparsity-promoting function for regularization. As alluded to in Remark 1, adding regularization often brings side-benefits in terms of accelerating convergence, avoiding swamps, and attaining a stationary point [38], [40]. For example, by adding the minimum-norm regularization $\|\mathbf{A}\|_F^2$, $\|\mathbf{B}\|_F^2$, and $\|\mathbf{C}\|_F^2$, it can be easily seen that each block always has a unique solution, and thus a stationary point of Problem (15) can be attained by alternating optimization, whenever a limit point exists. Given the overall objective (15), and by Lemma 1, we may consider the equivalent reformulation

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B},\mathbf{C},\{w_i\}} \quad & \sum_{i=1}^{I} \frac{w_i}{2} \left\| \underline{\mathbf{X}}^{(3)}(:,i) - (\mathbf{C} \odot \mathbf{B})\mathbf{A}^T(i,:) \right\|_2^2 \\ & + \frac{w_i}{2} \sum_{i=1}^{I} \phi_p(w_i) + \lambda_a f(\mathbf{A}) + \lambda_b g(\mathbf{B}) + \lambda_c h(\mathbf{C}), \\ \text{s.t.} \quad & \mathbf{A} \in \mathcal{A},\ \mathbf{B} \in \mathcal{B},\ \mathbf{C} \in \mathcal{C}, \\ & w_i \geq 0, \quad i = 1, \ldots, I. \end{aligned}$$

The subproblem w.r.t. $\{w_i\}$ admits the same solution as before. In addition, the subproblems w.r.t. the loading factors are constrained and regularized least squares problems. To describe our treatment, we begin with the subproblem w.r.t. $\mathbf{B}$:

$$\begin{aligned} \min_{\mathbf{B}} \quad & \frac{1}{2} \left\| (\mathbf{W} \otimes \mathbf{I}) \underline{\mathbf{X}}^{(1)} - ((\mathbf{W}\mathbf{A}) \odot \mathbf{C}) \mathbf{B}^T \right\|_F^2 + \lambda g(\mathbf{B}) \\ \text{s.t.} \quad & \mathbf{B} \in \mathcal{B}. \end{aligned}$$

To handle this problem, we propose the following *alternating direction method of multipliers* (ADMM) [19] based approach. We first rewrite the problem as

$$\begin{aligned} \min_{\mathbf{B},\, \mathbf{B}_1,\, \mathbf{B}_2} \quad & \frac{1}{2} \left\| (\mathbf{W} \otimes \mathbf{I}) \underline{\mathbf{X}}^{(1)} - ((\mathbf{W}\mathbf{A}) \odot \mathbf{C}) \mathbf{B}_1^T \right\|_F^2 \\ & + \lambda g(\mathbf{B}_2) + \mathbf{1}_{\mathcal{B}}(\mathbf{B}) \\ \text{s.t.} \quad & \mathbf{B} = \mathbf{B}_1 \\ & \mathbf{B} = \mathbf{B}_2. \end{aligned} \quad (17)$$

where $\mathbf{1}_{\mathcal{X}}(\mathbf{X})$ is 0 for $\mathbf{X} \in \mathcal{X}$ and ∞ otherwise. ADMM solves the following augmented Lagrangian dual of Problem (17) [19, Chapter 3]:

$$\begin{aligned} \max_{\mathbf{U}_1,\, \mathbf{U}_2} \min_{\mathbf{B},\, \mathbf{B}_1,\, \mathbf{B}_2} \quad & \frac{1}{2} \left\| (\mathbf{W} \otimes \mathbf{I}) \underline{\mathbf{X}}^{(1)} - ((\mathbf{W}\mathbf{A}) \odot \mathbf{C}) \mathbf{B}_1^T \right\|_F^2 \\ & + \lambda_b g(\mathbf{B}_2) + \mathbf{1}_{\mathcal{B}}(\mathbf{B}) + \frac{\rho}{2} \|\mathbf{B} - \mathbf{B}_1 + \mathbf{U}_1\|_F^2 \\ & + \frac{\rho}{2} \|\mathbf{B} - \mathbf{B}_2 + \mathbf{U}_2\|_F^2, \end{aligned} \quad (18)$$

where $\mathbf{U}_1$ and $\mathbf{U}_2$ are the dual variables, and $\rho > 0$ is the stepsize parameter that is pre-specified. The standard ADMM updates for Problem (18) are as follows [19, Chapter 3]:

$$\begin{aligned} \mathbf{B}_1 &:= \arg\min_{\mathbf{B}_1} \quad \frac{1}{2} \left\| (\mathbf{W} \otimes \mathbf{I}) \underline{\mathbf{X}}^{(1)} - ((\mathbf{W}\mathbf{A}) \odot \mathbf{C}) \mathbf{B}_1^T \right\|_F^2 \\ &\qquad + \frac{\rho}{2} \|\mathbf{B} - \mathbf{B}_1 + \mathbf{U}_1\|_F^2 && (19a) \\ \mathbf{B}_2 &:= \arg\min_{\mathbf{B}_2} \quad \lambda_b g(\mathbf{B}_2) + \frac{\rho}{2} \|\mathbf{B} - \mathbf{B}_2 + \mathbf{U}_2\|_F^2 && (19b) \\ \mathbf{B} &:= \arg\min_{\mathbf{B}} \quad \frac{\rho}{2} \|\mathbf{B} - \mathbf{B}_2 + \mathbf{U}_2\|_F^2 \\ &\qquad + \frac{\rho}{2} \|\mathbf{B} - \mathbf{B}_1 + \mathbf{U}_1\|_F^2 + \mathbf{1}_{\mathcal{B}}(\mathbf{B}), && (19c) \\ \mathbf{U}_1 &:= \mathbf{U}_1 + \mathbf{B} - \mathbf{B}_1, && (19d) \\ \mathbf{U}_2 &:= \mathbf{U}_2 + \mathbf{B} - \mathbf{B}_2, && (19e) \end{aligned}$$

The proposed variable-splitting strategy brings several advantages. First, the problem w.r.t. $\mathbf{B}_1$ (i.e., Problem (19a)) is a least squares problem, whose solution is

$$\mathbf{B}_1^T := (\mathbf{A}^T\mathbf{W}^2\mathbf{A} \circledast \mathbf{C}^T\mathbf{C} + \rho\mathbf{I})^{-1} \left((\mathbf{W}^2\mathbf{A} \odot \mathbf{C})^T \underline{\mathbf{X}}^{(1)} + \mathbf{M} \right),$$

where $\mathbf{M} = \rho(\mathbf{B} + \mathbf{U}_1)$. We see that the structure of $(\mathbf{W}^2\mathbf{A} \odot \mathbf{C})^T \underline{\mathbf{X}}^{(1)}$ has been preserved, and thus efficient solvers for this matrix multiplication problem can be applied when the tensor is large and sparse [9], [34]–[37]. Second, the $\mathbf{B}_2$ update is a proximal operator, which can be put in simple closed-form for many $g(\cdot)$'s. Let us consider $g(\mathbf{B}_2) = \frac{1}{2}\|\mathbf{T}\mathbf{B}_2\|_F^2$ as an example, which is often used for promoting smooth $\mathbf{B}$. The $\mathbf{B}_2$ update is then simply

$$\mathbf{B}_2 := (\lambda_b \mathbf{T}^T\mathbf{T} + \rho\mathbf{I})^{-1}(\mathbf{B} + \mathbf{U}_2).$$

Note that when $\mathbf{T} = \mathbf{I}$, this further reduces to $\mathbf{B}_2 = \frac{1}{\lambda_b + \rho}(\mathbf{B} + \mathbf{U}_2)$—which is useful to control the scaling of $\mathbf{B}$. Also, if one wants to promote sparsity in $\mathbf{B}$, several convex and nonconvex $g(\cdot)$'s that enable closed-form solution of Problem (19b) can be employed; see [41]. Third, the $\mathbf{B}$ update (Problem (19c)) also has a simple form:

$$\mathbf{B} := \left(\frac{1}{2}(\mathbf{B}_2 - \mathbf{U}_2 + \mathbf{B}_1 - \mathbf{U}_1) \right)_{\mathcal{B}},$$

where $(\cdot)_{\mathcal{B}}$ is a projector to the set $\mathcal{B}$. For many constraints $\mathcal{B}$, this projection step is fairly simple. For example, if $\mathcal{B} = \mathbb{R}_+$, the projection is

$$\mathbf{B} := \left(\frac{1}{2}(\mathbf{B}_2 - \mathbf{U}_2 + \mathbf{B}_1 - \mathbf{U}_1) \right)_{+},$$

where $(\cdot)_+$ is an element-wise operator such that $(x)_+ = \max\{x, 0\}$; see many other efficient projections in [19].

