A Bayesian Approach to Block-Term Tensor Decomposition Model Selection and Computation

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

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

2021 29th European Signal Processing Conference (EUSIPCO) | 978-9-0827-9706-0/21/$31.00 ©2021IEEE | DOI: 10.23919/EUSIPCO54536.2021.9616222

***Abstract*—The so-called block-term decomposition (BTD) tensor model, especially in its rank-$(L_r, L_r, 1)$ version, 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 and diverse applications. Its uniqueness and approximation have thus been thoroughly studied. Nevertheless, the challenging 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. In this work, a Bayesian approach is taken to addressing the problem of rank-$(L_r, L_r, 1)$ BTD model selection and computation, based on the idea of imposing column sparsity *jointly* on the factors and in a *hierarchical* manner and estimating the ranks as the numbers of factor columns of non-negligible energy. Approximate posterior inference for the proposed sophisticated Bayesian model is based on variational inference giving rise to an iterative algorithm that comprises closed-form updates. Its Bayesian nature completely avoids the ubiquitous in regularization-based methods task of hyper-parameter tuning. Simulation results with synthetic data are reported, which demonstrate the effectiveness of the proposed scheme in terms of both rank estimation and model fitting.**

***Index Terms*—Bayesian inference, block-term decomposition (BTD), hierarchical iterative reweighted least squares (HIRLS), rank, tensor, variational inference (VI)**

I. Introduction

Block-Term Decomposition (BTD) was introduced in [1] as a tensor model that combines the Canonical Polyadic Decomposition (CPD) and the Tucker decomposition (TD), in the sense that it decomposes a tensor in a sum of tensors (block terms) 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 (see [1, Fig. 2.3]). It can also be seen as a constrained CPD having factors with (some) colinear columns [1]. In a way, BTD 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 paper.

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. The work of E. Kofidis has been partly supported by the University of Piraeus Research Center.

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 work will also focus on this special yet very 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 $\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \mathbf{E}_r \circ \mathbf{c}_r$, 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. Eq. (1) can thus be re-written as

$$\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \mathbf{A}_r\mathbf{B}_r^{\mathrm{T}} \circ \mathbf{c}_r. \tag{1}$$

BTD has been successfully used in a wide range of application areas and its uniqueness and computation have been thoroughly studied (cf. [2] for an extensive review). In general, R and L_r, $r = 1, 2, \ldots, R$ are assumed *a-priori* known (and it is commonly assumed that all L_r are all equal to L, for simplicity). However, unless external information is given (such as in a telecommunications [3] or a hyperspectral image unmixing application with given or estimated ground truth [4]), there is no way to know these values beforehand. Model selection for BTD, that is estimating the number of block terms, R, and their ranks, L_r, $r = 1, 2, \ldots, R$, is clearly more challenging than in CPD and TD models and has only recently started to be studied (cf. [2] and references therein). The most recent contribution of this kind can be found in our work [2], which relies on a regularization of the squared approximation error function with the sum of the Frobenius norms of the factors reweighted by a diagonal weighting which jointly depends on the factors in two levels: the reweighted norms of $\mathbf{A} \triangleq \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_R \end{bmatrix}$ and $\mathbf{B} \triangleq \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 & \cdots & \mathbf{B}_R \end{bmatrix}$ are combined and then coupled with the reweighted norm of $\mathbf{C} \triangleq \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_R \end{bmatrix}$. This two-level coupling naturally matches the structure of the model

in (1), making explicit the different roles of $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$. This way, column sparsity is imposed *jointly* on the factors and in a *hierarchical* manner, which allows to estimate the ranks as the numbers of factor columns of non-negligible energy. Following a block coordinate descent solution approach, an alternating hierarchical iterative reweighted least squares (HIRLS) algorithm, called BTD-HIRLS, was developed in [2] that manages to both reveal the ranks and compute the BTD factors at a high convergence rate and low computational cost.

Nevertheless, BTD-HIRLS, being a regularization-based method, faces the same challenge that all such methods have to address, namely to appropriately tune the regularization parameter so as to achieve the best possible performance. Although a rough guideline for the parameter selection has been given and utilized in [2] as a reference point for the trial-end-error search, this is still only a rule of thumb, not completely relieving the algorithm from the need to spend resources on searching for the most appropriate parameter value. To overcome this difficulty, we take in this paper an alternative, Bayesian approach, viewing the unknowns as random variables and tackling the problem as one of Bayesian modeling and inference [5]. The idea is again to impose column sparsity jointly on the factors in a hierarchical, two-level manner. This is achieved through a Bayesian hierarchy of priors with sparsity inducing effect, that realize the coupling of the columns of $\mathbf{C}$ and the $\mathbf{A}_r, \mathbf{B}_r$ blocks at the outer level and that between the columns of corresponding blocks at the inner level. Thus, R is estimated as the number of columns of $\mathbf{C}$ of non-negligible energy while the L_r's are found similarly from the columns of the $\mathbf{A}_r, \mathbf{B}_r$ blocks. Approximate inference in the proposed probabilistic model is efficiently performed via variational inference [6] leading to an iterative algorithm that comprises closed-form updates and is fast converging. Its Bayesian nature completely avoids the need for parameter tuning. Simulation results with synthetic data are reported, which demonstrate the effectiveness of the proposed scheme in terms of both rank estimation and model fitting and in comparison with BTD-HIRLS. Bayesian methods for tensor decomposition model selection and computation have been already reported in the literature, with [7], [8], and [9] being recent examples concerning TD, CPD, and Tensor Train (TT) decomposition, respectively. To the best of our knowledge, however, the present work is the first of its kind for BTD in multilinear rank-$(L_r, L_r, 1)$ terms.

