Block-Term Tensor Decomposition Model Selection and Computation: The Bayesian Way

Paris V. Giampouras, Athanasios A. Rontogiannis, *Senior Member, IEEE,* and Eleftherios Kofidis, *Member, IEEE*

***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. Uniqueness conditions and fitting methods have thus been thoroughly studied. Nevertheless, the challenging problem of estimating the BTD model structure, namely the number of block terms, R, and their individual ranks, L_r, has only recently started to attract significant attention, mainly through regularization-based approaches which entail the need to tune the regularization parameter(s). In this work, we build on ideas of sparse Bayesian learning (SBL) and put forward a fully automated Bayesian approach. Through a suitably crafted multi-level *hierarchical* probabilistic model, which gives rise to heavy-tailed prior distributions for the BTD factors, structured sparsity is *jointly* imposed. Ranks are then estimated from the numbers of blocks (R) and columns (L_r) of non-negligible energy. Approximate posterior inference is implemented, within the variational inference framework. The resulting iterative algorithm completely avoids hyperparameter tuning, which is a significant defect of regularization-based methods. Alternative probabilistic models are also explored and the connections with their regularization-based counterparts are brought to light with the aid of the associated maximum a-posteriori (MAP) estimators. We report simulation results with both synthetic and real-word data, which demonstrate the merits of the proposed method in terms of both rank estimation and model fitting as compared to state-of-the-art relevant methods.**

***Index Terms*—Automatic relevance determination (ARD), Bayesian inference, block-term decomposition (BTD), hierarchical iterative reweighted least squares (HIRLS), rank, sparse Bayesian learning (SBL), 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) [2], in the sense that it decomposes a tensor in a sum of tensors (block terms) that have low multilinear rank (not necessarily 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) collinear columns [1]. In a way, BTD lies between the two extremes (in terms of core tensor structure), CPD and TD, and it is useful to recall here 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, via a probabilistic approach that relaxes the requirement for hyperparameters tuning, is the main subject of this paper.

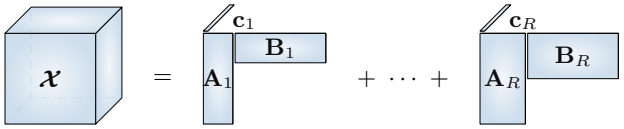


Fig. 1. Rank-$(L_r, L_r, 1)$ block-term decomposition.

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 a wide range of applications and the existence of more concrete and easier to check uniqueness conditions (cf. [3] for an extensive review). This special yet very popular BTD model is at the focus of the present work. 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

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

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. Eq. (1) can thus be re-written as

$$\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \left(\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}}\right) \circ \mathbf{c}_r. \tag{2}$$

A schematic representation of the rank-$(L_r, L_r, 1)$ BTD is given in Fig. 1. The rth term of this decomposition is a tensor whose frontal slices are all scalar multiples (with the entries of $\mathbf{c}_r$) of the low-rank matrix $\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}}$. It should be apparent from (2) and Fig. 1 that CPD results as a special case with all $L_r, r = 1, 2, \ldots, R$ equal to 1.

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 [4] or a hyperspectral

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

P. V. Giampouras is with the Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD 212 18, USA. E-mail: parisg@jhu.ed

A. A. Rontogiannis is with the IAASARS, National Observatory of Athens, 152 36 Penteli, Greece. E-mail: tronto@noa.gr.

E. Kofidis is with the Dept. of Statistics and Insurance Science, University of Piraeus, 185 34 Piraeus, Greece and the Computer Technology Institute & Press "Diophantus" (CTI), 265 04 Patras, Greece. E-mail: kofidis@unipi.gr

image unmixing application with given or estimated ground truth [5]), there is no way to know these values beforehand. Although overestimation of the ranks L_rs of the block terms has been observed not to be harmful in some blind source separation applications (e.g., [4]), this is not the case in general [3]. Besides, in addition to increasing the computational complexity, setting L_r too high may hinder interpretation of the results through letting noise/artifact sources interfere with the desired sources. This holds for R as well, whose choice is known to be more crucial to the obtained performance as it represents the number of "factors" that generate the data and its over/under-estimation will lead to over/under-fitting, with undesired consequences for the interpretability of the results (cf. [3] for related references).

A. Prior art

It is known that computing the number of rank-1 terms in a CPD model (i.e. the tensor rank) is NP-hard [6]. Model selection for BTD is clearly even more challenging than in CPD and TD models and has only recently started to be studied (cf. [3] for an extensive review of heuristic approaches and techniques). The most recent contribution of this kind can be found in our work [3], where the latent factors of BTD are recovered by solving a regularized minimization problem, namely,

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \frac{1}{2}\left\|\boldsymbol{\mathcal{Y}} - \sum_{r=1}^{R} \mathbf{A}_r\mathbf{B}_r^{\mathrm{T}} \circ \mathbf{c}_r\right\|_{\mathrm{F}}^2 + \lambda \sum_{r=1}^{R} \sqrt{\sum_{l=1}^{L} \sqrt{\|\mathbf{a}_{r,l}\|_2^2 + \|\mathbf{b}_{r,l}\|_2^2 + \eta^2} + \|\mathbf{c}_r\|_2^2 + \eta^2}, \quad (3)$$

where R, L are over-estimates of the rank and block ranks of the sought BTD model, $\mathbf{a}_{r,l}, \mathbf{b}_{r,l}$, $l = 1, 2, \ldots, L$ are the lth columns of $\mathbf{A}_r, \mathbf{B}_r$, respectively and η^2 is a small constant used to ensure smoothness at zero. Note that (3) is composed of the squared Frobenius norm of the error between the data and its BTD representation and an appropriately chosen regularization term whose minimization promotes structured sparsity over the latent factors of the model. The rank, R, and the block ranks, L_r, are then taken as the number of $\mathbf{c}_r$s of non-negligible magnitude and the numbers of non-negligible columns of the corresponding blocks, respectively. Structured sparsity is favored by the regularizer in a hierarchical, two-level manner, which is tailored to the form of the BTD model. Indeed, the inner sum of square roots is (excluding the smoothing constant) the sum of the ℓ_2 norms of the columns of $\left[\begin{array}{cc}\mathbf{A}_r^{\mathrm{T}} & \mathbf{B}_r^{\mathrm{T}}\end{array}\right]^{\mathrm{T}}$, for $r = 1, 2, \ldots, R$. The well-known column sparsity-promoting effect of this $\ell_{1,2}$ norm leads the superfluous columns of both matrices to be driven jointly to zero, thus providing a "relaxed" way of penalizing the block ranks of the BTD model. In an analogous manner, the outer sum of the regularizer penalizes jointly the number of nonzero columns of $\mathbf{C}$ along with the corresponding blocks $\mathbf{A}_r\mathbf{B}_r^{\top}$, which coincides with the number of block terms in the model. The hierarchical alternating iterative reweighted least squares algorithm, called BTD-HIRLS, proposed in [3] to solve the above problem has demonstrated its competence in revealing the true ranks and accurately computing the model parameters, while enjoying computational efficiency and fast convergence.

