Block-Term Tensor Decomposition: Model Selection and Computation

Athanasios A. Rontogiannis
IAASARS, National Observatory of Athens
152 36 Penteli, Greece
tronto@noa.gr

Eleftherios Kofidis*
Dept. of Statistics and Insurance Science
University of Piraeus
185 34 Piraeus, Greece
kofidis@unipi.gr

Paris V. Giampouras
Mathematical Inst. for Data Science
Johns Hopkins University
Baltimore, MD 21218, USA
parisg@jhu.edu

***Abstract*—The so-called block-term decomposition (BTD) tensor model has been recently receiving increasing attention due to its enhanced ability of representing systems and signals that are composed of *blocks* of rank higher than one, a scenario encountered in numerous diverse applications. Its uniqueness and approximation have thus been thoroughly studied. Nevertheless, the problem of estimating the BTD model structure, namely the number of block terms and their individual ranks, has only recently started to attract significant attention, as it is more challenging compared to more classical tensor models such as canonical polyadic decomposition (CPD) and Tucker decomposition (TD). This paper reports our recent results on this topic, which are based on an appropriate extension to the BTD model of our earlier rank-revealing work on low-rank matrix approximation. The idea is to impose column sparsity *jointly* on the factors and successively estimate the ranks as the numbers of factor columns of non-negligible magnitude, with the aid of alternating iteratively reweighted least squares (IRLS). Simulation results are reported that demonstrate the effectiveness of our method in accurately estimating both the ranks and the factors of the least squares BTD approximation.**

***Index Terms*—Alternating least squares (ALS), block coordinate descent (BCD), block successive upper bound minimization (BSUM), block-term tensor decomposition (BTD), iterative reweighted least squares (IRLS), rank, tensor**

I. Introduction

Block Term Decomposition (BTD) was introduced in [1] as a tensor model combining the Canonical Polyadic Decomposition (CPD) and the Tucker decomposition (TD), in the sense that it decomposes a tensor in a sum of tensors that have low multilinear rank (instead of rank one as in CPD). Hence a BTD can be seen as a *constrained* TD, with its core tensor being block diagonal [2]. BTD can also be seen as a *constrained* CPD having factors with (some) collinear columns [1]. In a way, it lies between the two extremes (in terms of core tensor structure), CPD and TD, and it is interesting to recall the related remark made in [1], namely that "“the” rank of a higher-order tensor is actually a combination of the two aspects: one should specify the number of blocks *and* their size." Accurately and efficiently estimating these numbers for a given tensor is the main subject of this work.

Although [1] introduced BTD as a sum of R rank-(L_r, M_r, N_r) terms ($r = 1, 2, \ldots, R$) in general, the special case of rank-$(L_r, L_r, 1)$ BTD has attracted a lot more of attention, because of both its more frequent occurrence in applications and the existence of more concrete and easier to check uniqueness conditions. This paper will also focus on this special yet much more popular BTD model. Consider a 3rd-order tensor, $\boldsymbol{\mathcal{X}} \in \mathbb{C}^{I\times J\times K}$. Then its rank-$(L_r, L_r, 1)$ decomposition is written as

$$\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \mathbf{E}_r \circ \mathbf{c}_r, \qquad (1)$$

where $\mathbf{E}_r$ is an $I \times J$ matrix of rank L_r, $\mathbf{c}_r$ is a nonzero column K-vector and $\circ$ denotes outer product. Clearly, $\mathbf{E}_r$ can be written as a matrix product $\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}}$ with the matrices $\mathbf{A}_r \in \mathbb{C}^{I\times L_r}$ and $\mathbf{B}_r \in \mathbb{C}^{J\times L_r}$ being of full column rank, L_r.

The most popular uniqueness theorem for rank-$(L_r, L_r, 1)$ BTD states that it is *sufficient* for the partitioned matrices $\mathbf{A} \triangleq \left[\begin{array}{cccc} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_R \end{array}\right]$ and $\mathbf{B} \triangleq \left[\begin{array}{cccc} \mathbf{B}_1 & \mathbf{B}_2 & \cdots & \mathbf{B}_R \end{array}\right]$ to be of full column rank and $\mathbf{C} \triangleq \left[\begin{array}{cccc} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_R \end{array}\right]$ not have any collinear columns [1, Theorem 4.1]. The generic version of the requirement for full column rank of $\mathbf{A}, \mathbf{B}$ is that $\min(I, J) \geq \sum_{r=1}^{R} L_r$, which can be easily met in applications where R and L_r are small. It should however be noted that this is not a necessary condition as our simulation results also demonstrate.

Alternating least squares (ALS) was extended to the approximation of a tensor BTD in [3]. Additional methods include non-linear least squares [4], ALS regularized through proximal point modifications [5], deflation-based [6], variable projection using Riemannian gradient for rank-(L_r, M_r, N_r) BTD with factors of orthonormal columns [7], and solving the equivalent matrix factorization problem with one of the factors constrained to have low-rank rows [8].