II. Problem Statement

Given the $I \times J \times K$ tensor

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} + \sigma \boldsymbol{\mathcal{N}}, \tag{2}$$

where $\boldsymbol{\mathcal{X}}$ is given by (1) and $\boldsymbol{\mathcal{N}}$ is the $I \times J \times K$ noise tensor of zero-mean unit variance i.i.d. Gaussian entries, we aim at estimating R, $L_r, r = 1, 2, \ldots, R$ and the factor matrices $\mathbf{A}_r = \begin{bmatrix} \mathbf{a}_{r1} & \mathbf{a}_{r2} & \cdots & \mathbf{a}_{rL_r} \end{bmatrix} \in \mathbb{C}^{I \times L_r}$, $\mathbf{B}_r = \begin{bmatrix} \mathbf{b}_{r1} & \mathbf{b}_{r2} & \cdots & \mathbf{b}_{rL_r} \end{bmatrix} \in \mathbb{C}^{J \times L_r}$, $\mathbf{C} \in \mathbb{C}^{K \times R}$, subject of course to the inherent ambiguity resulting from the fact that only the product $\mathbf{A}_r \mathbf{B}_r^{\mathrm{T}}$ can be uniquely identified modulo a scaling (with a counter-scaling of $\mathbf{c}_r$) [1]. In terms of its mode unfoldings $\mathbf{X}_{(1)} \in \mathbb{C}^{I \times JK}$, $\mathbf{X}_{(2)} \in \mathbb{C}^{J \times IK}$ and $\mathbf{X}_{(3)} \in \mathbb{C}^{K \times IJ}$, the tensor $\boldsymbol{\mathcal{X}}$ can be written as [1]

$$\begin{aligned} \mathbf{X}_{(1)}^{\mathrm{T}} &= (\mathbf{B} \odot \mathbf{C})\mathbf{A}^{\mathrm{T}} \triangleq \mathbf{P}\mathbf{A}^{\mathrm{T}}, && (3) \\ \mathbf{X}_{(2)}^{\mathrm{T}} &= (\mathbf{C} \odot \mathbf{A})\mathbf{B}^{\mathrm{T}} \triangleq \mathbf{Q}\mathbf{B}^{\mathrm{T}}, && (4) \\ \mathbf{X}_{(3)}^{\mathrm{T}} &= \begin{bmatrix} (\mathbf{A}_1 \odot_{\mathrm{c}} \mathbf{B}_1)\mathbf{1}_{L_1} & \cdots & (\mathbf{A}_R \odot_{\mathrm{c}} \mathbf{B}_R)\mathbf{1}_{L_R} \end{bmatrix} \mathbf{C}^{\mathrm{T}} \\ &\triangleq \mathbf{T}\mathbf{C}^{\mathrm{T}}, && (5) \end{aligned}$$

where $\odot$ denotes the Khatri-Rao product in its general (partition-wise) version and $\odot_{\mathrm{c}}$ is its column-wise version. In this paper, we follow a Bayesian approach to address the above problem. A fully Bayesian analysis is detailed next.

III. Proposed Bayesian Model

Let R and the L_rs be overestimated to R_{ini} and L_{ini}, respectively. Heavy-tailed multi-parameter Laplace distributions, known for their sparsity inducing effect, are placed over the columns of the $\mathbf{A}_r$s, $\mathbf{B}_r$s, and $\mathbf{C}$ in a way that implicitly implements a regularization analogous to that of the BTD-HIRLS method [2]. Namely, the number of block terms and the ranks of $\mathbf{A}_r$s and $\mathbf{B}_r$s are jointly penalized, while respecting the different role that these matrices play in the BTD model. This results in the nulling of all but R columns of $\mathbf{C}$, and the nulling of all but L_r columns of the corresponding "surviving" $\mathbf{A}_r, \mathbf{B}_r$ blocks. Following the premise of the well-known automatic relevance determination (ARD) framework [5], [10], the priors are assigned via a 3-level hierarchy of prior distributions outlined in the following.

The likelihood function, which encodes the underlying causal relation between the data and the latent variables, can be written with respect to (w.r.t.) the mode-1 unfolding of $\boldsymbol{\mathcal{Y}}$ (cf. (3)) as follows:

$$\begin{aligned} p(\mathbf{Y}_{(1)} \mid \mathbf{A}, \mathbf{B}, \mathbf{C}, \beta) &= \prod_{i=1}^{I} p(\boldsymbol{y}_{(1)i} \mid \mathbf{A}, \mathbf{B}, \mathbf{C}, \beta) \\ &= \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{y}_{(1)i} \mid \mathbf{P}\boldsymbol{a}_i, \beta^{-1}\mathbf{I}_{JK}) \end{aligned} \tag{6}$$