Nevertheless, being a regularization-based method, BTD-HIRLS 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 [3] as a reference point for the trial-and-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 regularization parameter value.

B. The Bayesian way

One would thus prefer to be able to automatically (not manually) select the value of the regularization parameter or, more generally, discover the columns of the factor matrices that should be kept, in an automatic, completely data-driven manner. Such a possibility is provided by what is known as *sparse Bayesian leaning (SBL)* [7], [8] following the *automatic relevance determination (ARD)* approach, first conceived for and applied in sparsifying the weights of a neural network [9]. Through this Bayesian perspective, the unknown parameters of the problem are viewed as random quantities and are each associated with a hyperparameter. Prior distributions suitably assigned to each hyperparameter are conducive to automatically determining the relevance of the associated parameters at inference time.

In the so-called ARD prior, the parameters are independent and zero-mean Gaussian if conditioned on the values of their hyperparameters, which are represented by the corresponding standard deviations. Hence if the hyperparameter is large enough, the parameter is important whereas for a small enough hyperparameter the corresponding parameter should be suppressed, thus revealing the true complexity (rank) of the model. As stated in [9], "the posterior distributions of these hyperparameters will reflect which of these situations is more probable, in light of the training data." If a parameter is relevant, this will influence the associated hyperparameter distribution which in turn will make the parameter more important, in an alternating update cycle between the parameter and hyperparameter posteriors.

ARD was first applied in automatic tensor rank learning on multi-way data modeled via TD in [10]. The hyperparameters (inverse powers of factor columns, also known as precisions) were modeled with Gamma priors, giving rise to the so-called Gauss-Gamma (GG) probabilistic model, where the marginal posterior of the parameters turns out to be a Laplacian, with its well-known sparsity-enforcing effect [8]. An analogous GG model was adopted in [11] for addressing the corresponding problem for incomplete tensors obeying a CPD model. A fully Bayesian inference approach was taken, in contrast to the *maximum a-posteriori (MAP)* estimation approach followed in [10]. The method proposed in [11] performs approximate variational inference (VI) [12], [13], in the sense that the posterior densities are found as the closest (in the sense of minimum

Kullback-Leibler (KL) divergence) to the true ones that meet the mean-field assumption of statistical independence of all parameters and hyperparameters. VI is known to be generally faster converging than sampling techniques and better suited to large datasets [13].

The method of [11] was later robustified to cope with incomplete tensors with outliers [14]. An online version, for tensors that may grow in time in all their modes and in any order, was reported in [15]. Since [11], several works on Bayesian tensor model selection and computation have been reported for both CPD (cf. [16] and references therein) and other tensor decomposition models including TD, tensor trains (TT), tensor rings (TR), and t-SVD, among others (see, e.g., [17]–[22]). In [23], a TT decomposition is employed to compress a deep neural network (DNN) during its training.[1] The TT ranks are automatically determined through a Bayesian GG modeling approach which models the powers of the slices of the TT cores by Gamma priors and couples consecutive cores through the product of their associated hyperparameters.

The GG model is generalized in [17] for Bayesian TD by replacing the Gamma hyperprior by an inverse Gamma (IG). This results in a multivariate Laplace marginal prior for the parameters, which also leads to a generalized inverse Gaussian (GIG) for the posterior of the sparsity-inducing precision hyperparameters. Similarly with [23], the core tensor is indirectly coupled with the matrix factors by using the product of these hyperparameters and the noise precision in the core's normal prior. A more recent generalization of the GG model, this time for CPD rank learning, is developed in [16] through a Gauss-GIG mixture that leads to a generalized hyperbolic (GH) marginal prior for the CPD factors. GH is known to be very flexible, including several other sparsity-enforcing distributions as special cases [25, Table I]. The value of this generalization is demonstrated by the fact that the resulting VI method outperforms [11] for high-rank tensors and/or low signal-to-noise ratio (SNR). It should be noted, however, that the algorithm in [16] is developed on the basis of a simplification of the GH distribution (cf. Section IV-C), which effectively leads again to a (generalized) Laplacian marginal prior.

C. Our contribution

In this paper, we also take a Bayesian approach, viewing the unknowns as random variables and tackling the problem as one of Bayesian modeling and inference [26]. The idea is again (as in BTD-HIRLS) 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. Our choices of priors fall in the class of the so-called exponential power distributions with GIG densities (EP-GIG) [27], which include the GG of [11] and the Gauss-GIG and GH of [16], [17] as special cases. Inspired by earlier work of ours [28] and in a manner analogous with the way coupling is achieved in [17] for TD, we realize the two-level coupling in the BTD model via appropriately defined products of the associated hyperparameters and the noise precision in the conditional priors of the factors. It is shown that, with our choices of priors, conjugacy is maintained, which allows the development of a tractable approximate inference, efficiently performed via VI [12], [13] and leading to an iterative algorithm that comprises closed-form updates and is fast converging. Overestimates of R and the L_rs are decreased in the course of the algorithm. This is in contrast to the rank incremental or greedy strategies followed in, e.g., [20] and [29]. 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. The Bayesian nature of our approach completely avoids the need for parameter tuning. We also present alternative Bayesian models that reflect simplified causal relationships among the latent variables and thus can be used for lending an insight into the incurred regularization effect through the lens of the MAP-based optimization problems. Simulation results with both synthetic and real 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. To the best of our knowledge, the present work is the first of its kind for BTD model selection and computation. A preliminary version can be found in [30]. In a shorter version, this work was accepted for presentation in EUSIPCO-2021 [31].

D. Organization of the paper

The rest of this paper is organized as follows. The adopted notation is described in the following subsection. The problem is mathematically stated in Section II, where useful expressions for the tensor unfoldings are also recalled. A Bayesian model that implements the idea underlying BTD-HIRLS is developed in Section III. The corresponding approximate inference method is presented and analyzed in Section IV. Alternative probabilistic models, that are inspired from deterministic criteria simpler than (3), are considered in Section V along with the associated MAP estimators, which clarify the connections with the regularization-based approach. Section VI reports and discusses the simulation results. Conclusions are drawn and future work plans are outlined in Section VII.