In all BTD methods mentioned above, R and L_r, $r = 1, 2, \ldots, R$ are assumed known[1] (and it is commonly assumed that all L_r are all equal to L, for simplicity). In fact, in

P. V. Giampouras is supported by the European Union under the Horizon 2020 Marie-Skłodowska-Curie Global Fellowship program: HyPPOCRATES — H2020-MSCA-IF-2018, Grant Agreement Number: 844290.

*This work has been partly supported by the University of Piraeus Research Center.

[1] In the non-iterative method of [9], and in the (almost) noise-free case, these are estimated (with the aid of SVDs) from a joint block matrix diagonalization problem. For noisier tensors, R and $\sum_r L_r$ are assumed known.

practice, this is a challenging question on its own. Unless external information is given (such as in a telecommunications [10] or in a hyperspectral imagery (HSI) unmixing application with given or estimated ground truth [11]), there is no way to know these values a priori.

A. Background

Model order selection techniques for BTD can be dictated by corresponding CPD techniques, as reviewed in [12, Section 4]. Schemes of multilinear rank estimation (largely based on matrix rank estimation and/or extensions of one-dimensional information-theoretic criteria) are also relevant in view of the constrained TD structure of BTD [13]–[16].

Model order selection can also be application-specific. For example, L_r's are estimated in [17] as the auto-regressive (AR) orders of the sources in ECG analysis, with R assumed known. In [18], and in the functional magnetic resonance imaging (fMRI) context, L_r is estimated as the number of statistically significant (bearing useful information) columns of $\mathbf{A}_r, \mathbf{B}_r$. [19] relies on the subspace-based method of [20] for estimating the number R of spectral signatures in BTD-based HSI denoising.

Alternative techniques rely on sparsity arguments for model selection. A greedy scheme, inspired from a sparse coding interpretation of the problem, is proposed in [21], for more general tensor decompositions [22] including BTD as a special case. Instead of building the model incrementally, however, one can follow the reverse way of starting from a rank *over-estimate* and arrive at the true rank(s) by eliminating negligible components, aided in this task by appropriate regularization. Such an approach is followed in [23], where the constrained CPD formulation of BTD is taken advantage of to first estimate R and then L_r's assumed all equal, before computing the model factors in (1). In each case, a regularization term is added to the tensor approximation cost, which is composed of mixed norms of the factor matrices and serves as upper bound on the tensor nuclear norm thus promoting column sparsity of the factors and hence low rankness. The augmented Lagrangian method is adopted for the computations.

Nevertheless, as demonstrated in [24], [25] for the CPD case, the problems of model rank estimation and approximation of factors can be addressed *jointly*, with significant gains in both accuracy and complexity (of particular interest in big data applications). This idea is proposed in [26] for the rank-$(L_r, L_r, 1)$ BTD model with not necessarily all equal block-term ranks L_r. A regularization term consisting of the sum of the mixed $\ell_{1,2}$ norms of the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ is added to the squared error of the tensor approximation, which helps promoting low rankness for the BTD factors and hence estimating R (as the number of non-zero columns of $\mathbf{C}$) and L_r's (as the number of non-zero columns of the rth blocks of $\mathbf{A}, \mathbf{B}$ that correspond to non-zero columns of $\mathbf{C}$). A block coordinate descent (BCD) solution approach is taken in [26], resembling a regularized version of the ALS procedure of [3].

The approach we propose in this paper also falls in the previous category. Yet, it has a number of very important new features, inherited from our earlier work on factorization-based low-rank approximation of matrices [27], from which it draws inspiration. In [27], the sum of reweighted Frobenius norms of the factors of the data matrix is used as regularization and, in particular, a diagonal weighting, jointly depending on the factors, is proposed, naturally leading to an iteratively reweighted least squares (IRLS) [28] solution approach, with fast convergence and low complexity. Here we generalize that idea in the BTD problem. The regularization of [27] is employed, in two levels: first, combining the reweighted norms of $\mathbf{A}$ and $\mathbf{B}$, and second, coupling these with the reweighted norm of $\mathbf{C}$. This two-level coupling naturally matches the structure of the model in (1), making explicit the different roles of $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, in contrast to previous related works [23], [26] that miss to exploit this relation. Furthermore, compared with previous works, the regularization proposed here has a stronger sparsity promoting action due to the properties of the norms adopted. Applying majorization with appropriate upper bounds and a BCD approach results in an alternating IRLS algorithm that manages to both reveal the ranks and compute the BTD factors at a high convergence rate and low computational cost. Notably, iterations involve updates that contain only matrix-matrix multiplications, which are optimally implemented on most modern computer systems (such as GPUs) and can be easily parallelized. The complexity can be reduced even more by eliminating negligible columns (column pruning) in the course of the iterations (although this option is not considered here). Simulation results are reported that demonstrate the effectiveness of the proposed method (in comparison with the classical BTD-ALS method [3]) in estimating both the model structure and its parameters.