where β is the noise precision and $\boldsymbol{a}_i$, $\boldsymbol{y}_{(1)i}$ are the ith rows[1] of $\mathbf{A}$, $\mathbf{Y}_{(1)}$, respectively, in column form. In an analogous way the likelihood function can be written in terms of the rows $\boldsymbol{b}_j, \boldsymbol{c}_k$ and $\boldsymbol{y}_{(2)j}, \boldsymbol{y}_{(3)k}$ of $\mathbf{B}, \mathbf{C}$ and $\mathbf{Y}_{(2)}$, $\mathbf{Y}_{(3)}$, respectively, if $\boldsymbol{\mathcal{Y}}$ is expressed w.r.t. its mode-2 and mode-3 unfoldings. $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are considered as unobserved random variables and are assigned 3-level hierarchical prior distributions. At the first level of the hierarchy, Gaussian distributions are placed over $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, namely,

$$p(\mathbf{A} \mid \mathbf{s}, \boldsymbol{\zeta}, \beta) = \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{a}_i \mid \mathbf{0}, \beta^{-1}\mathbf{S}^{-1}(\mathbf{Z}^{-1} \otimes \mathbf{I}_{L_{\mathrm{ini}}})), \tag{7}$$

$$p(\mathbf{B} \mid \mathbf{s}, \boldsymbol{\zeta}, \beta) = \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{b}_j \mid \mathbf{0}, \beta^{-1}\mathbf{S}^{-1}(\mathbf{Z}^{-1} \otimes \mathbf{I}_{L_{\mathrm{ini}}})), \tag{8}$$

[1]We denote matrix rows and columns with bold italic and roman letters, respectively.

and

$$p(\mathbf{C} \mid \boldsymbol{\zeta}, \beta) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{c}_k \mid \mathbf{0}, \beta^{-1}\mathbf{Z}^{-1}), \tag{9}$$

where $\mathbf{S} = \text{diag}(\mathbf{s})$ with $\mathbf{s} = [s_{rl}] \in \mathbb{R}^{L_{\text{ini}}R_{\text{ini}} \times 1}$ and $\mathbf{Z} = \text{diag}(\boldsymbol{\zeta})$, $\boldsymbol{\zeta} \in \mathbb{R}^{R_{\text{ini}} \times 1}$. Note that the priors of $\mathbf{A}$ and $\mathbf{B}$ are zero-mean with the same covariance matrix, which is formed from the diagonal matrices $\mathbf{Z}$ and $\mathbf{S}$. This particular selection is of critical importance from an implicit regularization perspective, since it induces identical sparsity patterns over columns/sub-blocks of $\mathbf{A}$ and $\mathbf{B}$. A sufficiently large value of ζ_r will lead the rth column of $\mathbf{C}$ (cf. (9)) and the entire set of the redundant L_{ini} columns of sub-matrices $\mathbf{A}_r, \mathbf{B}_r$ (cf. (7), (8)) to zero. At the same time, the superfluous lth columns of the "surviving" $\mathbf{A}_r, \mathbf{B}_r$ are *jointly* forced to be zero when the value of s_{rl} becomes sufficiently large (cf. (7), (8)). Thus, the R_{ini} diagonal values of $\mathbf{Z}$ act as weights that determine the number of nonzero block terms, R, while the $L_{\text{ini}}R_{\text{ini}}$ diagonal values of $\mathbf{S}$ determine the number of nonzero columns of $\mathbf{A}$ and $\mathbf{B}$. The sophisticated selection of (7), (8) leads to *joint* sparsity imposition on the columns of the factors, in two levels. This way, ideas of ARD and sparse Bayesian leaning (SBL) [10] are put into effect to address the challenging problem of BTD model selection.

At the second level of the hierarchy of priors, inverse Gamma priors are assigned over $\mathbf{s}$ and $\boldsymbol{\zeta}$,

$$p(\mathbf{s}) = \prod_{r=1}^{R_{\text{ini}}} \prod_{l=1}^{L_{\text{ini}}} \mathcal{IG}\left(s_{rl} \,\middle|\, \frac{I+J+1}{2}, \frac{\delta_{rl}}{2}\right), \tag{10}$$

$$p(\boldsymbol{\zeta}) = \prod_{r=1}^{R_{\text{ini}}} \mathcal{IG}\left(\zeta_r \,\middle|\, \frac{(I+J)L_{\text{ini}}+K+1}{2}, \frac{\rho_r}{2}\right), \tag{11}$$

where δ_{rl} and ρ_r are the scale parameters of the distributions over s_{rl} and ζ_r, respectively. The third level involves Gamma prior distributions over these variables, namely,

$$p(\delta_{rl}) = \mathcal{G}(\delta_{rl} \mid \psi, \tau), \tag{12}$$

$$p(\rho_r) = \mathcal{G}(\rho_r \mid \mu, \nu), \tag{13}$$

where ψ, τ, μ, ν take very small positive values rendering the respective priors non-informative. Note that these priors are conjugate w.r.t. the likelihood functions and w.r.t. each other, which guarantees that the posterior distributions will belong to the same class of distributions with the priors [5]. Moreover, one can show that heavy-tailed multi-parameter Laplace marginal distributions show up over the columns of $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ after integrating out the variables $\mathbf{s}$ and $\boldsymbol{\zeta}$. Finally, a non-informative Gamma prior is also placed on the noise precision variable β, namely,

$$p(\beta) = \mathcal{G}(\beta \mid \kappa, \theta). \tag{14}$$

The above Bayesian model is depicted in the form of a graphical model in Fig. 1.