E. Notation

Lower- and upper-case bold letters are used to denote vectors and matrices, respectively. We denote matrix rows with bold italic letters and we use roman letters for the matrix columns. 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. $*$ stands for the Hadamard product and $\otimes$ for the Kronecker product. The Khatri-Rao product is denoted

[1] In fact, the power of deep learning (in the form of a convolutional neural network trained on (rank,tensor) pairs) was also exploited to learn to estimate the rank of any given tensor in [24], with results that suggest an improvement over Bayesian schemes like [11]. Of course, one should also consider, in such a comparison, the well-known lack of interpretability of a trained deep neural network vis-à-vis the relatively well-understood principles underlying the purely Bayesian approach.

by $\odot$ in its general (partition-wise) version and by $\odot_{\rm c}$ in its column-wise version. $\circ$ denotes the outer product. The superscript $^{\rm T}$ stands for transposition. The identity matrix of order N and the all ones $M \times N$ matrix are respectively denoted by $\mathbf{I}_N$ and $\mathbf{1}_{M\times N}$. $\mathbf{1}_N$ stands for $\mathbf{1}_{N\times 1}$. $\text{diag}(\mathbf{x})$ is the diagonal matrix with the vector $\mathbf{x}$ on its main diagonal. The Euclidean vector norm and the Frobenius tensor norm are denoted by $\|\cdot\|_2$ and $\|\cdot\|_{\rm F}$, respectively. $\text{tr}\{\cdot\}$ stands for the trace operator. $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the normal probability density function (pdf) for a random vector $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. $\mathbf{x}$ is omitted when it is easily understood from the context. The generalized inverse Gaussian (GIG) pdf [32] is given by $\mathcal{GIG}(x|p,a,b) = \frac{(a/b)^{p/2}\exp\left[(p-1)\log x - \left(ax+\frac{b}{x}\right)/2\right]}{2\mathcal{K}_p(\sqrt{ab})}$, where $x > 0$, p is real, and $\mathcal{K}_p(\cdot)$ is the modified Bessel function of the second kind with index p. The Gamma pdf with shape ζ and rate τ results as a special case for $b \to 0, p > 0$ and is defined as $\mathcal{G}(x|\zeta,\tau) = \frac{\tau^\zeta}{\Gamma(\zeta)}x^{\zeta-1}e^{-\tau x} = \exp\left[(\zeta-1)\log x - x\tau - \log\Gamma(\zeta) + \zeta\log\tau\right]$, where $\Gamma(\cdot)$ is the Gamma function, $\Gamma(\zeta) = \int_0^\infty x^{\zeta-1}e^{-x}dx$. The inverse (or reciprocal) Gamma pdf also results from the GIG one as a special case (for $a \to 0$, $p < 0$) and, in its shape (ζ) and scale (τ) parametrization, is given by $\mathcal{IG}(x|\zeta,\tau) = \frac{\tau^\zeta}{\Gamma(\zeta)}x^{-(\zeta+1)}e^{-\tau/x} = \exp\left[-(\zeta+1)\log x - \frac{\tau}{x} - \log\Gamma(\zeta) + \zeta\log\tau\right]$, for $x > 0$. Sets are denoted by calligraphic letters. For a set $\mathcal{M}$, $|\mathcal{M}|$ is its cardinality. $\mathbb{R}$ and $\mathbb{C}$ are the fields of real and complex numbers, respectively.

II. Problem Statement

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

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} + \sigma\boldsymbol{\mathcal{N}}, \tag{4}$$

where $\boldsymbol{\mathcal{X}}$ is given by (2) and $\boldsymbol{\mathcal{N}}$ is a $I \times J \times K$ noise tensor of zero-mean unit variance i.i.d. Gaussian entries, with σ being the noise standard deviation, we aim at estimating R, $L_r, r = 1, 2, \ldots, R$ and the factor matrices $\mathbf{A}_r = \left[\begin{array}{cccc}\mathbf{a}_{r,1} & \mathbf{a}_{r,2} & \cdots & \mathbf{a}_{r,L_r}\end{array}\right] \in \mathbb{C}^{I\times L_r}$, $\mathbf{B}_r = \left[\begin{array}{cccc}\mathbf{b}_{r,1} & \mathbf{b}_{r,2} & \cdots & \mathbf{b}_{r,L_r}\end{array}\right] \in \mathbb{C}^{J\times L_r}$, $\mathbf{C} \in \mathbb{C}^{K\times R}$, subject of course to the inherent ambiguity resulting from the fact that only the product $\mathbf{A}_r\mathbf{B}_r^{\rm T}$ 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]

$$\mathbf{X}_{(1)}^{\rm T} = (\mathbf{B}\odot\mathbf{C})\mathbf{A}^{\rm T} \triangleq \mathbf{P}\mathbf{A}^{\rm T}, \tag{5}$$

$$\mathbf{X}_{(2)}^{\rm T} = (\mathbf{C}\odot\mathbf{A})\mathbf{B}^{\rm T} \triangleq \mathbf{Q}\mathbf{B}^{\rm T}, \tag{6}$$

$$\mathbf{X}_{(3)}^{\rm T} = \left[\begin{array}{cccc}(\mathbf{A}_1 \odot_{\rm c} \mathbf{B}_1)\mathbf{1}_{L_1} & \cdots & (\mathbf{A}_R \odot_{\rm c} \mathbf{B}_R)\mathbf{1}_{L_R}\end{array}\right]\mathbf{C}^{\rm T} \triangleq \mathbf{S}\mathbf{C}^{\rm T}. \tag{7}$$

In this paper, we follow a Bayesian approach to address the above problem, starting from overestimates of R and L_r, $r = 1, 2, \ldots, R$.

III. The Proposed Bayesian Model

Let R and the L_rs be overestimated to $R_{\rm ini}$ and $L_{\rm ini}$, respectively. We intend to place heavy-tailed distributions, known for their sparsity-inducing effect, over the columns of $\mathbf{A}_r$s, $\mathbf{B}_r$s, and $\mathbf{C}$ in a way that implicitly implements a regularization analogous to that of the BTD-HIRLS method [3]. 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 ARD framework and building upon ideas of SBL [7], [26], the priors are assigned via a 3-level hierarchy of conjugate prior distributions outlined next.