B. Notation

Lower- and upper-case bold letters are used to denote vectors and matrices, respectively. Higher-order tensors are denoted by upper-case bold calligraphic letters. For a tensor $\boldsymbol{\mathcal{X}}$, $\mathbf{X}_{(n)}$ stands for its mode-n unfolding. $\otimes$ stands for the Kronecker product. The Khatri-Rao product is denoted by $\odot$ in its general (partition-wise) version and by $\odot_{\mathrm{c}}$ in its column-wise version. $\circ$ denotes the outer product. The superscript $^{\mathrm{T}}$ stands for transposition. The identity matrix of order N and the all ones column N-vector are respectively denoted by $\mathbf{I}_N$ and $\mathbf{1}_N$. The Euclidean, mixed 1, 2 and Frobenius norms are denoted by $\|\cdot\|_2$, $\|\cdot\|_{1,2}$, and $\|\cdot\|_{\mathrm{F}}$, respectively. $\mathbb{C}$ is the field of complex numbers.

II. Problem Statement

Given an $I \times J \times K$ tensor $\boldsymbol{\mathcal{Y}}$, its best (in the least squares sense) rank-$(L_r, L_r, 1)$ approximation is sought for, namely

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} f(\mathbf{A}, \mathbf{B}, \mathbf{C}) \triangleq \frac{1}{2} \left\| \boldsymbol{\mathcal{Y}} - \sum_{r=1}^{R} \mathbf{A}_r \mathbf{B}_r^{\mathrm{T}} \circ \mathbf{c}_r \right\|_{\mathrm{F}}^2, \quad (2)$$

where the matrices $\mathbf{A}_r = \begin{bmatrix} \mathbf{a}_{r1} & \mathbf{a}_{r2} & \cdots & \mathbf{a}_{rL_r} \end{bmatrix} \in \mathbb{C}^{I \times L_r}$, $\mathbf{B}_r = \begin{bmatrix} \mathbf{b}_{r1} & \mathbf{b}_{r2} & \cdots & \mathbf{b}_{rL_r} \end{bmatrix} \in \mathbb{C}^{J \times L_r}$, $\mathbf{C} \in \mathbb{C}^{K \times R}$, and the ranks R and L_r, $r = 1, 2, \ldots, R$ are unknown. In terms of its mode unfoldings, the tensor $\boldsymbol{\mathcal{X}} \triangleq \sum_{r=1}^{R} \mathbf{A}_r\mathbf{B}_r^{\mathrm{T}} \circ \mathbf{c}_r$ can be written as [1]

$$\mathbf{X}_{(1)}^{\mathrm{T}} = (\mathbf{B} \odot \mathbf{C})\mathbf{A}^{\mathrm{T}}, \tag{3}$$

$$\mathbf{X}_{(2)}^{\mathrm{T}} = (\mathbf{C} \odot \mathbf{A})\mathbf{B}^{\mathrm{T}}, \tag{4}$$

$$\mathbf{X}_{(3)}^{\mathrm{T}} = \begin{bmatrix} (\mathbf{A}_1 \odot_{\mathrm{c}} \mathbf{B}_1)\mathbf{1}_{L_1} & \cdots & (\mathbf{A}_R \odot_{\mathrm{c}} \mathbf{B}_R)\mathbf{1}_{L_R} \end{bmatrix} \mathbf{C}^{\mathrm{T}} \tag{5}$$

These expressions can be used in alternatingly solving for $\mathbf{A}, \mathbf{B}, \mathbf{C}$, respectively.

We propose to consider the following modified problem

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} f(\mathbf{A},\mathbf{B},\mathbf{C}) + \lambda\|\mathbf{F}(\mathbf{A},\mathbf{B},\mathbf{C})\|_{1,2}, \tag{6}$$

in which regularization is done with the $\ell_{1,2}$ norm of the $2 \times R$ matrix $\mathbf{F}(\mathbf{A},\mathbf{B},\mathbf{C})$, constructed as follows. Let $\mathbf{G} \triangleq \begin{bmatrix} \mathbf{A}^{\mathrm{T}} & \mathbf{B}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ be the $(I+J) \times \sum_{r=1}^{R} L_r$ matrix resulting from stacking the factors $\mathbf{A}$ and $\mathbf{B}$ and $\mathbf{G}_r \triangleq \begin{bmatrix} \mathbf{A}_r^{\mathrm{T}} & \mathbf{B}_r^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ denote its rth $(I+J) \times L_r$ block. The matrix $\mathbf{F}(\mathbf{A},\mathbf{B},\mathbf{C})$ is then defined as

$$\mathbf{F}(\mathbf{A},\mathbf{B},\mathbf{C}) \triangleq \begin{bmatrix} \|\mathbf{G}_1\|_{1,2} & \|\mathbf{G}_2\|_{1,2} & \cdots & \|\mathbf{G}_R\|_{1,2} \\ \|\mathbf{c}_1\|_2 & \|\mathbf{c}_2\|_2 & \cdots & \|\mathbf{c}_R\|_2 \end{bmatrix}.$$