The ADMM updates w.r.t. $\mathbf{C}$ and $\mathbf{A}$ are similar; they are relegated to Appendix C. Overall, we solve the subproblems w.r.t. $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{W}$ cyclically as in the last section except that the former three are solved by ADMM. We should mention that the convergence properties of the described algorithm depend on the type of regularization and the constraints that are added. The reason is twofold. First, as previously mentioned, to ensure that every limit point of the solution sequence $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{W}\}$ is a stationary point of Problem (15), each subproblem w.r.t. $\mathbf{A}$,

$\mathbf{B}$ and $\mathbf{C}$ needs to be a convex problem admitting a unique solution, which depends on the type of regularization and the constraints. In addition, the convexity of a subproblem also affects the solution of ADMM—it guarantees that ADMM can attain the optimal solution of that subproblem.

B. Initialization Approaches

IRALS requires initial guesses of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. In practice, several initialization approaches can be considered:

• First, random initialization is viable. Since the considered problem is nonconvex, using random initialization may require restarting the algorithm several times from random initial points to attain a good solution, but it also helps the algorithm to avoid 'bad' local minima.

• Second, the loading factors estimated by algorithms that deal with the ℓ_2-norm fitting-based PARAFAC problem in (4) (or ℓ_2 PARAFAC for simplicity), e.g., TALS, can be used as starting points. Algorithms tackling the variants of Problem (4) with constraints and regularization on the loading factors can also be employed. This approach is effective when those algorithms are not totally thrown off by the outlying slabs.

• Third, when I is larger than JK, one can first estimate an orthogonal basis $\mathbf{U} \in \mathbb{R}^{KJ \times R}$ such that $\mathcal{R}(\mathbf{U}) = \mathcal{R}(\mathbf{C} \odot \mathbf{B})$, and then apply a Khatri-Rao subspace-based PARAFAC algorithm, such as those in [10], [22], [42], [43], on the extracted $\mathbf{U}$ to get an initial guess of $(\mathbf{B}, \mathbf{C})$. Khatri-Rao subspace-based initialization is effective with large I since the procedure can 'compress' the original tensor substantially,[2] and PARAFAC algorithms empirically work better when the data size is smaller. When there is no outlying slab, and when both $\mathbf{C} \odot \mathbf{B}$ and $\mathbf{A}$ have full-column rank, a basis of $\mathcal{R}(\mathbf{C} \odot \mathbf{B})$ can be obtained by applying singular value decomposition (SVD) on $\underline{\mathbf{X}}^{(3)}$. Here, since there are outlying columns of $\underline{\mathbf{X}}^{(3)}$, we can estimate $\mathbf{U}$ using robust SVD, which has been intensively studied in the recent literature; see, e.g., [44], [45].

VII. Numerical Results

In this section, we first use synthetic data to verify our ideas. Then, real-data experiments will be presented to show the effectiveness of the proposed algorithmic framework in practice. The algorithms presented in this section are all implemented in Matlab, and all simulations and experiments were carried out on a desktop computer with an i7 3.4 GHz quad-core CPU and 8 GB RAM.

A. Synthetic Data Simulations

In this subsection, we generate the non-negative loading factors of three-way tensors following the exponential distribution with $\mu = 1$. The outlier elements are uniformly distributed within zero and one, and then are scaled to satisfy the specified simulation conditions (see below). To quantify the corruption level, we define the signal-to-outlier ratio (SOR) as

$$\mathrm{SOR(dB)} = 10 \log_{10} \left(\frac{(1/I) \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \underline{\mathbf{X}}^2(i,j,k)}{(1/|\mathcal{N}|) \sum_{i \in \mathcal{N}} \|\mathbf{O}_i\|_F^2} \right)$$

[2]To be specific, $\mathbf{U} = (\mathbf{C} \odot \mathbf{B})\mathbf{\Theta}$, for some $\mathbf{\Theta} \in \mathbb{R}^{R \times R}$, can be considered as a compressed tensor which has only R slabs, whereas the original tensor has I slabs. If $R \ll I$, the compressed tensor has many fewer slabs.

TABLE I
The Average MSEs of the Estimated $\mathbf{B}$ and $\mathbf{C}$ by the Algorithms Under Various SORs; $(I, J, K) = (20, 20, 20)$; no. of outlying slabs $= 6$

$R = 5$						
Algorithm	Measure	SOR				
		-10	-5	0	5	10
TALS	MSE (dB)	-10.3345	-13.9955	-20.6065	-28.3189	-34.2078
	TIME (sec.)	0.0655	0.0635	0.0579	0.0534	0.0534
L1 PARAFAC	MSE (dB)	-11.6272	-19.6811	-25.499	-28.156	-66.0477
	TIME (sec.)	15.1952	12.8005	10.9255	10.4323	10.1826
IRALS	MSE (dB)	**-28.6011**	-46.3832	**-76.4109**	**-129.469**	-127.115
	TIME (sec.)	0.2576	0.1857	0.1496	0.1452	0.1423
IRALS w./ nn	MSE (dB)	-28.5889	**-64.3808**	-73.2943	-129.468	**-127.125**
	TIME (sec.)	8.2756	5.7996	4.6465	4.2905	4.1048
$R = 10$						
Algorithm	Measure	SOR				
		-10	-5	0	5	10
TALS	MSE (dB)	-9.6438	-12.9671	-16.4158	-24.1415	-27.1827
	TIME (sec.)	0.1955	0.139	0.1174	0.0959	0.0869
L1 PARAFAC	MSE (dB)	-7.5341	-10.0898	-13.4314	-15.9167	-17.7212
	TIME (sec.)	79.3323	67.8216	52.7992	44.2532	36.9568
IRALS	MSE (dB)	-19.4927	-29.2948	-39.594	-38.0696	**-68.5139**
	TIME (sec.)	0.573	0.5305	0.4496	0.3681	0.305
IRALS w./ nn	MSE (dB)	**-22.5658**	**-33.7498**	**-54.5883**	**-60.9308**	-67.8565
	TIME (sec.)	16.3145	16.4172	14.061	11.8115	9.6166

To benchmark our algorithm, we employ TALS for ℓ_2 PARAFAC fitting (i.e., Problem (4)) and the ℓ_1-norm fitting based PARAFAC (ℓ_1 PARAFAC) [16] with the alternating weighted median filtering realization. We fix $p = 0.5$ throughout this section; our experience is that that the results obtained for different $p \in [0.1, 1]$ are qualitatively similar to those obtained for $p = 0.5$. IRALS is stopped when the absolute change of the objective value is less than 10^{-8} or the number of iterations reaches 1000. For IRALS with constraints, we stop the ADMM algorithms for the subproblems when $\|\mathbf{B} - \mathbf{B}_1\|_2 + \|\mathbf{B} - \mathbf{B}_2\|_2 \leq 10^{-3}$ following the guidelines in [19]. IRALS and IRALS with constraints are initialized by plain TALS in this subsection.