The likelihood function, which encodes the underlying causal relation between the data and the latent variables, can be written in three equivalent forms, with respect to (w.r.t.) the three unfoldings of $\boldsymbol{\mathcal{Y}}$ (cf. (5), (6), (7)), as follows:

$$p(\mathbf{Y}_{(1)}^{\rm T} \mid \mathbf{A},\mathbf{B},\mathbf{C},\beta) = \prod_{i=1}^{I} p(\boldsymbol{y}_{(1)i} \mid \mathbf{A},\mathbf{B},\mathbf{C},\beta) = \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{y}_{(1)i} \mid \mathbf{P}\boldsymbol{a}_i, \beta^{-1}\mathbf{I}_{JK}), \tag{8}$$

$$p(\mathbf{Y}_{(2)}^{\rm T} \mid \mathbf{A},\mathbf{B},\mathbf{C},\beta) = \prod_{j=1}^{J} p(\boldsymbol{y}_{(2)i} \mid \mathbf{A},\mathbf{B},\mathbf{C},\beta) = \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{y}_{(2)j} \mid \mathbf{Q}\boldsymbol{b}_j, \beta^{-1}\mathbf{I}_{IK}), \tag{9}$$

$$p(\mathbf{Y}_{(3)}^{\rm T} \mid \mathbf{A},\mathbf{B},\mathbf{C},\beta) = \prod_{k=1}^{K} p(\boldsymbol{y}_{(3)k} \mid \mathbf{A},\mathbf{B},\mathbf{C},\beta) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{y}_{(3)k} \mid \mathbf{S}\boldsymbol{c}_k, \beta^{-1}\mathbf{I}_{IJ}), \tag{10}$$

where β is the noise precision (i.e., the inverse of the noise variance) and $\boldsymbol{a}_i, \boldsymbol{b}_j, \boldsymbol{c}_k$ and $\boldsymbol{y}_{(1)i}, \boldsymbol{y}_{(2)j}, \boldsymbol{y}_{(3)k}$ are the ith, jth, kth rows of $\mathbf{A},\mathbf{B},\mathbf{C}$ and $\mathbf{Y}_{(1)}$, $\mathbf{Y}_{(2)}$, $\mathbf{Y}_{(3)}$, respectively, in column form. The matrices $\mathbf{A},\mathbf{B},\mathbf{C}$ are considered as unobserved 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,

$$p(\mathbf{A} \mid \mathbf{t},\boldsymbol{\zeta},\beta) = \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{a}_i \mid \mathbf{0}, \beta^{-1}\mathbf{T}^{-1}(\mathbf{Z}^{-1} \otimes \mathbf{I}_{L_{\rm ini}}), \tag{11}$$

$$p(\mathbf{B} \mid \mathbf{t},\boldsymbol{\zeta},\beta) = \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{b}_j \mid \mathbf{0}, \beta^{-1}\mathbf{T}^{-1}(\mathbf{Z}^{-1} \otimes \mathbf{I}_{L_{\rm ini}})), \tag{12}$$

$$p(\mathbf{C} \mid \boldsymbol{\zeta},\beta) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{c}_k \mid \mathbf{0}, \beta^{-1}\mathbf{Z}^{-1}), \tag{13}$$

where $\mathbf{T} = \text{diag}(\mathbf{t})$ with $\mathbf{t} \in \mathbb{R}^{L_{\rm ini}R_{\rm ini}\times 1}$ and $\mathbf{Z} = \text{diag}(\boldsymbol{\zeta})$, $\boldsymbol{\zeta} \in \mathbb{R}^{R_{\rm ini}\times 1}$. Note that the priors of $\mathbf{A}$ and $\mathbf{B}$ are zero-mean with a *common covariance matrix*, which is essentially formed by the product of the inverses of the diagonal precision matrices $\mathbf{Z}$ and $\mathbf{T}$. This particular selection is critical from an implicit regularization perspective, since it induces *identical sparsity patterns* over columns/sub-blocks of $\mathbf{A}$ and $\mathbf{B}$. More specifically, the use of products of precision hyperparameters

in (11) and (12) serves our aim to impose at the same time two types of sparsity (i.e., block and column sparsity) on the factors $\mathbf{A}$ and $\mathbf{B}$.[2] In addition, by assigning the same precision hyperparameters $\mathbf{Z}$ to the columns of $\mathbf{C}$, we intend to achieve simultaneous elimination of columns $\mathbf{c}_r$'s of $\mathbf{C}$ and their corresponding blocks $\mathbf{A}_r$'s and $\mathbf{B}_r$'s. In the next section, we will see how the posterior covariance matrices of the latent BTD factors will determine the redundant block terms and columns of $\mathbf{A}_r, \mathbf{B}_r$s after the inference process. We can thus claim that, by combining $\mathbf{Z}$ and $\mathbf{T}$ as in the proposed priors (11)–(13), we may realize the two components of the regularizer in (3). Namely, sufficiently large values of ζ_r and its respective $t_{r,l}$'s will lead the rth column of $\mathbf{C}$ (cf. (13)) and the entire set of the redundant L_{ini} columns of sub-matrices $\mathbf{A}_r, \mathbf{B}_r$ (cf. (11), (12)) to zero, acting like the outer sum of square roots in (3). Moreover, the superfluous lth columns of the "surviving" $\mathbf{A}_r, \mathbf{B}_r$ are *jointly* forced to be zero when the value of $t_{r,l}$ becomes sufficiently large (cf. (11), (12)). Hence $\mathbf{T}$ plays a role similar to that of the inner sum of square roots of the regularizer in (3). Interestingly, $\mathbf{Z}$ and $\mathbf{T}$ are learned from data, thus providing a compelling way to perform BTD model selection.

At the second level of the hierarchy of priors, IG priors are assigned over $\mathbf{t}$ and $\boldsymbol{\zeta}$,

$$p(\mathbf{t}) = \prod_{r=1}^{R_{\text{ini}}} \prod_{l=1}^{L_{\text{ini}}} \mathcal{IG}\left(t_{r,l} \,\middle|\, \frac{I+J+1}{2}, \frac{\delta_{r,l}}{2}\right), \tag{14}$$

$$p(\boldsymbol{\zeta}) = \prod_{r=1}^{R_{\text{ini}}} \mathcal{IG}\left(\zeta_r \,\middle|\, \frac{(I+J)L_{\text{ini}}+K+1}{2}, \frac{\rho_r}{2}\right), \tag{15}$$

leading to hierarchical Gaussian-IG priors for $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$.[3] $\delta_{r,l}$ and ρ_r are the scale parameters of the distributions over $t_{r,l}$ and ζ_r, respectively. To be able to also infer these parameters from the data, we define a third hierarchical level that involves Gamma prior distributions, namely,

$$p(\delta_{r,l}) = \mathcal{G}(\delta_{r,l} \mid \psi, \tau), \tag{16}$$

$$p(\rho_r) = \mathcal{G}(\rho_r \mid \mu, \nu), \tag{17}$$

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 as the priors [26]. Finally, we assign a Gamma distribution to the noise precision β as follows:

$$p(\beta) = \mathcal{G}(\beta \mid \kappa, \theta). \tag{18}$$

Similarly to the hyperparameters of the variables δ and ρ, κ and θ are being set to small positive values rendering the prior non-informative, in the sense that the influence of the prior upon conditioning on the data and the inference process becomes negligible.