The minimization of the $\ell_{1,2}$ norm of a vector or matrix subject to a data proximity criterion has been widely utilized in the literature for enforcing group sparsity in vector/matrix recovery problems [29]. This property of the $\ell_{1,2}$ norm was exploited in our earlier work [27] for model order selection in low-rank matrix factorization applications. In the present work, we extend that idea to the BTD problem by employing a two-level hierarchical $\ell_{1,2}$ norm-based regularization scheme. At the upper level, the $\ell_{1,2}$ norm of the matrix $\mathbf{F}(\mathbf{A},\mathbf{B},\mathbf{C})$ above promotes the elimination of whole blocks of $\mathbf{A}$ and $\mathbf{B}$ (which are tied together by the mixed norms $\|\mathbf{G}_r\|_{1,2}$, $r = 1, 2, \ldots, R$) and the corresponding columns of $\mathbf{C}$. At the lower level, the $\ell_{1,2}$ norms $\|\mathbf{G}_r\|_{1,2}$ induce column sparsity to the "surviving" blocks of $\mathbf{A}, \mathbf{B}$. Hence, we have the flexibility to overestimate the ranks R and L_r, $r = 1, 2, \ldots, R$ as $R = R_{\text{ini}}$ and $L_r = L_{\text{ini}}$ in the unknown BTD model, since this regularization can reduce them towards their actual values with a proper selection of the regularization parameter λ. The problem in (6) can be solved with an alternating IRLS algorithm, as described in the next section.

III. Proposed Method

We write the problem in (6) more explicitly in terms of the BTD factors $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ as

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \frac{1}{2}\left\|\boldsymbol{\mathcal{Y}} - \sum_{r=1}^{R} \mathbf{A}_r\mathbf{B}_r^{\mathrm{T}} \circ \mathbf{c}_r\right\|_{\mathrm{F}}^2 + \tag{7}$$

$$\lambda \sum_{r=1}^{R} \sqrt{\left(\sum_{l=1}^{L} \sqrt{\|\mathbf{a}_{rl}\|_2^2 + \|\mathbf{b}_{rl}\|_2^2 + \eta^2}\right)^2 + \|\mathbf{c}_r\|_2^2 + \eta^2},$$

where a very small positive constant η^2 was added to ensure smoothness and R and L here stand for the initial (over)estimates of the model rank parameters. It can be shown that the objective function is convex with respect to (w.r.t.) each one of the factors $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ separately but not w.r.t. all of them. Moreover, due to the regularization term, it is non-separable w.r.t to each one of the matrix factors. Capitalizing on the ideas appearing in [30], we curb the problem in (7) by suitably extending the well-known IRLS method. Note that, in contrast to the classical IRLS, the proposed regularization term hints at the use of two separate reweighting least squares steps. Each step gives rise to a distinct reweighting matrix. Namely, the first one is composed of the inverses of the outer summation terms of the regularizer in (7). This matrix weights jointly the blocks of $\mathbf{A}, \mathbf{B}$, i.e., $\mathbf{A}_r$s, $\mathbf{B}_r$s, and the respective columns of $\mathbf{C}$. The second reweighting matrix contains the inverses of the terms of the inner summation in (7) and jointly balances the corresponding columns of $\mathbf{A}_r$s and $\mathbf{B}_r$s.

The BTD factors are estimated in an iterative alternating fashion. At first, in order to estimate $\mathbf{A}^{k+1}$ at iteration $k+1$, we formulate the following minimization problem

$$\mathbf{A}^{k+1} = \arg\min_{\mathbf{A}} \frac{1}{2}\left\|\mathbf{Y}_{(1)}^{\mathrm{T}} - \mathbf{P}^k\mathbf{A}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \tag{8}$$

$$\frac{\lambda}{2}\sum_{r=1}^{R} \frac{\left(\frac{1}{2}\sum_{l=1}^{L} \frac{\|\mathbf{a}_{rl}\|_2^2 + \|\mathbf{b}_{rl}^k\|_2^2 + \eta^2}{\sqrt{\|\mathbf{a}_{rl}^k\|_2^2 + \|\mathbf{b}_{rl}^k\|_2^2 + \eta^2}}\right)^2 + \|\mathbf{c}_r^k\|_2^2 + \eta^2}{\sqrt{\left(\sum_{l=1}^{L} \sqrt{\|\mathbf{a}_{rl}^k\|_2^2 + \|\mathbf{b}_{rl}^k\|_2^2 + \eta^2}\right)^2 + \|\mathbf{c}_r^k\|_2^2 + \eta^2}},$$

where $\mathbf{P}^k \triangleq \mathbf{B}^k \odot \mathbf{C}^k$. Note that the above objective function is obtained by generalizing the classical IRLS objective to the two-step reweighting form described above. Setting its derivative w.r.t. $\mathbf{A}$ to zero we get the unique solution (of ridge regression form) given by

$$\mathbf{A}^{k+1} = \mathbf{Y}_{(1)}\mathbf{P}^k(\mathbf{P}^{k\mathrm{T}}\mathbf{P}^k + \lambda\mathbf{D}^k)^{-1}, \tag{9}$$

where $\mathbf{D}^k \triangleq (\mathbf{D}_1^k \otimes \mathbf{I}_L)\mathbf{D}_2^k$. $\mathbf{D}_1^k$ is an $R \times R$ diagonal matrix, whose rth diagonal entry is