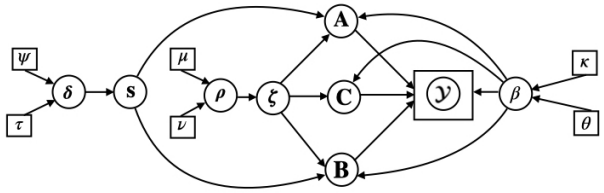

Fig. 1. The proposed Bayesian model.

IV. Approximate Posterior Inference

Let $\boldsymbol{\Theta}$ be the cell array which includes all unobserved variables, that is, $\boldsymbol{\Theta} \triangleq \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{s}, \boldsymbol{\zeta}, \beta, \boldsymbol{\rho}, \boldsymbol{\delta}\}$. The exact joint posterior of the variables of the adopted Bayesian model is given by

$$p(\boldsymbol{\Theta} \mid \boldsymbol{\mathcal{Y}}) = \frac{p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta})}{\int p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta}) d\boldsymbol{\Theta}}. \tag{15}$$

Due to the complexity of the model, the marginal distribution of $\boldsymbol{\mathcal{Y}}$ in the denominator is computationally intractable. Therefore, we follow a variational inference (VI) approach for approximating (15). The idea is to approximate the posterior by a distribution which is as close as possible to the exact posterior in terms of the Kullback-Leibler divergence [6]. VI allows for an efficient approximate inference process even in vastly complicated Bayesian models that involve high-dimensional variables. It is usually coupled with mean-field approximation, namely, the assumption that the posterior distribution can be factorized w.r.t. the involved variables, implying statistical independence among them. In our case, the approximate posterior $q(\boldsymbol{\Theta})$ of $p(\boldsymbol{\Theta} \mid \boldsymbol{\mathcal{Y}})$ is written in the form

$$q(\boldsymbol{\Theta}) = q(\beta) \prod_{i=1}^{I} q(\boldsymbol{a}_i) \prod_{j=1}^{J} q(\boldsymbol{b}_j) \prod_{k=1}^{K} q(\boldsymbol{c}_k) \times \prod_{r=1}^{R_{\text{ini}}} \prod_{l=1}^{L_{\text{ini}}} q(\mathbf{s}_{rl}) q(\delta_{rl}) \prod_{r=1}^{R_{\text{ini}}} q(\zeta_r) q(\rho_r) \tag{16}$$

Denoting the individual variables above by $\boldsymbol{\theta}_i$, the corresponding VI-based posteriors are known to satisfy [6]

$$q(\boldsymbol{\theta}_i) = \frac{\exp(\langle \ln(p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta})) \rangle_{i \neq j})}{\int \exp(\langle \ln(p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta})) \rangle_{i \neq j} d\boldsymbol{\theta}_i}, \tag{17}$$

where $\langle \cdot \rangle_{i \neq j}$ denotes expectation w.r.t. all $q(\boldsymbol{\theta}_j)$s but $q(\boldsymbol{\theta}_i)$. To solve (17) a block coordinate ascent approach is taken, employing the cyclic update rule, namely solving for $q(\boldsymbol{\theta}_i)$ given $q(\boldsymbol{\theta}_j)$, $j \neq i$. It follows that the posterior distribution of $\boldsymbol{a}_i$ is given by

$$q(\boldsymbol{a}_i) = \mathcal{N}(\langle \boldsymbol{a}_i \rangle, \Sigma_{\boldsymbol{a}}), \tag{18}$$

with[2]

$$\langle \boldsymbol{a}_i \rangle = \langle \beta \rangle \Sigma_{\boldsymbol{a}} \langle \mathbf{P} \rangle^{\mathrm{T}} \boldsymbol{y}_{(1)i}, \tag{19}$$

$$\Sigma_{\boldsymbol{a}} = \langle \beta \rangle^{-1} (\langle \mathbf{P}^{\mathrm{T}} \mathbf{P} \rangle + \langle \mathbf{S} \rangle (\langle \mathbf{Z} \rangle \otimes \mathbf{I}_{L_{\text{ini}}}))^{-1}, \tag{20}$$

[2]All $\boldsymbol{a}_i$s turn out to have the same covariance matrix, $\Sigma_{\boldsymbol{a}}$, and similarly for the $\boldsymbol{b}_j$s and the $\boldsymbol{c}_k$s.

where $\langle\cdot\rangle$ denotes expectation w.r.t the posterior of the involved variable. The posterior of $\boldsymbol{b}_j$ results in an analogous manner:

$$q(\mathbf{b}_j) = \mathcal{N}(\langle \boldsymbol{b}_j \rangle, \Sigma_{\boldsymbol{b}}), \tag{21}$$

with

$$\langle \boldsymbol{b}_j \rangle = \langle\beta\rangle \Sigma_{\boldsymbol{b}} \langle \mathbf{Q} \rangle^{\mathrm{T}} \boldsymbol{y}_{(2)j} \tag{22}$$

$$\Sigma_{\boldsymbol{b}} = \langle\beta\rangle^{-1} (\langle \mathbf{Q}^{\mathrm{T}}\mathbf{Q} \rangle + \langle \mathbf{S} \rangle (\langle \mathbf{Z} \rangle \otimes \mathbf{I}_{L_{\mathrm{ini}}}))^{-1}. \tag{23}$$