Table I shows the average mean-squared-errors (MSEs) of the estimated $\mathbf{B}$ and $\mathbf{C}$ by the algorithms under various SORs; the runtime performance is also presented in this table. The MSE of the estimated $\mathbf{B}$ is defined as

$$\mathrm{MSE} = \min_{\substack{\boldsymbol{\pi} \in \Pi, \\ c_1, \ldots, c_J \in \{\pm 1\}}} \frac{1}{K} \sum_{j=1}^{J} \left\| \frac{\mathbf{B}(:,j)}{\|\mathbf{B}(:,j)\|_2} - c_k \frac{\hat{\mathbf{B}}(:,\pi_j)}{\left\|\hat{\mathbf{B}}(:,\pi_j)\right\|_2} \right\|_2^2,$$

where Π is the set of all permutations of $\{1, 2, \ldots, K\}$, and $\mathbf{B}(:,j)$ and $\hat{\mathbf{B}}(:,j)$ are the ground truth of the jth column of $\mathbf{B}$ and the corresponding estimate, respectively; the same definition of MSE holds for $\hat{\mathbf{C}}$. We see that for $R = 5$, IRALS and IRALS with non-negativity constraints (denoted by 'IRALS w./nn') both exhibit much lower MSEs compared to TALS and ℓ_1 PARAFAC. When $R = 10$, IRALS with non-negativity constraints gives the best MSE performance in general. In terms of runtime, the unconstrained IRALS and the IRALS with non-negativity constraints are both faster than ℓ_1 PARAFAC. Notably, unconstrained IRALS is more than 50 times faster than ℓ_1 PARAFAC in the presented simulations in this table.

Table II shows the MSEs and runtimes versus the number of outlying slabs. In many cases of this simulation, ℓ_1 PARAFAC could not yield a reasonable result, and thus it was removed from the comparison. For the other three algorithms, we see that IRALS and IRALS with non-negativity constraints can

TABLE II
THE AVERAGE MSES OF THE ESTIMATED $\mathbf{B}$ AND $\mathbf{C}$ BY THE ALGORITHMS VERSUS THE NUMBER OF OUTLYING SLABS; $(I, J, K) = (20, 20, 20)$; SOR $= 0$ dB; $R = 5$

Algorithm	Measure	number of outlying slabs				
		3	5	7	9	11
TALS	MSE (dB)	-15.0081	-11.4041	-9.4724	-8.4138	-8.0331
	TIME (sec.)	0.0592	0.063	0.0635	0.0681	0.0718
IRALS	MSE (dB)	-30.4836	-31.3318	-24.839	-23.3083	-23.1527
	TIME (sec.)	0.1723	0.217	0.2349	0.2474	0.2461
IRALS w./ nn	MSE (dB)	**-37.743**	**-40.8854**	**-40.8**	**-40.4385**	**-26.3071**
	TIME (sec.)	3.4883	4.5614	5.3182	5.7222	5.2961

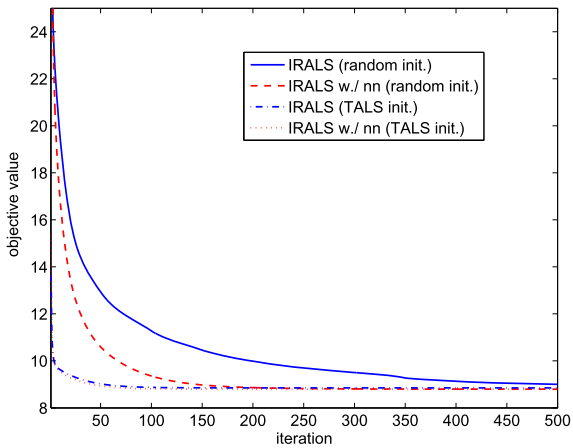


Fig. 3. The convergence curves of the objective value when applying IRALS with different initializations and constraints.

yield reasonable estimation of the loading factors even when the number of outlying slabs exceeds a half of the total number of slabs, but TALS gives very poor estimation in this case.

Fig. 3 presents the objective values of (13) against the iterations when applying IRALS and IRALS with nonnegativity constraints with different initializations. This simulation is under the settings $I = J = K = 20$, $R = 5$, and $|\mathcal{N}| = 6$. Each curve is averaged from 100 trials. We see that, when using random initialization, the cost function of IRALS with nonnegativity constraints on the loading factors converges much faster than that of IRALS with no constraints. In addition, using the output of TALS helps the cost functions of both algorithms converge faster. Specifically, under such an initialization scheme, the objective values given by the algorithms both converge within 100 iterations.

B. Blind Speech Separation

In this subsection, we revisit the blind speech separation problem that has been mentioned in Section III. We first show a simulation using instantaneously mixed speech sources, where the mixtures follow the signal model in (6). The sources are randomly picked from a database that consists of 23 speech segments; each source has a length of 3 second, and is sampled at a rate of 16 KHz. We use $I = 5$ sensors and $R = 6$ sources, which poses a challenging under-determined blind separation problem. Each time frame consists of 200 samples—this results in $K = 239$ time frames (slabs). Spatially and temporally white Gaussian noise is added to the received signals. Each local covariance of the received signals (i.e., each slab of the

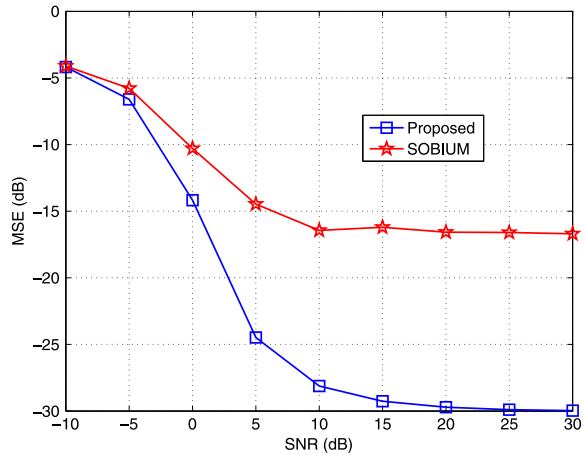


Fig. 4. The MSEs of the estimated mixing systems obtained by SOBIUM and the proposed algorithm under various SNRs.

PARAFAC model) is calculated using the local sample mean of $\mathbf{x}(t)(\mathbf{x}(t))^T$, and the noise variance is estimated by

$$\hat{\sigma}^2 = \min_{k=1,\ldots,K} \lambda_{\min}\left(\underline{\mathbf{X}}(:,:,k)\right),$$

where $\lambda_{\min}(\mathbf{X})$ denotes the smallest eigenvalue of $\mathbf{X}$. The estimated noise variance is then removed from the data; see [10], [11] for details. The mixing system estimation problem can be formulated as

$$\min_{\mathbf{A},\mathbf{C}} \sum_{k=1}^{K} \left(\left\| \underline{\mathbf{X}}(:,:,k) - \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{A}^T \right\|_F^2 + \epsilon \right)^{\frac{p}{2}},$$

and we apply IRALS to the above by treating $\mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{A}^T$ as $\mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{B}^T$. In this subsection, we use the Khatri-Rao subspace-based initialization as mentioned in Section VI-B, since the number of slabs (K in this case) is large.

Fig. 4 shows the average MSEs of the estimated mixing system obtained by several algorithms; the result is averaged from 100 independent trials. The benchmarked PARAFAC algorithm is SOBIUM [22], which is known as a state-of-the-art blind source separation algorithm for the under-determined case (i.e., $I < J$). We see that the proposed algorithm consistently yields around 15 dB lower MSE than that of SOBIUM, which is a significant performance boost. This phenomenon verifies the existence of (significant) modeling error at some slabs, and also shows the effectiveness of our proposed algorithm.