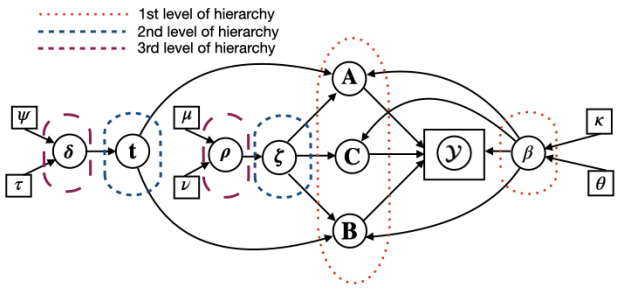


Fig. 2. The proposed Bayesian model.

The adopted Bayesian model is depicted in Fig. 2 in the form of a graphical model (with the meaning of $\boldsymbol{\delta}, \boldsymbol{\rho}$ being obvious) that manifests the causal relationships of the involved random variables. The proposed 3-level hierarchy of priors leads to a heavy-tailed distribution over the columns $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, thus allowing for simultaneously learning the latent factors of the BTD model and revealing their ranks. Note that the joint marginal pdf of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ resulting from the hierarchical distributions assigned to the latent factors and their hyperparameters cannot be analytically obtained due to the complexity of the model. Namely, the interrelation of the variables $\mathbf{A}, \mathbf{B}, \mathbf{C}$ with both $\mathbf{t}$ and $\boldsymbol{\zeta}$ renders the derivation of their joint pdf an infeasible task. In an effort to provide an insight into the heavy-tailed properties of the distributions assigned over the columns of $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, we give in Section V the analytical expression of the joint marginal pdf for a similar but slightly "relaxed" hierarchical Bayesian model. We would like to stress again at this point that the model described in this section allows us to follow a hyperparameter tuning-free approach since all involved parameters are treated as random variables. The way this is done is detailed next.

IV. Approximate Posterior Inference

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

$$p(\boldsymbol{\Theta} \mid \boldsymbol{\mathcal{Y}}) = \frac{p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta})}{\int p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta}) d\boldsymbol{\Theta}}. \tag{19}$$

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 (19). The idea is to approximate the posterior by a distribution which is as close as possible to the exact posterior in terms of the KL divergence [12]. 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

[2] An analogous idea, namely expressing the goal of sparsity enforcement as product of precision hyperparameters, has also appeared independently in, e.g., [17] in the context of low-rank Tucker decomposition, and in our earlier work [28] on low-rank matrix factorization with one factor being sparse.

[3] Other members of the EP-GIG family (i.e., for other values of q in [27]) might be also considered to serve as the priors of the BTD factors. Such a study, however, and the possible gains or losses from such choices, are beyond the scope of this paper and can be included in future related work.

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

$$q(\boldsymbol{\Theta}) = q(\beta)\prod_{i=1}^{I} q(\boldsymbol{a}_i)\prod_{j=1}^{J} q(\boldsymbol{b}_j)\prod_{k=1}^{K} q(\boldsymbol{c}_k)\times \prod_{r=1}^{R_{\text{ini}}}\prod_{l=1}^{L_{\text{ini}}} q(t_{r,l})q(\delta_{r,l})\prod_{r=1}^{R_{\text{ini}}} q(\zeta_r)q(\rho_r). \quad (20)$$

Denoting the individual variables above by $\boldsymbol{\theta}_i$, the corresponding VI-based posteriors are known to satisfy [12]

$$q(\boldsymbol{\theta}_i) = \frac{\exp(\langle \ln(p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta}))\rangle_{i\neq j})}{\int \exp(\langle \ln(p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta}))\rangle_{i\neq j} d\boldsymbol{\theta}_i}, \quad (21)$$

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 (21) a block coordinate ascent approach is taken, employing the cyclic update rule, namely solving for $q(\boldsymbol{\theta}_i)$ given $q(\boldsymbol{\theta}_j)$, $j \neq i$ and continuing for all i in a cyclic manner. More specifically, from (21) and using the expression for the likelihood which is based on the mode-1 unfolding of $\boldsymbol{\mathcal{Y}}$ (cf. (8)) the posterior distribution of $\boldsymbol{a}_i$ turns out to be

$$q(\boldsymbol{a}_i) = \mathcal{N}(\langle \boldsymbol{a}_i\rangle, \Sigma_{\boldsymbol{a}}), \quad (22)$$

with[4]

$$\langle \boldsymbol{a}_i\rangle = \langle\beta\rangle\Sigma_{\boldsymbol{a}}\langle \mathbf{P}\rangle^{\mathrm{T}}\boldsymbol{y}_{(1)i}, \quad (23)$$

$$\Sigma_{\boldsymbol{a}} = \langle\beta\rangle^{-1}(\langle \mathbf{P}^{\mathrm{T}}\mathbf{P}\rangle + \langle \mathbf{T}\rangle(\langle \mathbf{Z}\rangle \otimes \mathbf{I}_{L_{\text{ini}}}))^{-1}, \quad (24)$$

where $\langle\cdot\rangle$ denotes expectation w.r.t the posterior of the involved variable. Now, by employing (9), the posterior of $\boldsymbol{b}_j$ results in an analogous manner as:

$$q(\mathbf{b}_j) = \mathcal{N}(\langle \boldsymbol{b}_j\rangle, \Sigma_{\boldsymbol{b}}), \quad (25)$$

with

$$\langle \boldsymbol{b}_j\rangle = \langle\beta\rangle\Sigma_{\boldsymbol{b}}\langle \mathbf{Q}\rangle^{\mathrm{T}}\boldsymbol{y}_{(2)j} \quad (26)$$

$$\Sigma_{\boldsymbol{b}} = \langle\beta\rangle^{-1}(\langle \mathbf{Q}^{\mathrm{T}}\mathbf{Q}\rangle + \langle \mathbf{T}\rangle(\langle \mathbf{Z}\rangle \otimes \mathbf{I}_{L_{\text{ini}}}))^{-1}. \quad (27)$$