The posterior of $\boldsymbol{c}_k$ is

$$q(\boldsymbol{c}_k) = \mathcal{N}(\langle \boldsymbol{c}_k \rangle, \Sigma_{\boldsymbol{c}}) \tag{24}$$

with

$$\langle \boldsymbol{c}_k \rangle = \langle\beta\rangle \Sigma_{\boldsymbol{c}} \langle \mathbf{T} \rangle^{\mathrm{T}} \boldsymbol{y}_{(3)k} \tag{25}$$

$$\Sigma_{\boldsymbol{c}} = \langle\beta\rangle^{-1} (\langle \mathbf{T}^{\mathrm{T}}\mathbf{T} \rangle + \langle \mathbf{Z} \rangle)^{-1}. \tag{26}$$

Next, the approximate posteriors of the variables belonging to the second level of hierarchy are given. The posterior of s_{rl} turns out [11] to be a GIG pdf,

$$q(s_{rl}) = \mathcal{GIG}\left(s_{rl} \middle| -\frac{1}{2}, \langle\beta\rangle \langle\zeta_r\rangle (\langle \mathbf{a}_{rl}^{\mathrm{T}} \mathbf{a}_{rl} \rangle + \langle \mathbf{b}_{rl}^{\mathrm{T}} \mathbf{b}_{rl} \rangle), \langle \delta_{rl} \rangle \right) \tag{27}$$

with mean

$$\langle s_{rl} \rangle = \sqrt{\frac{\langle \delta_{rl} \rangle}{\langle\beta\rangle \langle\zeta_r\rangle (\langle \mathbf{a}_{rl}^{\mathrm{T}} \mathbf{a}_{rl} \rangle + \langle \mathbf{b}_{rl}^{\mathrm{T}} \mathbf{b}_{rl} \rangle)}}, \tag{28}$$

where $\langle \mathbf{a}_{rl}^{\mathrm{T}} \mathbf{a}_{rl} \rangle$ and $\langle \mathbf{b}_{rl}^{\mathrm{T}} \mathbf{b}_{rl} \rangle$ are the $((r-1)L_{\mathrm{ini}} + l, (r-1)L_{\mathrm{ini}} + l)$ entries of

$$\langle \mathbf{A}^{\mathrm{T}} \mathbf{A} \rangle = \langle \mathbf{A} \rangle^{\mathrm{T}} \langle \mathbf{A} \rangle + I \Sigma_{\boldsymbol{a}} \tag{29}$$

and

$$\langle \mathbf{B}^{\mathrm{T}} \mathbf{B} \rangle = \langle \mathbf{B} \rangle^{\mathrm{T}} \langle \mathbf{B} \rangle + J \Sigma_{\boldsymbol{b}}, \tag{30}$$

respectively. Similarly, the approximate posterior of ζ_r is also GIG, with $\langle \zeta_r \rangle$ given by

$$\langle \zeta_r \rangle = \sqrt{\frac{\langle \rho_r \rangle}{\langle\beta\rangle (\sum_{l=1}^{L_{\mathrm{ini}}} \langle s_{rl} \rangle (\langle \mathbf{a}_{rl}^{\mathrm{T}} \mathbf{a}_{rl} \rangle + \langle \mathbf{b}_{rl}^{\mathrm{T}} \mathbf{b}_{rl} \rangle) + \langle \mathbf{c}_r^{\mathrm{T}} \mathbf{c}_r \rangle)}} \tag{31}$$

and $\langle \mathbf{c}_r^{\mathrm{T}} \mathbf{c}_r \rangle$ denoting the (r, r) entry of

$$\langle \mathbf{C}^{\mathrm{T}} \mathbf{C} \rangle = \langle \mathbf{C} \rangle^{\mathrm{T}} \langle \mathbf{C} \rangle + K \Sigma_{\boldsymbol{c}}. \tag{32}$$

It can be shown (as in [11]) that, at the third level of hierarchy, the approximate posteriors of δ_{rl}, ρ_r and β are Gamma distributions with $\langle \delta_{rl} \rangle, \langle \rho_r \rangle$ and $\langle \beta \rangle$ given in Table I, where the resulting *Bayesian-BTD (BBTD)* algorithm is summarized (cf. [12] for a more detailed presentation). The rest of the first- and second-order statistics that are required in the algorithm implementation are computed as in Table II, based on the assumption of statistically independent $\mathbf{A}, \mathbf{B}, \mathbf{C}$ (cf. (16)) and making use of the identities for the Grammians of Khatri-Rao products proved in [2, Appendix C]. $*$ stands for the Hadamard product.