$$\mathbf{D}_1^k(r,r) = \left[\left(\sum_{l=1}^{L} \sqrt{\|\mathbf{a}_{rl}^k\|_2^2 + \|\mathbf{b}_{rl}^k\|_2^2 + \eta^2}\right)^2 + \|\mathbf{c}_r^k\|_2^2 + \eta^2\right]^{-1/2} \tag{10}$$

and $\mathbf{D}_2^k$ is an $RL \times RL$ diagonal matrix, whose $((r-1)L+l)$th diagonal entry is

$$\mathbf{D}_2^k((r-1)L+l, (r-1)L+l) = (\|\mathbf{a}_{rl}^k\|_2^2 + \|\mathbf{b}_{rl}^k\|_2^2 + \eta^2)^{-1/2}. \tag{11}$$

It is clear from (10) and (11) that the diagonal matrix $\mathbf{D}_1^k \otimes \mathbf{I}_L$ realizes the first reweighting step mentioned previously and the diagonal matrix $\mathbf{D}_2^k$ the second one.

In an analogous manner, $\mathbf{B}^{k+1}$ can be obtained from the solution of the minimization problem

$$\mathbf{B}^{k+1} = \arg\min_{\mathbf{B}} \frac{1}{2}\left\|\mathbf{Y}_{(2)}^{\mathrm{T}} - \mathbf{Q}^k\mathbf{B}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \tag{12}$$

$$\frac{\lambda}{2}\sum_{r=1}^{R} \frac{\left(\frac{1}{2}\sum_{l=1}^{L} \frac{\|\mathbf{a}_{rl}^k\|_2^2 + \|\mathbf{b}_{rl}\|^2 + \eta^2}{\sqrt{\|\mathbf{a}_{rl}^k\|_2^2 + \|\mathbf{b}_{rl}^k\|_2^2 + \eta^2}}\right)^2 + \|\mathbf{c}_r^k\|_2^2 + \eta^2}{\sqrt{\left(\sum_{l=1}^{L} \sqrt{\|\mathbf{a}_{rl}^k\|_2^2 + \|\mathbf{b}_{rl}^k\|_2^2 + \eta^2}\right)^2 + \|\mathbf{c}_r^k\|_2^2 + \eta^2}},$$

where (cf. (4)) $\mathbf{Q}^k \triangleq \mathbf{C}^k \odot \mathbf{A}^k$. The unique solution to this problem is now expressed as $\mathbf{B}^{k+1} = \mathbf{Y}_{(2)}\mathbf{Q}^k(\mathbf{Q}^{k\mathrm{T}}\mathbf{Q}^k + \lambda\mathbf{D}^k)^{-1}$. Finally, the factor $\mathbf{C}^{k+1}$ is estimated from

Algorithm 1: BTD-IRLS algorithm

Input: $\boldsymbol{\mathcal{Y}}, \lambda, R_{\rm ini}, L_{\rm ini}$
Initialize: $k=0, \mathbf{A}^0, \mathbf{B}^0, \mathbf{C}^0$
repeat
  Compute $\mathbf{D}_1^k, \mathbf{D}_2^k$ from (10) and (11)
  $\mathbf{D}^k \leftarrow (\mathbf{D}_1^k \otimes \mathbf{I}_L)\mathbf{D}_2^k$
  $\mathbf{P}^k \leftarrow \mathbf{B}^k \odot \mathbf{C}^k$
  $\mathbf{A}^{k+1} \leftarrow \mathbf{Y}_{(1)}\mathbf{P}^k(\mathbf{P}^{k\mathrm{T}}\mathbf{P}^k + \lambda\mathbf{D}^k)^{-1}$
  $\mathbf{Q}^k \leftarrow \mathbf{C}^k \odot \mathbf{A}^{k+1}$
  $\mathbf{B}^{k+1} \leftarrow \mathbf{Y}_{(2)}\mathbf{Q}^k(\mathbf{Q}^{k\mathrm{T}}\mathbf{Q}^k + \lambda\mathbf{D}^k)^{-1}$
  $\mathbf{S}^k \leftarrow \left[\begin{array}{ccc} (\mathbf{A}_1^{k+1} \odot_{\rm c} \mathbf{B}_1^{k+1})\mathbf{1}_L & \cdots & (\mathbf{A}_R^{k+1} \odot_{\rm c} \mathbf{B}_R^{k+1})\mathbf{1}_L \end{array}\right]$
  $\mathbf{C}^{k+1} \leftarrow \mathbf{Y}_{(3)}\mathbf{S}^k(\mathbf{S}^{k\mathrm{T}}\mathbf{S}^k + \lambda\mathbf{D}_1^k)^{-1}$
  $k \leftarrow k+1$
until *convergence*