We also consider the convolutive mixture case, in which the signal model can be represented as

$$\mathbf{x}(t) = \sum_{\ell=0}^{\ell_{\max}-1} \mathbf{H}(\ell)\mathbf{s}(t-\ell),$$

where $\mathbf{x}(t)$ and $\mathbf{s}(t)$ are defined as before, and $\mathbf{H}(\ell)$ denotes the mixing system impulse response at time lag ℓ. The convolutive mixture model is more realistic, since it captures the multi-path reverberation characteristics of real acoustic environments; but is also far more challenging to deal with, compared to the instantaneous mixture case. We build up the convolutive mixtures by setting up a simulated room with multiple paths between the speakers and receivers following the image method [46]. To separate the sources, we follow the frequency-domain approach [6], [7]—the basic idea is to transform the mixtures to the frequency domain, where the per-fre-

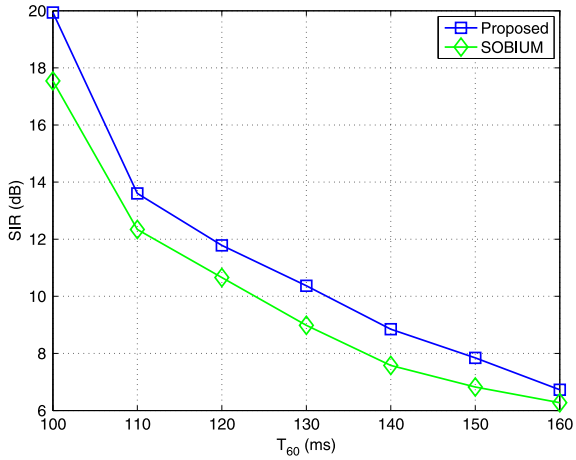


Fig. 5. The SIRs obtained by applying SOBIUM and the proposed algorithm to convolutive mixtures under various T_{60}'s.

quency (bin) mixtures follow an approximately instantaneous mixing model. Thus, PARAFAC algorithms can be applied at each frequency to obtain the source components at that frequency, and the time-domain sources can be obtained subsequently using certain post-processing steps, the most critical of which are permutation and scaling alignment across the different frequency bins. We measure the quality of the unmixed speech signals using the signal-to-interference ratio (SIR) criterion as in [6], [7]; higher SIR means better separation performance. Fig. 5 shows the results of using $I = 4$ sensors to separate $J = 3$ sources; the result is also averaged from 100 trials with randomly picked sources. We see that, under different reverberation conditions for the simulated room (a larger T_{60} means a more severe multipath effect, thereby a more challenging environment for speech separation), the proposed algorithm consistently outperforms SOBIUM by around 2 dB.

C. Fluorescence Data Analysis

In this subsection, we deal with a real fluorescence EEM data set—the Dorrit data that is available online at http://www.models.life.ku.dk/dorrit. Our working data set has 116 spectral emissions, 18 excitations, and 27 samples, which is a tensor with $I = 27$, $J = 116$ and $K = 18$. The Dorrit data set is known for containing some badly contaminated slabs, even after pre-processed by some automatic scattering removal algorithm [47], and there are also some relatively clean samples in this data set; see Fig. 6. We formulate the problem of estimating the spectral emissions ($\mathbf{B}$) and excitations ($\mathbf{C}$) as

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{i=1}^{I} \left(\left\| \underline{\mathbf{X}}(i,:,:) - \mathbf{B}\mathbf{D}_i(\mathbf{A})\mathbf{C}^T \right\|_F^2 + \epsilon \right)^{\frac{p}{2}} + \lambda_a \|\mathbf{A}\|_F^2 + \lambda_b \|\mathbf{T}\mathbf{B}\|_F^2 + \lambda_c \|\mathbf{T}\mathbf{C}\|_F^2 \quad \text{s.t. } \mathbf{A} \geq \mathbf{0},\ \mathbf{B} \geq \mathbf{0},\ \mathbf{C} \geq \mathbf{0},$$

where $\mathbf{T}$ is defined in (16) with appropriate dimensions. We add smoothness regularization on $\mathbf{B}$ and $\mathbf{C}$ since we know that the emission and the excitation spectra are smooth in practice; also, non-negativity constraints are added to all three loading factors. We should point out that adding $\|\mathbf{A}\|_F^2$ is important; otherwise, the scaling of $\mathbf{B}$ and $\mathbf{C}$ can be 'absorbed' by $\mathbf{A}$, and the smoothness regularization (or, any other scaling-sensitive regularization) may not work.

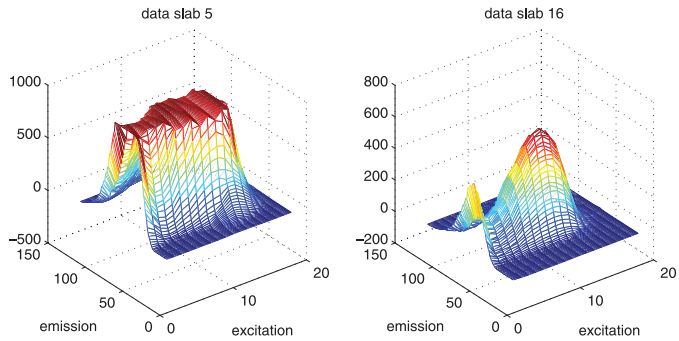


Fig. 6. An outlying slab (left) and a relatively clean slab (right) of the Dorrit data.

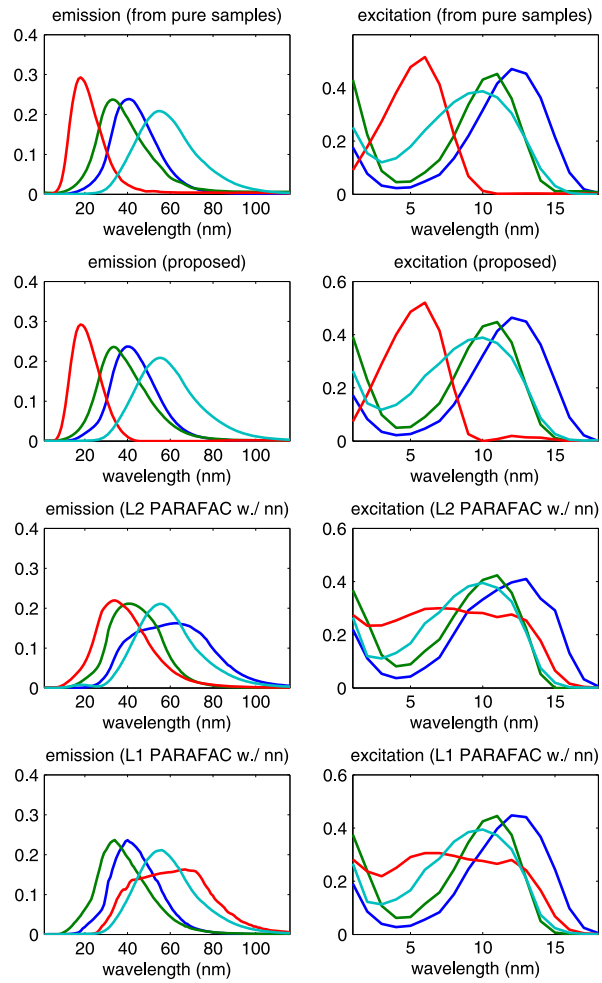


Fig. 7. The estimated emission and excitation curves obtained using the proposed algorithm, as well as nonnegativity-constrained ℓ_2 and ℓ_1 PARAFAC fitting.

In this experiment, we set $\lambda_b = \lambda_c = 10$ and $\lambda_a = 10^{-2}$ and $R = 4$. Here, we use the ℓ_1 and ℓ_2 PARAFAC algorithms with nonnegativity constraints as benchmarks, which are both implemented in the N-way toolbox [48] (available at http://www.models.life.ku.dk/source/nwaytoolbox/). The result of the nonnegativity-constrained ℓ_2 PARAFAC algorithm is used to initialize the proposed algorithm. The estimated $\mathbf{B}$ and $\mathbf{C}$ by the algorithms are shown in Fig. 7. We also provide the emission

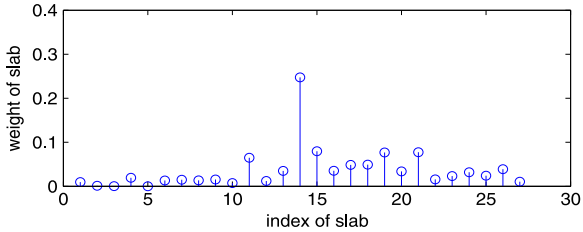


Fig. 8. The normalized weights of the samples obtained via IRALS.

and excitation spectra obtained from certain 'pure samples' containing only a single compound. These pure samples are known from prior studies with this particular dataset, and thus the recovered spectra are believed to be close to the ground truth—see the row tagged as 'from pure samples' in Fig. 7. We see that the spectra estimated by the proposed algorithm are visually very similar to those measured from the pure samples. However, both of the nonnegativity-constrained ℓ_1 and ℓ_2 PARAFAC algorithms yield clearly worse results—for both of them, an estimated emission spectrum and an estimated excitation spectrum are highly inconsistent with the results measured from the pure samples. It is also interesting to observe the weights of the slabs given by the proposed algorithm in Fig. 8. One can see that the algorithm automatically fully downweights slab 5, which is consistent with our observation (consistent with domain expert knowledge) that slab 5 is an extreme outlying sample (cf. Fig. 6). This verifies the effectiveness of our algorithm for joint slab selection and model fitting.