TABLE I
THE BBTD ALGORITHM

Input: $\boldsymbol{\mathcal{Y}}, R_{\mathrm{ini}}, L_{\mathrm{ini}}$
Output: $\hat{R}$, $\hat{L}_r, r = 1, 2, \ldots, \hat{R}$, $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$
Initialize $\langle \mathbf{B} \rangle, \langle \mathbf{C} \rangle, \langle \beta \rangle, \langle \mathbf{S} \rangle, \langle \mathbf{Z} \rangle, \langle \boldsymbol{\delta} \rangle, \langle \boldsymbol{\rho} \rangle, \Sigma_{\boldsymbol{b}}, \Sigma_{\boldsymbol{c}}$
repeat
$\Sigma_{\boldsymbol{a}} = \langle\beta\rangle^{-1} (\langle \mathbf{P}^{\mathrm{T}}\mathbf{P} \rangle + \langle \mathbf{S} \rangle (\langle \mathbf{Z} \rangle \otimes \mathbf{I}_{L_{\mathrm{ini}}}))^{-1}$
$\langle \mathbf{A} \rangle = \langle\beta\rangle \mathbf{Y}_{(1)} \langle \mathbf{P} \rangle \Sigma_{\boldsymbol{a}}$
$\Sigma_{\boldsymbol{b}} = \langle\beta\rangle^{-1} (\langle \mathbf{Q}^{\mathrm{T}}\mathbf{Q} \rangle + \langle \mathbf{S} \rangle (\langle \mathbf{Z} \rangle \otimes \mathbf{I}_{L_{\mathrm{ini}}}))^{-1}$
$\langle \mathbf{B} \rangle = \langle\beta\rangle \mathbf{Y}_{(2)} \langle \mathbf{Q} \rangle \Sigma_{\boldsymbol{b}}$
$\Sigma_{\boldsymbol{c}} = \langle\beta\rangle^{-1} (\langle \mathbf{T}^{\mathrm{T}}\mathbf{T} \rangle + \langle \mathbf{Z} \rangle)^{-1}$
$\langle \mathbf{C} \rangle = \langle\beta\rangle \mathbf{Y}_{(3)} \langle \mathbf{T} \rangle \Sigma_{\boldsymbol{c}}$
$r = 1, 2, \ldots, R_{\mathrm{ini}}, l = 1, 2, \ldots, L_{\mathrm{ini}}$
$\langle s_{rl} \rangle = \sqrt{\frac{\langle \delta_{rl} \rangle}{\langle\beta\rangle \langle\zeta_r\rangle (\langle \mathbf{a}_{rl}^{\mathrm{T}} \mathbf{a}_{rl} \rangle + \langle \mathbf{b}_{rl}^{\mathrm{T}} \mathbf{b}_{rl} \rangle)}}$
$\left\langle \frac{1}{s_{rl}} \right\rangle = \frac{1}{\langle \delta_{rl} \rangle} + \frac{1}{\langle s_{rl} \rangle}$
$\langle \delta_{rl} \rangle = \frac{2\psi + I + J + 1}{2\tau + \langle \frac{1}{s_{rl}} \rangle}$
$r = 1, 2, \ldots, R_{\mathrm{ini}}$
$\langle \zeta_r \rangle = \sqrt{\frac{\langle \rho_r \rangle}{\langle\beta\rangle (\sum_{l=1}^{L_{\mathrm{ini}}} \langle s_{rl} \rangle (\langle \mathbf{a}_{rl}^{\mathrm{T}} \mathbf{a}_{rl} \rangle + \langle \mathbf{b}_{rl}^{\mathrm{T}} \mathbf{b}_{rl} \rangle)) + \langle \mathbf{c}_r^{T} \mathbf{c}_r \rangle}}$
$\left\langle \frac{1}{\zeta_r} \right\rangle = \frac{1}{\langle \rho_r \rangle} + \frac{1}{\langle \zeta_r \rangle}$
$\langle \rho_r \rangle = \frac{2\mu + (I+J) L_{\mathrm{ini}} + K + 1}{2\nu + \langle \frac{1}{\zeta_r} \rangle}$
$\langle\beta\rangle = (2\kappa + (I+J) L_{\mathrm{ini}} R_{\mathrm{ini}} + K R_{\mathrm{ini}} + IJK + 1) / (2\theta + \left\langle \left\| \mathbf{Y}_{(1)}^{\mathrm{T}} - \mathbf{P}\mathbf{A}^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \right\rangle + \sum_{r=1}^{R_{\mathrm{ini}}} \langle\zeta_r\rangle [\sum_{l=1}^{L_{\mathrm{ini}}} \langle s_{rl} \rangle (\langle \mathbf{a}_{rl}^{\mathrm{T}} \mathbf{a}_{rl} \rangle + \langle \mathbf{b}_{rl}^{\mathrm{T}} \mathbf{b}_{rl} \rangle) + \langle \mathbf{c}_r^{\mathrm{T}} \mathbf{c}_r \rangle])$
until ***convergence***