$$\mathbf{C}^{k+1} = \arg\min_{\mathbf{C}} \frac{1}{2}\left\|\mathbf{Y}_{(3)}^{\mathrm{T}} - \mathbf{S}^k\mathbf{C}^{\mathrm{T}}\right\|_{\rm F}^2 + \frac{\lambda}{2}\sum_{r=1}^{R} \frac{\left(\sum_{l=1}^{L}\sqrt{\|\mathbf{a}_{rl}^k\|_2^2 + \|\mathbf{b}_{rl}^k\|_2^2 + \eta^2}\right)^2 + \|\mathbf{c}_r\|_2^2 + \eta^2}{\sqrt{\left(\sum_{l=1}^{L}\sqrt{\|\mathbf{a}_{rl}^k\|_2^2 + \|\mathbf{b}_{rl}^k\|_2^2 + \eta^2}\right)^2 + \|\mathbf{c}_r^k\|_2^2 + \eta^2}} \quad (13)$$

as $\mathbf{C}^{k+1} = \mathbf{Y}_{(3)}\mathbf{S}^k(\mathbf{S}^{k\mathrm{T}}\mathbf{S}^k + \lambda\mathbf{D}_1^k)^{-1}$, where (cf. (5)) $\mathbf{S}^k \triangleq \left[\begin{array}{ccc} (\mathbf{A}_1^k \odot_{\rm c} \mathbf{B}_1^k)\mathbf{1}_L & \cdots & (\mathbf{A}_R^k \odot_{\rm c} \mathbf{B}_R^k)\mathbf{1}_L \end{array}\right]$, i.e., a single reweighting via $\mathbf{D}_1^k$ is sufficient for estimating factor $\mathbf{C}$, as expected.

Summarizing the above, the steps of the proposed BTD iterative reweighted least squares (BTD-IRLS) algorithm, which alternatingly solves for $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, in that order, are tabulated as Algorithm 1. Note that, if R and L are overestimated, reweighting via $\mathbf{D}_1$ imposes *jointly* block sparsity on $\mathbf{A}$ and $\mathbf{B}$ and column sparsity on $\mathbf{C}$, hence helping in estimating R. In addition, reweighting via $\mathbf{D}_2$ promotes column sparsity *jointly* to the corresponding blocks of $\mathbf{A}$ and $\mathbf{B}$, thus estimating L_r's. This mechanism, combined with an appropriate selection of λ, can reveal the actual value of R and the true block-term ranks L_r's, as empirically shown in the next section.

A notable feature of the proposed algorithm is that it comprises matrix operations only and relatively small-size matrix inversions, which is translated to relatively low computational complexity. Further reduction is possible by eliminating negligible columns (pruning) in the course of the iterations [27].

As pointed out in [31], the conventional IRLS algorithm may also be seen through a block successive upper bound minimization (BSUM) viewpoint. It can be shown that the objective functions of (8), (12), and (13), corresponding to factors $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ respectively, satisfy the conditions of BSUM, thus ensuring convergence of BTD-IRLS to stationary points. Details are omitted due to lack of space.

IV. Numerical Examples

In this section, we report indicative simulation results for evaluating the performance of the proposed BTD-IRLS algorithm. For comparison purposes, the classical BTD-ALS algorithm of [3], which makes no use of any low-rank regularizer, is also tested.

In all experiments, we generate BTD tensors $\boldsymbol{\mathcal{X}}$ contaminated by additive noise, i.e., $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} + \sigma\boldsymbol{\mathcal{N}}$, where $\boldsymbol{\mathcal{N}}$ contains zero-mean, independent and identically distributed (i.i.d) Gaussian entries of unit variance and σ is set so that we get a given signal-to-noise ratio (SNR), with SNR in dB defined as $\mathrm{SNR} = 10\log_{10}(\|\boldsymbol{\mathcal{X}}\|_{\rm F}^2/(\sigma^2\|\boldsymbol{\mathcal{N}}\|_{\rm F}^2))$. The entries of the matrices $\mathbf{A}_r$ and $\mathbf{B}_r$ and the vectors $\mathbf{c}_r$ have been also sampled from i.i.d. zero-mean Gaussian distributions of unit variance. The tensor approximation accuracy is measured with the normalized mean squared error (NMSE) over the blocks, defined as $\mathrm{NMSE}(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}) = \frac{1}{R}\sum_{r=1}^{R}\frac{\|\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}}\circ\mathbf{c}_r - \hat{\mathbf{A}}_r\hat{\mathbf{B}}_r^{\mathrm{T}}\circ\hat{\mathbf{c}}_r\|_{\rm F}^2}{\|\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}}\circ\mathbf{c}_r\|_{\rm F}^2}$, where $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ denote the true and the estimated factors, respectively. For the stopping criterion, we use the relative difference between two consecutive estimates of the reconstruction error. The algorithms stop either when the relative difference becomes less than 10^{-5} or a maximum of 200 iterations is reached. The regularization parameter λ of BTD-IRLS is fine-tuned so that the minimum NMSE is attained. For each realization, both algorithms are randomly initialized for 10 times, and the best run among them, in terms of the NMSE, is kept. We report the medians of the results obtained over 20 independent realizations.

TABLE I
NMSE comparison between BTD-IRLS and BTD-ALS in the presence of noise for different SNR values.