D. ENRON E-mail Data Mining

In this subsection, we apply the proposed algorithm on the celebrated ENRON E-mail corpus. This data set contains the e-mail communications between 184 persons within 44 months. Specifically, $\mathbf{X}(i,j,k)$ denotes the number of e-mails sent by person i to person j within month k. Many studies have been done for mining the social groups out of this data set [26], [27], [49]. In particular, [27] applied a sparsity-regularized and nonnegativity-constrained PARAFAC algorithm on this data set, and some interesting (and interpretable) results have been obtained. In particular, the significant non-zero elements of $\mathbf{A}(:,r)$ usually correspond to persons with similar 'social' positions such as lawyers or executives.

Here, we also aim at mining the social groups out of the ENRON data, while taking data for 'outlying months' into consideration. It is well known that the ENRON company went through a criminal investigation and finally filed for bankruptcy. Hence, one may conjecture that the e-mail interaction patterns between the social groups might be irregular during the outbreak of the crisis. We fit the data using the following formulation:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{k=1}^{K} \left(\left\| \underline{\mathbf{X}}(:,:,k) - \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{B}^T \right\|_F^2 + \epsilon \right)^{\frac{p}{2}} + \lambda_a f(\mathbf{A}) + \lambda_b \|\mathbf{B}\|_F^2 + \lambda_c \|\mathbf{C}\|_F^2 \quad \text{s.t. } \mathbf{A} \geq \mathbf{0},\ \mathbf{B} \geq \mathbf{0},\ \mathbf{C} \geq \mathbf{0},$$

where $f(\mathbf{A})$ is a function that promotes sparsity following the insight in [27]; $\|\mathbf{B}\|_F^2$ and $\|\mathbf{C}\|_F^2$ are added to avoid scaling/counter-scaling issues, as in the previous example. Notice that here we use an aggressive sparsity promoting function $f(\mathbf{A})$ from [41], which itself cannot be put in closed form—notwithstanding, the proximal operator of $f(\mathbf{A})$ can be written in closed-form, and thus is easy to incorporate into our ADMM framework. We fit the ENRON data with $R = 5$ as in [27], and set $\lambda_a = 6.5 \times 10^{-2}$, $\lambda_b = \lambda_c = 10^{-3}$. The same pre-processing as in [27], [49] is applied to the non-zero data to compress the dynamic range; i.e., all the non-zero raw data elements are transformed by an element-wise mapping $x' = \log_2(x) + 1$. As in the last subsection, the proposed algorithm is initialized by the nonnegativity-constrained ℓ_2 PARAFAC algorithm.

Table III shows the five social groups mined from the data, corresponding to the non-zero elements in the five columns of $\mathbf{A}$. We see that these five groups are quite clean, covering 73 ('important') persons out of 184 in total. More interestingly, the algorithm automatically downweights the slabs corresponding to the period when the company was having a crisis—see Fig. 9. This verifies our guess: The interaction pattern during this particular period is not regular, and downweighting these slabs can give us more clean social groups.

VIII. Conclusion

In this work, we considered the problem of low-rank tensor decomposition in the presence of outlying slabs. Several practical motivating applications have been introduced. A conjugate augmented optimization framework has been proposed to deal with the formulated ℓ_p minimization-based factorization problem. The proposed algorithm features similar complexity as the classic TALS algorithm that is not robust to outlying slabs. Regularized and constrained optimization has also been considered by employing an ADMM update scheme. Simulations using synthetic data and experiments using real data have shown that the proposed approach is promising in different pertinent applications such as blind speech separation, fluorescence data spectroscopy, and social network mining.

Appendix

A. Proof of Claim 1

Consider a feasible solution $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$, where $\tilde{\mathbf{A}}(\mathcal{N}_c,:) = \mathbf{A}(\mathcal{N}_c,:)\mathbf{\Pi}\mathbf{\Delta}_a$, $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{\Pi}\mathbf{\Delta}_b$, and $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Delta}_c$. Consequently, it can be seen that for all $i \in \mathcal{N}_c$ we have

$$\mathcal{I}\left(\left\| \underline{\mathbf{X}}^{(3)}(:,i) - (\tilde{\mathbf{C}} \odot \tilde{\mathbf{B}})\tilde{\mathbf{A}}(i,:)^T \right\|_2 \right) = 0.$$

Hence, the optimal value of the cost function satisfies

$$v_{\min} \leq I - |\mathcal{N}_c| \leq \frac{I-c}{2}.$$

Now, we show that there is no other solution that leads to a smaller objective value. Suppose that there exists an index set $\mathcal{S} \subseteq \mathcal{N}_c$ such that (some of) the slabs indexed by $i \in \mathcal{S} \bigcup \mathcal{N}$ constitute a PARAFAC model whose loading matrices do not contain $\mathbf{B}\mathbf{\Pi}\mathbf{\Delta}_b$ or $\mathbf{C}\mathbf{\Pi}\mathbf{\Delta}_c$. We show that $|\mathcal{S}| < c$. In fact, if $|\mathcal{S}| \geq c$, then, with probability one, the slabs that belong to $\mathcal{S}$ can only be decomposed using $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{A}(\mathcal{S},:)$ with a common column permutation and scaling. The reason is as follows. By the assumption that $\mathbf{A}$ is drawn from some absolutely continuous distribution, we see that $k_{\mathbf{A}(\mathcal{S},:)} = \min\{|\mathcal{S}|, R\} \geq \min\{c, R\} = c$ holds with probability one, and thus $k_{\mathbf{A}(\mathcal{S},:)} + \min\{J, R\} + \min\{K, R\} \geq 2R + 2$ holds almost surely. Hence, by the