Concluding the first level of the hierarchy, the posterior of $\boldsymbol{c}_k$ is

$$q(\boldsymbol{c}_k) = \mathcal{N}(\langle \boldsymbol{c}_k\rangle, \Sigma_{\boldsymbol{c}}), \quad (28)$$

with

$$\langle \boldsymbol{c}_k\rangle = \langle\beta\rangle\Sigma_{\boldsymbol{c}}\langle \mathbf{S}\rangle^{\mathrm{T}}\boldsymbol{y}_{(3)k} \quad (29)$$

$$\Sigma_{\boldsymbol{c}} = \langle\beta\rangle^{-1}(\langle \mathbf{S}^{\mathrm{T}}\mathbf{S}\rangle + \langle \mathbf{Z}\rangle)^{-1}. \quad (30)$$

Next, the approximate posteriors of the variables belonging to the second level of hierarchy are given. Following similar arguments with [28], the posterior of $t_{r,l}$ turns out to be a GIG pdf,

$$q(t_{r,l}) = \mathcal{GIG}\left(t_{r,l}\,\middle|\, -\frac{1}{2}, \langle\beta\rangle\langle\zeta_r\rangle(\langle \mathbf{a}_{r,l}^{\mathrm{T}}\mathbf{a}_{r,l}\rangle + \langle \mathbf{b}_{r,l}^{\mathrm{T}}\mathbf{b}_{r,l}\rangle), \langle\delta_{r,l}\rangle\right) \quad (31)$$

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

with mean

$$\langle t_{r,l}\rangle = \sqrt{\frac{\langle\delta_{r,l}\rangle}{\langle\beta\rangle\langle\zeta_r\rangle(\langle \mathbf{a}_{r,l}^{\mathrm{T}}\mathbf{a}_{r,l}\rangle + \langle \mathbf{b}_{r,l}^{\mathrm{T}}\mathbf{b}_{r,l}\rangle)}}, \quad (32)$$

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

$$\langle \mathbf{A}^{\mathrm{T}}\mathbf{A}\rangle = \langle \mathbf{A}\rangle^{\mathrm{T}}\langle \mathbf{A}\rangle + I\Sigma_{\boldsymbol{a}} \quad (33)$$

and

$$\langle \mathbf{B}^{\mathrm{T}}\mathbf{B}\rangle = \langle \mathbf{B}\rangle^{\mathrm{T}}\langle \mathbf{B}\rangle + J\Sigma_{\boldsymbol{b}}, \quad (34)$$

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

$$\langle\zeta_r\rangle = \sqrt{\frac{\langle\rho_r\rangle}{\langle\beta\rangle(\sum_{l=1}^{L_{\text{ini}}}\langle t_{r,l}\rangle(\langle \mathbf{a}_{r,l}^{\mathrm{T}}\mathbf{a}_{r,l}\rangle + \langle \mathbf{b}_{r,l}^{\mathrm{T}}\mathbf{b}_{r,l}\rangle) + \langle \mathbf{c}_r^{\mathrm{T}}\mathbf{c}_r\rangle)}} \quad (35)$$

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

$$\langle \mathbf{C}^{\mathrm{T}}\mathbf{C}\rangle = \langle \mathbf{C}\rangle^{\mathrm{T}}\langle \mathbf{C}\rangle + K\Sigma_{\boldsymbol{c}}. \quad (36)$$

In addition, by employing the pdfs of $t_{r,l}$ and ζ_r, the expectations $\langle\frac{1}{t_{r,l}}\rangle$ and $\left\langle\frac{1}{\zeta_r}\right\rangle$, required in the posteriors at the third level of hierarchy, can be expressed as

$$\left\langle\frac{1}{t_{r,l}}\right\rangle = \frac{1}{\langle\delta_{r,l}\rangle} + \frac{1}{\langle t_{r,l}\rangle}, \quad \left\langle\frac{1}{\zeta_r}\right\rangle = \frac{1}{\langle\rho_r\rangle} + \frac{1}{\langle\zeta_r\rangle}. \quad (37)$$

Finally, it can be shown (as in [28]) that, at the third level of hierarchy, the approximate posteriors of $\delta_{r,l}, \rho_r$ and β are Gamma distributions with $\langle\delta_{r,l}\rangle$, $\langle\rho_r\rangle$ and $\langle\beta\rangle$ given in Algorithm 1, where the resulting *Bayesian-BTD (BBTD)* algorithm is summarized. The rest of the first- and second-order statistics that are required in the algorithm implementation are computed as in Table I, based on the assumption of statistically independent $\mathbf{A}, \mathbf{B}, \mathbf{C}$ (cf. (20)) and making use of the identities for the Grammians of Khatri-Rao products proved in [3, Appendix C].

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 $\langle\mathbf{A}\rangle, \langle\mathbf{B}\rangle$, as detailed in Algorithm 1. 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.

The algorithm can be randomly initialized and, as empirically demonstrated in Section VI, it converges fast and is very robust to initialization. Moreover, in view of its mean-field VI nature, the method is guaranteed to converge to a stationary point of the KL divergence function.

As far as the computational complexity of the algorithm is concerned, the computational cost of a BBTD iteration is similar to that of BTD-HIRLS (cf. [3, Appendix C]), with $\mathcal{O}((I+J)L^2+K)R^2)$ extra multiplications required to compute $\langle t_{r,l}\rangle$, $\langle\delta_{r,l}\rangle$, $\langle\zeta_r\rangle$ and $\langle\rho_r\rangle$. $\mathcal{O}(IJK + IJKLR + I(LR)^2 + (LR)^3 + LR + R)$ additional multiplications are needed in the computation of $\langle\beta\rangle$. Therefore, as in BTD-HIRLS, and for the more realistic case of tensors with dimensions much larger than R and L, the number of multiplications required per iteration of BBTD is $\mathcal{O}(IJKLR)$, i.e.,