	SNR (dB)			
	5	10	15	20
BTD-IRLS	0.0035	0.0011	0.0003	0.0001
BTD-ALS	0.0374	0.0284	0.0129	0.0100

A. Performance in the presence of noise

First, we test the proposed algorithm for different SNR values. We set $I = 120$, $J = 100$ and $K = 110$. The true R was set to 5 and the L_r's were integer numbers chosen uniformly at random from the set $\{2, 3, 4, 5, 6, 7, 8\}$. Since the ranks are in general unknown, we initialized BTD-IRLS with overestimates of the true ones, namely $R_{\rm ini} = 10$ and $L_{\rm ini} = 10$ for all blocks. For BTD-ALS it was assumed that the true R is known, while all L_r's were overestimated to 10. As shown in Table I, the proposed method clearly outperforms BTD-ALS in terms of the NMSE. This is also verified in Fig. 1, where the evolution of the NMSE w.r.t. the number of iterations is plotted, for SNR=10 dB and a single realization of the experiment. Clearly, BTD-IRLS outperforms BTD-ALS despite the fact that the latter is given the exact knowledge of R. This can be attributed to the way that the ranks are penalized in BTD-IRLS, which helps the algorithm arrive at the true R and L_r's in most of the realizations. This property of BTD-IRLS is further highlighted next.

B. Performance for different values of R and L_r's

With SNR=15 dB, we estimate the success rates in the estimation of R and L_r's for two different scenarios, namely, for a) $\sum_{r=1}^{R} L_r > \min(I, J)$ and b) $\sum_{r=1}^{R} L_r \leq \min(I, J)$. As mentioned earlier, in the former case, the uniqueness condition is no longer met, which makes it more interesting. In both scenarios, $R_{\rm ini} = 12$, and $L_{\rm ini} = 10$ for all terms.

In the first scenario, we test with tensors of size $(16, 16, 10)$, while varying the values of R and L_rs. Again, for each

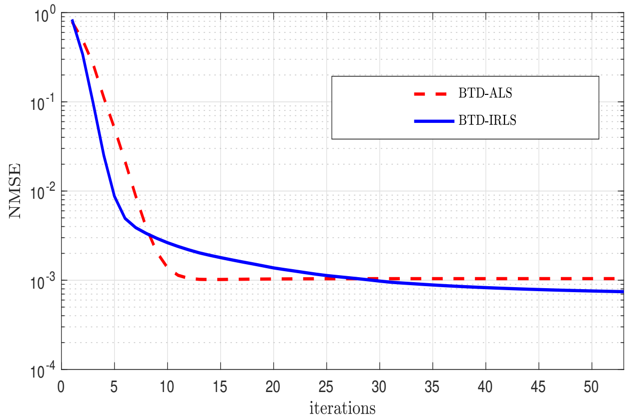

Fig. 1. NMSE of BTD-IRLS and BTD-ALS vs. iterations for SNR=10 dB.

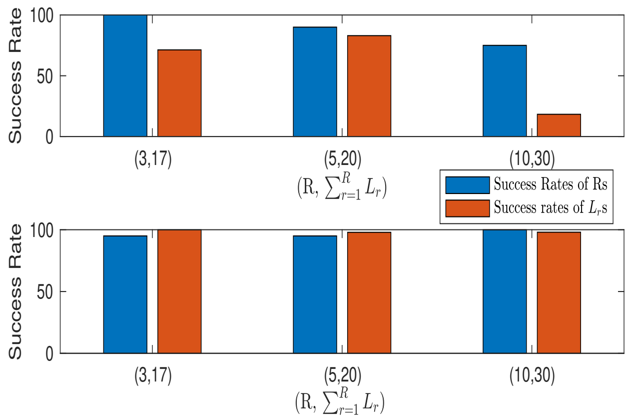

Fig. 2. Success rate (%) of BTD-IRLS in estimating the true R and L_rs. SNR=15 dB. Top: $\sum_{r=1}^{R} L_r > \min(I, J)$, Bottom: $\sum_{r=1}^{R} L_r \leq \min(I, J)$.

combination of the rank values, we consider 20 independent realizations and count the number of times that *both* R and *all* L_rs have been correctly estimated. As shown in Fig. 2 (top), BTD-IRLS achieves high success rates (%) for $R = 3$ and $R = 5$ and exhibits a good performance in estimating R even in the higher-rank case, i.e., for $R = 10$ and $\sum_{r=1}^{R} L_r = 30$. As expected, the success rate decreases with the increase of rank values, since this also increases the model complexity.

Finally, we evaluate the performance of BTD-IRLS for tensors of size $(I, J, K) = (120, 100, 110)$ when the sufficient uniqueness condition is satisfied. As it can be seen in Fig. 2 (bottom), BTD-IRLS achieves high success rates for all combinations of R and L_rs tested in this scenario.

Future work will include more extensive performance evaluation and comparisons with alternative methods as well as the development of constrained variants of the method and (semi-)automatic ways of tuning its regularization parameter.

References

[1] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms — Part II: Definitions and uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1033–1066, 2008.

[2] N. D. Sidiropoulos *et al.*, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, Jul. 2017.

[3] L. De Lathauwer and D. Nion, "Decompositions of a higher-order tensor in block terms — Part III: Alternating least squares algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1067–1083, 2008.