TABLE III
MINING THE ENRON E-MAIL CORPUS USING THE PROPOSED ALGORITHM

cluster 1 (Legal; 16 persons)	cluster 2 (Excecutive; 18 persons)	cluster 3 (Executive; 25 persons)
Brenda Whitehead, N/A	David Delainey, CEO ENA and Enron Energy Services	Andy Zipper , VP Enron Online
Dan Hyvl, N/A	Drew Fossum, VP Transwestern Pipeline Company (ETS)	Jeffrey Shankman, President Enron Global Markets
Debra Perlingiere, Legal Specialist ENA Legal	Elizabeth Sager, VP and Asst Legal Counsel ENA Legal	Barry Tycholiz, VP Marketing
Elizabeth Sager, VP and Asst Legal Counsel ENA Legal	James Steffes, VP Government Affairs	Richard Sanders, VP Enron Wholesale Services
Jeff Hodge, Asst General Counsel ENA Legal	Jeff Dasovich, Employee Government Relationship Executive	James Steffes, VP Government Affairs
Kay Mann, Lawyer	John Lavorato, CEO Enron America	Mark Haedicke, Managing Director ENA Legal
Louise Kitchen, President Enron Online	Kay Mann, Lawyer	Greg Whalley, President
Marie Heard, Senior Legal Specialist ENA Legal	Kevin Presto, VP East Power Trading	Jeff Dasovich, Employee Government Relationship Executive
Mark Haedicke, Managing Director ENA Legal	Margaret Carson, Employee Corporate and Environmental Policy	Jeffery Skilling, CEO
Mark Taylor , Manager Financial Trading Group ENA Legal	Mark Haedicke, Managing Director ENA Legal	Vince Kaminski, Manager Risk Management Head
Richard Sanders, VP Enron Wholesale Services	Philip Allen, VP West Desk Gas Trading	Steven Kean, VP Chief of Staff
Sara Shackleton, Employee ENA Legal	Richard Sanders, VP Enron Wholesale Services	Joannie Williamson, Executive Assistant
Stacy Dickson, Employee ENA Legal	Richard Shapiro , VP Regulatory Affairs	John Arnold, VP Financial Enron Online
Stephanie Panus, Senior Legal Specialist ENA Legal	Sally Beck, COO	John Lavorato, CEO Enron America
Susan Bailey, Legal Assistant ENA Legal	Shelley Corman, VP Regulatory Affairs	Jonathan McKa, Director Canada Gas Trading
Tana Jones, Employee Financial Trading Group ENA Legal	Steven Kean, VP Chief of Staff	Kenneth Lay, CEO
	Susan Scott, Employee Transwestern Pipeline Company (ETS)	Liz Taylor, Executive Assistant to Greg Whalley
	Vince Kaminski, Manager Risk Management Head	Louise Kitchen, President Enron Online
cluser 4 (Trading; 12 persons)	cluster 5 (Pipeline; 15 persons)	Michelle Cash, N/A
Chris Dorland, Manager	Bill Rapp, N/A	Mike McConnel, Executive VP Global Markets
Eric Bas, Trader Texas Desk Gas Trading	Darrell Schoolcraft, Employee Gas Control (ETS)	Kevin Presto, VP East Power Trading
Philip Allen, Manager	Drew Fossum, VP Transwestern Pipeline Company (ETS)	Richard Shapiro, VP Regulatory Affairs
Kam Keiser, Employee Gas	Kevin Hyatt, Director Asset Development TW Pipeline Business (ETS)	Rick Buy, Manager Chief Risk Management Officer
Mark Whitt, Director Marketing	Kimberly Watson, Employee Transwestern Pipeline Company (ETS)	Sally Beck, COO
Martin Cuilla, Manager Central Desk Gas Trading	Lindy Donoho, Employee Transwestern Pipeline Company (ETS)	Hunter Shively, VP Central Desk Gas Trading
Matthew Lenhart, Analyst West Desk Gas Trading	Lynn Blair, Employee Northern Natural Gas Pipeline (ETS)	
Michael Grigsby, Director West Desk Gas Trading	Mark McConnell, Employee Transwestern Pipeline Company (ETS)	
Monique Sanchez, Associate West Desk Gas Trader (EWS)	Michelle Lokay, Admin. Asst. Transwestern Pipeline Company (ETS)	
Susan Scott, Employee Transwestern Pipeline Company (ETS)	Rod Hayslett, VP Also CFO and Treasurer	
Jane Tholt, VP West Desk Gas Trading	Shelley Corman, VP Regulatory Affairs	
Philip Allen, VP West Desk Gas Trading	Stanley Horton, President Enron Gas Pipeline	
	Susan Scott, Employee Transwestern Pipeline Company (ETS)	
	Teb Lokey, Manager Regulatory Affairs	
	Tracy Geaccone, Manager (ETS)	

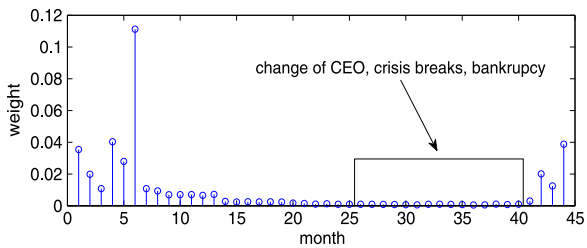


Fig. 9. The normalized weights obtained by the proposed algorithm when applied on the ENRON e-mail data.

uniqueness condition mentioned in (3), the PARAFAC decomposition of $\underline{\mathbf{X}}(\mathcal{S},:,:)$ is essentially unique with probability one. Thus, it can be seen that if the solution to Problem (8) does not satisfy $\mathbf{B}^\star = \mathbf{B}\mathbf{\Pi}\mathbf{\Delta}_b$, and $\mathbf{C}^\star = \mathbf{C}\mathbf{\Pi}\mathbf{\Delta}_c$, we must have

$$v_{\min} \geq I - \left|\mathcal{N}\bigcup\mathcal{S}\right| > I - \frac{I+c}{2} = \frac{I-c}{2},$$

where we have used the fact that $|\mathcal{N}\bigcup\mathcal{S}| < (I+c)/2$.

It remains to show that $\mathbf{A}^\star(\mathcal{N}_c,:) = \mathbf{A}(\mathcal{N}_c,:)\mathbf{\Pi}\mathbf{\Delta}_a$. In fact, given that the optimal solution satisfies $\mathbf{B}^\star = \mathbf{B}\mathbf{\Pi}\mathbf{\Delta}_b$, and $\mathbf{C}^\star = \mathbf{C}\mathbf{\Pi}\mathbf{\Delta}_c$, the optimal $\mathbf{A}^\star$ should be able to make

$$\mathcal{I}\left(\left\|\underline{\mathbf{X}}^{(3)}(:,i) - (\mathbf{C}^\star \odot \mathbf{B}^\star)\mathbf{A}^\star(i,:)^T\right\|_2\right) = 0, \qquad (20)$$

for as many as possible i's. For $i \in \mathcal{N}_c$, we conclude $\mathbf{A}^\star(\mathcal{N}_c,:) = \mathbf{A}(\mathcal{N}_c,:)\mathbf{\Pi}\mathbf{\Delta}_a$. The reason, again, lies in the uniqueness result in (3): Since $|\mathcal{N}_c| \geq (I+c)/2 \geq c$, we have $k_{\mathbf{A}(\mathcal{N}_c,:)} \geq c$ with probability one, since $\mathbf{A}$ is drawn from an absolutely continuous distribution over $\mathbb{R}^{I\times R}$. Hence, $k_{\mathbf{A}(\mathcal{N}_c,:)} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2R+2$ holds with probability one. Consequently, the PARAFAC decomposition of $\underline{\mathbf{X}}(\mathcal{N}_c,:,:)$ is essentially unique. This implies $\mathbf{A}^\star(\mathcal{N}_c,:) = \mathbf{A}(\mathcal{N}_c,:)\mathbf{\Pi}\mathbf{\Delta}_a$.

B. Proof of Claim 2

To relate the stationary points of Problem (13) to the stationary points of Problem (10), let us denote the cost functions of Problem (10) and Problem (13) as $\Psi_1(\mathbf{A},\mathbf{B},\mathbf{C})$ and $\Psi_2(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{W})$, respectively. We see that

$$\Psi_1(\mathbf{A},\mathbf{B},\mathbf{C}) = \min_{w_1,\ldots,w_I\geq 0} \Psi_2\left(\mathbf{A},\mathbf{B},\mathbf{C},\{w_i\}_{i=1}^I\right).$$

Let us consider $(\mathbf{A}^\star,\mathbf{B}^\star,\mathbf{C}^\star,\{w_i^\star\}_{i=1}^I)$ as a stationary point of Problem (13). Following Lemma 1, a direct observation is that

$$\Psi_1(\mathbf{A}^\star,\mathbf{B}^\star,\mathbf{C}^\star) = \Psi_2\left(\mathbf{A}^\star,\mathbf{B}^\star,\mathbf{C}^\star,\{w_i^\star\}_{i=1}^I\right), \qquad (21)$$

since $\Psi_2(\mathbf{A},\mathbf{B},\mathbf{C},\{w_i\}_{i=1}^I)$ has a unique stationary point w.r.t. $\{w_i\}_{i=1}^I$ on the interior of the nonnegative orthant, which is the optimal solution w.r.t. $\{w_i\}_{i=1}^I$. Hence, one can see that $\mathbf{A}^\star,\mathbf{B}^\star,\mathbf{C}^\star$ is also a stationary point of $\Psi_1(\mathbf{A},\mathbf{B},\mathbf{C})$. In fact, taking $\mathbf{A}^\star$ as an example, we see that, following (21),

$$\mathrm{Tr}\left(\nabla_{\mathbf{A}}\Psi_2\left(\mathbf{A}^\star,\mathbf{B}^\star,\mathbf{C}^\star,\{w_i^\star\}_{i=1}^I\right)^T(\mathbf{A}-\mathbf{A}^\star)\right) \geq \mathbf{0}$$
$$\Rightarrow \mathrm{Tr}\left(\nabla_{\mathbf{A}}\Psi_1(\mathbf{A}^\star,\mathbf{B}^\star,\mathbf{C}^\star)^T(\mathbf{A}-\mathbf{A}^\star)\right) \geq \mathbf{0},$$

which implies that $\mathbf{A}$ is also a stationary point of Problem (10). The same proof applies to $\mathbf{B}$ and $\mathbf{C}$.