TABLE II
FIRST- AND SECOND-ORDER STATISTICS REQUIRED IN BBTD

$\langle \mathbf{P} \rangle = \langle \mathbf{B} \rangle \odot \langle \mathbf{C} \rangle$
$\langle \mathbf{Q} \rangle = \langle \mathbf{C} \rangle \odot \langle \mathbf{A} \rangle$
$\langle \mathbf{T} \rangle = \left[(\langle \mathbf{A}_1 \rangle \odot_{\mathrm{c}} \langle \mathbf{B}_1 \rangle) \mathbf{1}_{L_{\mathrm{ini}}} \cdots (\langle \mathbf{A}_{R_{\mathrm{ini}}} \rangle \odot_{\mathrm{c}} \langle \mathbf{B}_{R_{\mathrm{ini}}} \rangle) \mathbf{1}_{L_{\mathrm{ini}}} \right]$
$\langle \mathbf{P}^{\mathrm{T}}\mathbf{P} \rangle = \langle \mathbf{B}^{\mathrm{T}}\mathbf{B} \rangle * (\langle \mathbf{C}^{\mathrm{T}}\mathbf{C} \rangle \otimes \mathbf{1}_{L_{\mathrm{ini}} \times L_{\mathrm{ini}}})$
$\langle \mathbf{Q}^{\mathrm{T}}\mathbf{Q} \rangle = \langle \mathbf{A}^{\mathrm{T}}\mathbf{A} \rangle * (\langle \mathbf{C}^{\mathrm{T}}\mathbf{C} \rangle \otimes \mathbf{1}_{L_{\mathrm{ini}} \times L_{\mathrm{ini}}})$
$\langle \mathbf{T}^{\mathrm{T}}\mathbf{T} \rangle = (\mathbf{I}_{R_{\mathrm{ini}}} \otimes \mathbf{1}_{L_{\mathrm{ini}}}^{\mathrm{T}}) (\langle \mathbf{A}^{\mathrm{T}}\mathbf{A} \rangle * \langle \mathbf{B}^{\mathrm{T}}\mathbf{B} \rangle) (\mathbf{I}_{R_{\mathrm{ini}}} \otimes \mathbf{1}_{L_{\mathrm{ini}}})$
$\langle \| \mathbf{Y}_{(1)} - \mathbf{P}\mathbf{A}^{\mathrm{T}} \|_{\mathrm{F}}^2 \rangle = \| \mathbf{Y}_{(1)} \|_{\mathrm{F}}^2 - 2\mathrm{tr}\{ \langle \mathbf{A} \rangle^{\mathrm{T}} \mathbf{Y}_{(1)}^{\mathrm{T}} \langle \mathbf{P} \rangle \} + \mathrm{tr}\{ \langle \mathbf{A}^{\mathrm{T}}\mathbf{A} \rangle \langle \mathbf{P}^{\mathrm{T}}\mathbf{P} \rangle \}$

R is estimated as the number of columns of $\langle \mathbf{C} \rangle$ of non-negligible energy and similarly for the L_rs and the corresponding blocks of $\langle \mathbf{A} \rangle, \langle \mathbf{B} \rangle$. The iterations stop when a convergence criterion is met (e.g., the relative difference of the tensor reconstruction errors in two consecutive iterations becomes less than a user-defined threshold) or the maximum number of iterations is reached.

V. SIMULATION RESULTS

In this section, we evaluate the effectiveness of the BBTD algorithm in revealing the ranks and computing the model parameters, in comparison with the regularization-based BTD-HIRLS scheme with its regularization parameter being selected so as to achieve the minimum approximation error. A $55 \times 55 \times 20$ tensor $\boldsymbol{\mathcal{Y}}$ is generated as in (2), with $R = 5$ and the L_rs being set as $L_1 = 8, L_2 = 6, L_3 = 4, L_4 = 5$ and $L_5 = 3$. The entries of $\mathbf{A}_r, \mathbf{B}_r$ and $\mathbf{C}$ are i.i.d., sampled from the standard Gaussian distribution. The noise power is set so as to result in a signal-to-noise ratio $\mathrm{SNR} = 10 \log_{10} \|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}^2 / (\sigma^2 \|\boldsymbol{\mathcal{N}}\|_{\mathrm{F}}^2)$ of only 5 dB. Both R

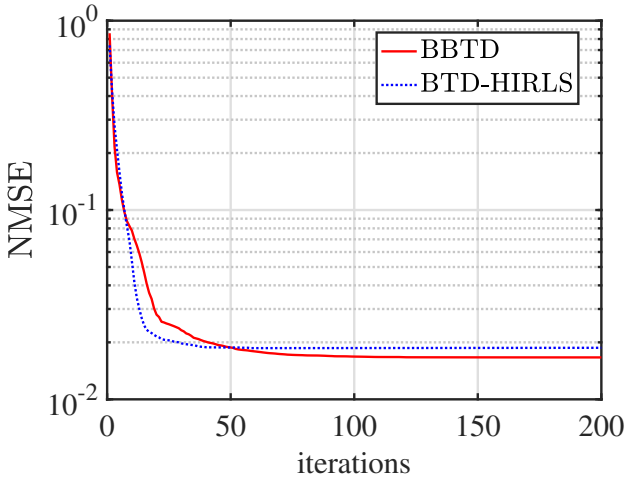

Fig. 2. NMSE vs. iterations.

and all L_rs are overestimated as 10. The evolution of the median of the NMSE $= \sum_{r=1}^{R} \frac{\|\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}} \circ \mathbf{c}_r - \hat{\mathbf{A}}_r\hat{\mathbf{B}}_r^{\mathrm{T}} \circ \hat{\mathbf{c}}_r\|_{\mathrm{F}}^2}{\|\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}} \circ \mathbf{c}_r\|_{\mathrm{F}}^2}$ values[3] obtained for 100 independent realizations of $\mathcal{Y}$ is plotted in Fig. 2. It should be stressed that both BTD-HIRLS (as demonstrated in [2]) and BBTD appear to be insensitive to different initializations, hence a single random initialization is used here. Observe that both algorithns converge in less than 100 iterations.