Algorithm 1: The BBTD algorithm

Data: $\mathcal{Y}, R_{\text{ini}}, L_{\text{ini}}$
Result: $\hat{R}$, $\hat{L}_r, r = 1, 2, \ldots, \hat{R}$, $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$
Initialize $\langle \mathbf{B} \rangle, \langle \mathbf{C} \rangle, \langle \beta \rangle, \langle \mathbf{T} \rangle, \langle \mathbf{Z} \rangle, \langle \boldsymbol{\delta} \rangle, \langle \boldsymbol{\rho} \rangle$;
repeat
  $\Sigma_{\boldsymbol{a}} \leftarrow \langle \beta \rangle^{-1}(\langle \mathbf{P}^{\mathrm{T}}\mathbf{P} \rangle + \langle \mathbf{T} \rangle(\langle \mathbf{Z} \rangle \otimes \mathbf{I}_{L_{\text{ini}}}))^{-1}$;
  $\langle \mathbf{A} \rangle \leftarrow \langle \beta \rangle \mathbf{Y}_{(1)} \langle \mathbf{P} \rangle \Sigma_{\boldsymbol{a}}$;
  $\Sigma_{\boldsymbol{b}} \leftarrow \langle \beta \rangle^{-1}(\langle \mathbf{Q}^{\mathrm{T}}\mathbf{Q} \rangle + \langle \mathbf{T} \rangle(\langle \mathbf{Z} \rangle \otimes \mathbf{I}_{L_{\text{ini}}}))^{-1}$;
  $\langle \mathbf{B} \rangle \leftarrow \langle \beta \rangle \mathbf{Y}_{(2)} \langle \mathbf{Q} \rangle \Sigma_{\boldsymbol{b}}$;
  $\Sigma_{\boldsymbol{c}} \leftarrow \langle \beta \rangle^{-1}(\langle \mathbf{S}^{\mathrm{T}}\mathbf{S} \rangle + \langle \mathbf{Z} \rangle)^{-1}$;
  $\langle \mathbf{C} \rangle \leftarrow \langle \beta \rangle \mathbf{Y}_{(3)} \langle \mathbf{S} \rangle \Sigma_{\boldsymbol{c}}$;
  for $r = 1, 2, \ldots, R_{\text{ini}}$ **do**
    for $l = 1, 2, \ldots, L_{\text{ini}}$ **do**
      $\langle t_{r,l} \rangle \leftarrow \sqrt{\frac{\langle \delta_{r,l} \rangle}{\langle \beta \rangle \langle \zeta_r \rangle (\langle \mathbf{a}_{r,l}^{\mathrm{T}} \mathbf{a}_{r,l} \rangle + \langle \mathbf{b}_{r,l}^{\mathrm{T}} \mathbf{b}_{r,l} \rangle)}}$;
      $\left\langle \frac{1}{t_{r,l}} \right\rangle \leftarrow \frac{1}{\langle \delta_{r,l} \rangle} + \frac{1}{\langle t_{r,l} \rangle}$;
      $\langle \delta_{r,l} \rangle \leftarrow \frac{2\psi + I + J + 1}{2\tau + \langle \frac{1}{t_{r,l}} \rangle}$;
    end
  end
  for $r = 1, 2, \ldots, R_{\text{ini}}$ **do**
    $\langle \zeta_r \rangle \leftarrow \sqrt{\frac{\langle \rho_r \rangle}{\langle \beta \rangle [\sum_{l=1}^{L_{\text{ini}}} \langle t_{r,l} \rangle (\langle \mathbf{a}_{r,l}^{\mathrm{T}} \mathbf{a}_{r,l} \rangle + \langle \mathbf{b}_{r,l}^{\mathrm{T}} \mathbf{b}_{r,l} \rangle) + \langle \mathbf{c}_r^{T} \mathbf{c}_r \rangle]}}$;
    $\left\langle \frac{1}{\zeta_r} \right\rangle \leftarrow \frac{1}{\langle \rho_r \rangle} + \frac{1}{\langle \zeta_r \rangle}$;
    $\langle \rho_r \rangle \leftarrow \frac{2\mu + (I+J)L_{\text{ini}} + K + 1}{2\nu + \langle \frac{1}{\zeta_r} \rangle}$;
  end
  $\langle \beta \rangle \leftarrow (2\kappa + (I+J)L_{\text{ini}}R_{\text{ini}} + KR_{\text{ini}} + IJK)/(2\theta + \left\langle \left\| \mathbf{Y}_{(1)}^{\mathrm{T}} - \mathbf{P}\mathbf{A}^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \right\rangle + \sum_{r=1}^{R_{\text{ini}}} \langle \zeta_r \rangle [\sum_{l=1}^{L_{\text{ini}}} \langle t_{r,l} \rangle (\langle \mathbf{a}_{r,l}^{\mathrm{T}} \mathbf{a}_{r,l} \rangle + \langle \mathbf{b}_{r,l}^{\mathrm{T}} \mathbf{b}_{r,l} \rangle) + \langle \mathbf{c}_r^{\mathrm{T}} \mathbf{c}_r \rangle])$;
until *convergence*;
$\mathcal{I} \leftarrow \{i \in \{1, 2, \ldots, R_{\text{ini}}\} \mid$ ith column of $\langle \mathbf{C} \rangle$ is of non-negligible energy$\}$;
$\hat{R} \leftarrow |\mathcal{I}|$;
$\hat{\mathbf{C}} \leftarrow \langle \mathbf{C} \rangle(:, \mathcal{I})$;
Let the elements of $\mathcal{I}$ be sorted in increasing order as $i_1, i_2, \ldots, i_{\hat{R}}$;
for $r = 1, 2, \ldots, \hat{R}$ **do**
  $\mathcal{I}_r \leftarrow \{l \in \{1, 2, \ldots, L_{\text{ini}}\} \mid$ lth column of $\langle \mathbf{A} \rangle_r \triangleq \langle \mathbf{A} \rangle(:, (i_r - 1)L_{\text{ini}} + 1 : i_r L_{\text{ini}})$ is of non-negligible energy$\}$
  $= \{l \in \{1, 2, \ldots, L_{\text{ini}}\} \mid$ lth column of $\langle \mathbf{B} \rangle_r \triangleq \langle \mathbf{B} \rangle(:, (i_r - 1)L_{\text{ini}} + 1 : i_r L_{\text{ini}})$ is of non-negligible energy$\}$;
  $\hat{L}_r \leftarrow |\mathcal{I}_r|$;
  $\hat{\mathbf{A}}_r \leftarrow \langle \mathbf{A} \rangle_r(:, \mathcal{I}_r)$, $\hat{\mathbf{B}}_r \leftarrow \langle \mathbf{B} \rangle_r(:, \mathcal{I}_r)$;
end
$\hat{\mathbf{A}} \leftarrow [\ \hat{\mathbf{A}}_1 \quad \hat{\mathbf{A}}_2 \quad \cdots \quad \hat{\mathbf{A}}_{\hat{R}}\ ]$;
$\hat{\mathbf{B}} \leftarrow [\ \hat{\mathbf{B}}_1 \quad \hat{\mathbf{B}}_2 \quad \cdots \quad \hat{\mathbf{B}}_{\hat{R}}\ ]$;

TABLE I
FIRST- AND SECOND-ORDER STATISTICS REQUIRED IN THE BBTD ALGORITHM

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

of the same order as the computational cost of a BTD-HIRLS iteration [3]. Clearly, R and L here refer to their overestimates, R_{ini} and L_{ini}, respectively. The cost can be reduced if *pruning* of the nulled columns of $\langle \mathbf{C} \rangle$ and the corresponding blocks of $\langle \mathbf{A} \rangle$ and $\langle \mathbf{B} \rangle$ in the course of the algorithm is included. For the sake of the simplicity of presentation, this is only performed in Algorithm 1 at the end of the iterative inference.