C. ADMM Updates w.r.t. $\mathbf{C}$ and $\mathbf{A}$

Now, let us consider the update of $\mathbf{C}$:

$$\min_{\mathbf{C}} \quad \frac{1}{2}\left\|(\mathbf{I}\otimes\mathbf{W})\underline{\mathbf{X}}^{(2)} - (\mathbf{B}\odot\mathbf{W}\mathbf{A})\mathbf{C}^T\right\|_F^2 + \lambda_c h(\mathbf{C})$$
$$\text{s.t.} \quad \mathbf{C} \geq \mathbf{0}. \qquad (22)$$

By applying the same structure of ADMM, we come up with

$$\mathbf{C}_1^T := \left(\mathbf{B}^T\mathbf{B} \circledast \mathbf{A}^T\mathbf{W}^2\mathbf{A} + \rho\mathbf{I}\right)^{-1} \times \left(\left(\mathbf{B} \odot \mathbf{W}^2\mathbf{A}\right)^T \underline{\mathbf{X}}^{(2)} + \rho(\mathbf{C} + \mathbf{V}_1)^T\right) \quad (23a)$$

$$\mathbf{C}_2 := \arg\min_{\mathbf{C}_2} \lambda_c h(\mathbf{C}_2) + \frac{\rho}{2}\|\mathbf{C}_1 - \mathbf{C} + \mathbf{V}_2\|_F^2 \quad (23b)$$

$$\mathbf{C} := \left(\frac{1}{2}(\mathbf{C}_2 - \mathbf{V}_2 + \mathbf{C}_1 - \mathbf{V}_1)\right)_{\mathcal{C}} \quad (23c)$$

$$\mathbf{V}_1 := \mathbf{V}_1 + \mathbf{C} - \mathbf{C}_1 \quad (23d)$$

$$\mathbf{V}_2 := \mathbf{V}_2 + \mathbf{C} - \mathbf{C}_2. \quad (23e)$$

The update w.r.t. $\mathbf{A}$ is even simpler:

$$\mathbf{A}_1^T := \left(\mathbf{C}^T\mathbf{C} \circledast \mathbf{B}^T\mathbf{B} + \rho\mathbf{I}\right)^{-1} \times \left((\mathbf{C} \odot \mathbf{B})^T \underline{\mathbf{X}}^{(3)} + \rho(\mathbf{A} + \mathbf{Z}_1)^T\right) \quad (24a)$$

$$\mathbf{A}_2 := \arg\min_{\mathbf{A}_2} \lambda_a f(\mathbf{A}_2) + \frac{\rho}{2}\|\mathbf{A} - \mathbf{A}_1 + \mathbf{Z}_2\|_F^2 \quad (24b)$$

$$\mathbf{A} = \left(\frac{1}{2}(\mathbf{A}_1 - \mathbf{Z}_1 + \mathbf{A}_2 - \mathbf{Z}_2)\right)_{\mathcal{A}} \quad (24c)$$

$$\mathbf{Z}_1 := \mathbf{Z}_1 + \mathbf{A} - \mathbf{A}_1 \quad (24d)$$

$$\mathbf{Z}_2 := \mathbf{Z}_2 + \mathbf{A} - \mathbf{A}_2. \quad (24e)$$

REFERENCES

[1] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 810–823, 2000.

[2] Y. Rong, S. A. Vorobyov, A. B. Gershman, and N. D. Sidiropoulos, "Blind spatial signature estimation via time-varying user power loading and parallel factor analysis," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1697–1710, May 2005.

[3] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 2377–2388, Aug. 2000.

[4] N. D. Sidiropoulos and X.-Q. Liu, "Identifiability results for blind beamforming in incoherent multipath with small delay spread," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 228–236, Jan. 2001.

[5] A. Smilde, R. Bro, and P. Geladi, *Multi-Way Analysis: Applications in the Chemical Sciences*. New York, NY, USA: Wiley, 2005.

[6] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures," *IEEE Audio, Speech, Langu. Process.*, vol. 18, no. 6, pp. 1193–1207, Aug. 2010.

[7] K. Rahbar and J. Reilly, "A frequency domain method for blind source separation of convolutive audio mixtures," *IEEE Speech Audio Process.*, vol. 13, no. 5, pp. 832–844, Sep. 2005.

[8] X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Tensor-based power spectrum separation and emitters localization for cognitive radio," in *Proc. IEEE SAM 2014*, 2014, pp. 421–424.

[9] E. E. Papalexakis, U. Kang, C. Faloutsos, N. D. Sidiropoulos, and A. Harpale, "Large scale tensor decompositions: Algorithmic developments and applications," *IEEE Data Eng. Bull.*, vol. 36, no. 3, pp. 59–66, 2013.

[10] K.-K. Lee, W.-K. Ma, X. Fu, T.-H. Chan, and C.-Y. Chi, "A Khatri-Rao subspace approach to blind identification of mixtures of quasi-stationary sources," *Signal Process.*, vol. 93, no. 12, pp. 3515–3527, 2013.

[11] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, May 2015.

[12] S. Engelen and M. Hubert, "Detecting outlying samples in a parallel factor analysis model," *Analytica Chimica Acta*, vol. 705, no. 1, pp. 155–165, 2011.

[13] M. Hubert, J. Van Kerckhoven, and T. Verdonck, "Robust PARAFAC for incomplete data," *J. Chemometrics*, vol. 26, no. 6, pp. 290–298, 2012.

[14] R. Bro and M. Vidal, "EEMizer: Automated modeling of fluorescence eem data," *Chemometrics Intell. Lab. Syst.*, vol. 106, no. 1, pp. 86–92, 2011.

[15] S. Engelen, S. Frosch, and B. M. Jrgensen, "A fully robust PARAFAC method for analyzing fluorescence data," *J. Chemometrics* vol. 23, no. 3, pp. 124–131, 2009 [Online]. Available: http://dx.doi.org/10.1002/cem.1208

[16] S. A. Vorobyov, Y. Rong, N. D. Sidiropoulos, and A. B. Gershman, "Robust iterative fitting of multilinear models," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2678–2689, Aug. 2005.

[17] D. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

[18] B. Chen, S. He, Z. Li, and S. Zhang, "Maximum block improvement and polynomial optimization," *SIAM J. Optim.*, vol. 22, no. 1, pp. 87–107, 2012.

[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, pp. 1–122, 2011.

[20] J. B. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra Appl.*, vol. 18, pp. 95–138, 1977.

[21] L.-H. Lim and P. Comon, "Blind multilinear identification," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1260–1280, 2014.

[22] L. D. Lathauwer and J. Castaing, "Blind identification of underdetermined mixtures by simultaneous matrix diagonalization," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1096–1105, Mar. 2008.

[23] D. Nion and L. Lathauwer, "An enhanced line search scheme for complex-valued tensor decompositions. Application in DS-CDMA," *Signal Process.*, vol. 88, no. 3, pp. 749–755, 2008.

[24] P. Tichavský and Z. Koldovský, "Weight adjusted tensor method for blind separation of underdetermined mixtures of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1037–1047, Mar. 2011.

[25] A. Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1545–1553, Jul. 2002.

[26] J. Diesner, T. L. Frantz, and K. M. Carley, "Communication networks from the Enron email corpus "it's always about the people. enron is no different"," *Comput. Math. Org. Theory*, vol. 11, no. 3, pp. 201–228, 2005.

[27] E. E. Papalexakis, N. D. Sidiropoulos, and R. Bro, "From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 493–506, 2013.

[28] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, no. 3, 2008 [Online]. Available: http://stacks.iop.org/0266-5611/24/i=3/a=035020

[29] B. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, Jan. 1999.

[30] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. ICASSP*, April 4, 2008, pp. 3869–3872.

[31] R. Chartrand, "Nonconvex compressed sensing and error correction," in *Proc. ICASSP*, April 2007, vol. 3, pp. 889–892.

[32] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, 1992.

[33] J. Idier, "Convex half-quadratic criteria and interacting auxiliary variables for image restoration," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1001–1009, 2001.