Fig. 3 depicts the success rates in the recovery of the true R and L_rs. The Bayesian algorithm is seen to recover the correct number of block terms in this scenario with a slightly lower probability than the state-of-the-art BTD-HIRLS algorithm. It should be noted however that in the rare cases that BBTD fails, this happens only with a small deviation from the true $R = 5$, namely 6. As it can be observed from Figs. 2(b)–(f), where the success rates in recovering the individual ranks given a correctly estimated R are shown, BBTD performs similarly with or better than BTD-HIRLS, achieving almost 100% accuracy in all terms. Instead BTD-HIRLS exhibits lower accuracy in estimating $L_1 = 8$. That BBTD performs *overall* (in estimating *both* R and the L_rs) better than BTD-HIRLS can be attributed to the probabilistic model that is adopted, which allows BBTD to better capture the structure of BTD when it comes to the low-rankness of the $\mathbf{A}_r, \mathbf{B}_r$. Namely, BBTD implicitly weighs differently each pair $\mathbf{A}_r, \mathbf{B}_r$, as opposed to BTD-HIRLS, which uniformly weighs all block terms through a single regularization parameter. This also explains the fact that the BTD-HIRLS estimation for R is relatively more accurate.

[3]The Hungarian algorithm is employed to match the $\hat{R}$ estimated non-zero block terms with the true ones.

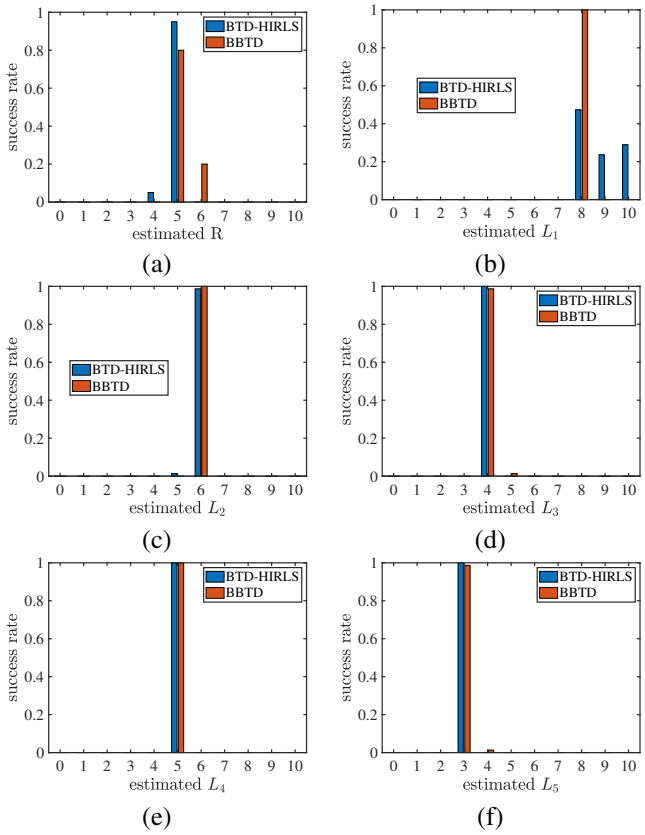

Fig. 3. Success rates of recovering (a) R and (b)–(f) L_rs via the BBTD and BTD-HIRLS algorithms.

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] A. A. Rontogiannis, E. Kofidis, and P. V. Giampouras, "Block-term tensor decomposition: Model selection and computation," *IEEE J. Sel. Topics Signal Process.*, Jan. 2021.

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

[4] Y. Qian *et al.*, "Matrix-vector nonnegative tensor factorization for blind unmixing of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1776–1792, Mar. 2017.

[5] S. Theodoridis, *Machine Learning — A Bayesian and Optimization Perspective*, 2nd ed. Academic Press, 2020.

[6] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, pp. 131–146, Nov. 2008.

[7] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian sparse Tucker models for dimension reduction and tensor completion," arXiv:1505.02343v1 [cs.LG], May 2015.

[8] L. Cheng *et al.*, "Towards probabilistic tensor canonical polyadic decomposition 2.0: Automatic tensor rank learning using generalized hyperbolic prior," arXiv:2009.02472v1 [cs.LG], Sep. 2020.

[9] L. Xu *et al.*, "Learning tensor train representation with automatic rank determination from incomplete noisy data," arXiv:2010.06564v1 [eess.SP], Oct. 2020.

[10] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jun. 2001.

[11] P. V. Giampouras *et al.*, "Online sparse and low-rank subspace learning from incomplete data: A Bayesian view," *Signal Process.*, vol. 137, pp. 199–212, 2017.

[12] P. V. Giampouras, A. A. Rontogiannis, and E. Kofidis, "A Bayesian approach to block-term tensor decomposition model selection and computation," arXiv:2101.02931v1 [stat.ME], Jan. 2021.