[4] L. Sorber, M. Van Barel, and L. De Lathauwer, "Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-$(L_r, L_r, 1)$ terms, and a new generalization," *SIAM J. Optim.*, vol. 23, no. 2, 2013.

[5] N. Li, "Variants of ALS on tensor decompositions and applications," Ph.D. dissertation, Clarkson University, 2013.

[6] J. H. de M. Goulart and P. Comon, "Non-iterative low-multilinear-rank tensor approximation with application to decomposition in rank-$(1, L, L)$ terms," *hal-01516167*, Apr. 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01516167

[7] G. Olikier, P.-A. Absil, and L. De Lathauwer, "A variable projection method for block term decomposition of higher-order tensors," in *Proc. LVA/ICA-2018*, Guildford, UK, Jul. 2018.

[8] X. Fu and K. Huang, "Block-term tensor decomposition via constrained matrix factorization," in *Proc. MLSP-2019*, Pittsburg, PA, Oct. 2019.

[9] I. Domanov and L. De Lathauwer, "On uniqueness and computation of the decomposition of a tensor into multilinear rank-$(1, L_r, L_r)$ terms," *SIAM J. Matrix Anal. Appl.*, vol. 41, no. 2, pp. 747—-803, 2020.

[10] L. De Lathauwer, "Block component analysis: a new concept for blind source separation," in *Proc. LVA/ICA-2012*, Tel Aviv, Israel, Mar. 2012.

[11] Y. Qian *et al.*, "Matrix-vector nonnegative tensor factorization for blind unmixing of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, Mar. 2017.

[12] B. Hunyadi *et al.*, "Block term decomposition for modelling epileptic seizures," *EURASIP J. Adv. Signal Processing*, no. 139, 2014.

[13] E. E. Papalexakis, "Automatic unsupervised tensor mining with quality assessment," in *Proc. SDM-2016*, Miami, FA, May 2016.

[14] T. Yokota, N. Lee, and A. Cichocki, "Robust multilinear tensor rank estimation using higher order singular value decomposition and information criteria," *IEEE Trans. Signal Process.*, Mar. 2017.

[15] Q. Xie *et al.*, "Kronecker-basis-representation based tensor sparsity and its applications to tensor recovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, Aug. 2018.

[16] X. Han *et al.*, "Robust multilinear decomposition of low rank tensors," in *Proc. LVA/ICA-2018*, Guilford, UK, Jul. 2018.

[17] V. Zarzoso, "Parameter estimation in block term decomposition for non-invasive atrial fibrillation analysis," in *Proc. CAMSAP-2017*, Curaçao, Dutch Antilles, Dec. 2017.

[18] C. Chatzichristos *et al.*, "Blind fMRI source unmixing via higher-order tensor decompositions," *J. Neurosci. Methods*, Mar. 2019.

[19] F. Xiong, J. Zhou, and Y. Qian, "Hyperspectral restoration via L_0 gradient regularized low-rank tensor factorization," *IEEE Trans. Geosci. Remote Sens.*, Dec. 2019.

[20] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. Remote Sens.*, Aug. 2008.

[21] A. J. Brockmeier *et al.*, "Greedy algorithm for model selection of tensor decompositions," in *Proc. ICASSP-2013*, Vancouver, Canada, May 2013.

[22] A. H. Phan *et al.*, "From basis components to complex structural patterns," in *Proc. ICASSP-2013*, Vancouver, Canada, May 2013.

[23] X. Han *et al.*, "Block term decomposition with rank estimation using group sparsity," in *Proc. CAMSAP-2017*, Curaçao, Dutch Antilles, Dec. 2017.

[24] B. Yang, G. Wang, and N. D. Sidiropoulos, "Tensor completion via group-sparse regularization," in *Proc. ACSSC-2016*, Asilomar Conf. Grounds, Pacific Grove, CA, Nov. 2016.

[25] I. C. Tsaknakis *et al.*, "A computationally efficient tensor completion algorithm," *IEEE Signal Process. Lett.*, Aug. 2018.

[26] J. H. de M. Goulart *et al.*, "Alternating group lasso for block-term tensor decomposition and application to ECG source separation," *IEEE Trans. Signal Process.*, Apr. 2020.

[27] P. V. Giampouras, A. A. Rontogiannis, and K. D. Koutroumbas, "Alternating iteratively reweighted least squares minimization for low-rank matrix factorization," *IEEE Trans. Signal Process.*, Jan. 2019.

[28] I. Daubechies *et al.*, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure and Appl. Math.*, vol. 63, pp. 1–38, 2010.

[29] Y. Hu *et al.*, "Group sparse optimization via $\ell_{p,q}$ regularization," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 960–1011, 2017.

[30] A. Beck, "On the convergence of alternating minimization with applications to iteratively reweighted least squares and decomposition schemes," *SIAM J. Optim.*, vol. 25, no. 1, pp. 185–209, 2015.

[31] M. Hong *et al.*, "A unified algorithmic framework for block-structured optimization involving big data," *IEEE Signal Process. Mag.*, Jan. 2016.