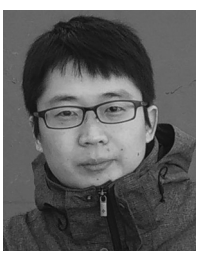
[34] B. Bader and T. Kolda, "Efficient MATLAB computations with sparse and factored tensors," *SIAM J. Sci. Comput.* vol. 30, no. 1, pp. 205–231, 2008 [Online]. Available: http://dx.doi.org/10.1137/060676489

[35] T. Kolda and J. Sun, "Scalable tensor decompositions for multi-aspect data mining," in *Proc. IEEE ICDM 2008*, pp. 363–372.

[36] B. W. Bader and T. G. Kolda *et al.*, Matlab Tensor Toolbox Version 2.5 January 2012 [Online]. Available: http://www.sandia.gov/~tgkolda/TensorToolbox/

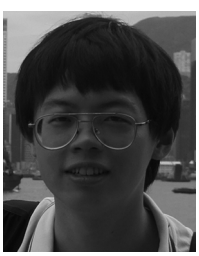
[37] N. Ravindran, N. D. Sidiropoulos, S. Smith, and G. Karypis, "Memory-efficient parallel computation of tensor and matrix products for big tensor decomposition," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Nov. 3–5, 2014.

[38] N. Li, S. Kindermann, and C. Navasca, "Some convergence results on the regularized alternating least-squares method for tensor decomposition," *Linear Algebra Appl.*, vol. 438, no. 2, pp. 796–812, 2013.
[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Press, 2004.
[40] C. Navasca, L. D. Lathauwer, and S. Kindermann, "Swamp reducing technique for tensor decomposition," in *Proc. EUSIPCO 2008*, 2008.
[41] R. Chartrand, "Shrinkage mappings and their induced penalty functions," in *Proc. ICASSP*, 2014, pp. 1026–1029.
[42] A.-J. van der Veen, "Constant modulus beamforming," in *Robust Adaptive Beamforming*. New York, NY, USA: Wiley, 2006, p. 299.
[43] A.-J. van der Veen, "Joint diagonalization via subspace fitting techniques," in *Proc. ICASSP*, May 2001, vol. 5, pp. 2773–2776.
[44] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *Proc. 31st Int. Con. Mach. Learn. (ICML)*, 2014, pp. 1062–1070.
[45] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2496–2504.
[46] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, Apr. 1979.
[47] M. Bahram, R. Bro, C. Stedmon, and A. Afkhami, "Handling of rayleigh and raman scatter for PARAFAC modeling of fluorescence data using interpolation," *J. Chemometrics*, vol. 20, no. 3–4, pp. 99–105, 2006.
[48] C. A. Andersson and R. Bro, "The n-way toolbox for matlab," *Chemometrics Intel. Lab. Syst.*, vol. 52, no. 1, pp. 1–4, 2000.
[49] B. W. Bader, R. A. Harshman, and T. G. Kolda, "Temporal analysis of social networks using three-way dedicom," Sandia National Labs., TR SAND2006-2161, 2006, vol. 119.



Xiao Fu (S'12–M'15) received his B.Eng and M.Eng degrees in communication and information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2005 and 2010, respectively. In 2014, he received his Ph.D. degree in electronic engineering from the Chinese University of Hong Kong (CUHK), Hong Kong. From 2005 to 2006, he was an assistant engineer at China Telecom Co. Ltd., Shenzhen, China. He is currently a Postdoctoral Associate at the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, United States. His research interests include signal processing and machine learning, with a recent emphasis on factor analysis and its applications.

Dr. Fu was an awardee of the Oversea Research Attachment Programme (ORAP) 2013 of the Engineering Faculty, CUHK, which sponsored his visit to the Department of Electrical and Computer Engineering, University of Minnesota, from September 2013 to February 2014. He received a Best Student Paper Award at ICASSP 2014.



Kejun Huang (S'13) received the B.Eng. in Communication Engineering from Nanjing University of Information Science and Technology, Nanjing, China in 2010. Since September 2010, he has been working towards his Ph.D. degree in the Department of Electrical and Computer Engineering, University of Minnesota. His research interests include signal processing, machine learning, and data analytics. His current research focuses on identifiability, algorithms, and performance analysis for factor analysis of big matrix and tensor data.

Wing-Kin Ma (M'01–SM'11) received the B.Eng. degree in electrical and electronic engineering from the University of Portsmouth, Portsmouth, U.K., in 1995, and the M.Phil. and Ph.D. degrees, both in electronic engineering, from The Chinese University of Hong Kong (CUHK), Hong Kong, in 1997 and 2001, respectively. He is currently an Associate Professor with the Department of Electronic Engineering, CUHK. From 2005 to 2007, he was also an Assistant Professor with the Institute of Communications Engineering, National Tsing Hua University, Taiwan, R.O.C. Prior to becoming a faculty member, he held various research positions with McMaster University, Canada; CUHK; and the University of Melbourne, Australia. His research interests are in signal processing and communications, with a recent emphasis on optimization, MIMO transceiver designs and interference management, blind signal processing theory, methods and applications, and hyperspectral unmixing in remote sensing.

Dr. Ma is currently serving or has served as Associate Editor and Guest Editor of several journals, which include IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE SIGNAL PROCESSING LETTERS, SIGNAL PROCESSING, IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS and IEEE SIGNAL PROCESSING MAGAZINE. He was a tutorial speaker in EUSIPCO 2011 and ICASSP 2014. He is currently a Member of the Signal Processing Theory and Methods Technical Committee (SPTM-TC) and the Signal Processing for Communications and Networking Technical Committee (SPCOM-TC). Dr. Ma's students have won ICASSP Best Student Paper Awards in 2011 and 2014, respectively, and he is co-recipient of a WHISPERS 2011 Best Paper Award. He received Research Excellence Award 2013–2014 by CUHK.

Nicholas D. Sidiropoulos (F'09) received the Diploma in Electrical Engineering from the Aristotelian University of Thessaloniki, Greece, and M.S. and Ph.D. degrees in Electrical Engineering from the University of Maryland-College Park, in 1988, 1990 and 1992, respectively. He served as Assistant Professor at the University of Virginia (1997–1999); Associate Professor at the University of Minnesota-Minneapolis (2000–2002); Professor at the Technical University of Crete, Greece (2002–2011); and Professor at the University of Minnesota-Minneapolis (2011-). His current research focuses primarily on signal and tensor analytics, with applications in cognitive radio, big data, and preference measurement. He received the NSF/CAREER award (1998), the IEEE Signal Processing Society (SPS) Best Paper Award (2001, 2007, 2011), and the IEEE SPS Meritorious Service Award (2010). He has served as IEEE SPS Distinguished Lecturer (2008–2009), and Chair of the IEEE Signal Processing for Communications and Networking Technical Committee (2007–2008). He received the Distinguished Alumni Award of the Department of Electrical and Computer Engineering, University of Maryland, College Park in 2013, and was elected EURASIP Fellow in 2014.

Rasmus Bro received the M.Sc. degree in mathematics and analytical chemistry from the Technical University of Denmark, Kgs. Lyngby, Denmark, 1139 in 1994 and the Ph.D. degree (cum laude) in multiway analysis from the University of Amsterdam, 1141 Amsterdam, The Netherlands, in 1998. Since 1994, he has been with the Department of Food Science, Quality and Technology, University of Copenhagen, Copenhagen, Denmark (former Royal Veterinary & Agricultural University). In 2002, he was appointed Full Professor of Chemometrics. He has had several stays abroad at research institutions in The Netherlands, Norway, France, and the United States. He has authored more than 100 peer-reviewed scientific papers, two books on chemometrics, and more than 20 proceedings papers, book contributions, reviews, and patents. His current research interests include chemometrics, multivariate calibration, multiway analysis, exploratory analysis, experimental design, numerical analysis, blind source separation, curve resolution, MATLAB programming, and constrained regression. Dr. Bro received the third Elsevier Chemometrics Award for noteworthy accomplishments in the field of chemometrics by younger scientists, in 2000 and the Eastern Analytical Symposium Award for Achievements in Chemometrics in 2004.