V. Alternative Bayesian Models and Their Regularization-based Counterparts

In this section, we present two alternative Bayesian models, which can be viewed as simplified versions of the model introduced in Section III. The main goal here is to manifest the role that specific aspects of the adopted model (e.g., the number of the levels in the hierarchy) play in the regularization that is induced to the latent BTD factors at inference time. Both models presented next assume the same likelihood function as the more composite model presented previously. That being said, the main difference between the two models lies in the priors placed over $\mathbf{A}, \mathbf{B}, \mathbf{C}$, as detailed next.

a) Model I: This model consists of a single level of hierarchy, with Gaussian priors being assigned to the rows of $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$:

$$p(\mathbf{A}|t, \beta) = \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{a}_i \mid \mathbf{0}, \beta^{-1} t^{-1} \mathbf{I}_{L_{\text{ini}} R_{\text{ini}}}), \quad (38)$$

$$p(\mathbf{B}|t, \beta) = \prod_{j=1}^{J} \mathcal{N}(\boldsymbol{b}_j \mid \mathbf{0}, \beta^{-1} t^{-1} \mathbf{I}_{L_{\text{ini}} R_{\text{ini}}}), \quad (39)$$

$$p(\mathbf{C}|t, \beta) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{c}_k \mid \mathbf{0}, \beta^{-1} t^{-1} \mathbf{I}_{R_{\text{ini}}}), \quad (40)$$

where t is now a deterministic parameter, intended to play the role of the regularization parameter in the associated deterministic regularization-based problem. The corresponding graphical model is given in Fig. 3-I. It is obviously a simplified, single-level version of the 3-level hierarchical model introduced in Section III.

To perform point estimation of $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, the corresponding MAP estimator is derived next. The joint posterior pdf of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ can be expressed as follows:

$$p(\mathbf{A}, \mathbf{B}, \mathbf{C}|\mathcal{Y}, \beta) \propto p(\mathcal{Y}|\mathbf{A}, \mathbf{B}, \mathbf{C}, \beta) p(\mathbf{A}, \mathbf{B}, \mathbf{C}|\beta). \quad (41)$$

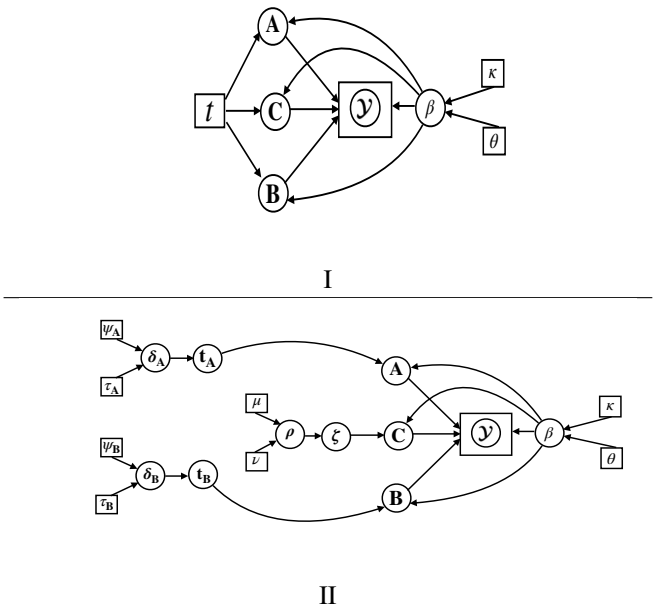


Fig. 3. Alternative probabilistic models.

Due to the Gaussianity of the noise, the likelihood function $p(\boldsymbol{\mathcal{Y}}|\mathbf{A},\mathbf{B},\mathbf{C},\beta)$ can be written from (4) as

$$p(\boldsymbol{\mathcal{Y}}|\mathbf{A},\mathbf{B},\mathbf{C},\beta) \propto \exp\left(-\frac{\beta}{2}\left\|\boldsymbol{\mathcal{Y}}-\sum_{r=1}^{R_{\text{ini}}}\mathbf{A}_r\mathbf{B}_r^{\text{T}}\circ\mathbf{c}_r\right\|_{\text{F}}^2\right). \tag{42}$$

In addition, from (38), (39), and (40), the prior distribution of $\mathbf{A},\mathbf{B}$ and $\mathbf{C}$ takes the following form

$$p(\mathbf{A},\mathbf{B},\mathbf{C}|\beta) \propto \exp\left[-\frac{\beta t}{2}\sum_{r=1}^{R_{\text{ini}}}\sum_{l=1}^{L_{\text{ini}}}(\|\mathbf{a}_{r,l}\|_2^2+\|\mathbf{b}_{r,l}\|_2^2)\right] \times \exp\left(-\frac{\beta t}{2}\sum_{r=1}^{R_{\text{ini}}}\|\mathbf{c}_r\|_2^2\right). \tag{43}$$

Combining (42) with (43) and taking the logarithm of their product we end up with the following MAP-type optimization problem:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}}\left\|\boldsymbol{\mathcal{Y}}-\sum_{r=1}^{R_{\text{ini}}}\mathbf{A}_r\mathbf{B}_r^{\text{T}}\circ\mathbf{c}_r\right\|_{\text{F}}^2+t\left(\|\mathbf{A}\|_{\text{F}}^2+\|\mathbf{B}\|_{\text{F}}^2+\|\mathbf{C}\|_{\text{F}}^2\right). \tag{44}$$

This implies that the regularizer induced by the single-level Gaussian priors favors smooth solutions in terms of the latent factors. This is deduced from the fact that $\mathbf{A},\mathbf{B}$ and $\mathbf{C}$ can be updated using an alternating minimization strategy which gives rise to ridge regression-type subproblems (as it is done in, e.g., [5]). In the light of this feature, no distinction between the columns of each of the factor matrices is being made and hence Model I is expected to have a weaker rank revelation effect than the one of the main model introduced in Section III.

b) Model II: With the latter observation in mind, we now introduce the second alternative Bayesian model whose graphical model is depicted in Fig. 3-II. Model II places heavy-tailed multi-parameter Laplace priors, known for their sparsity-inducing effect, over the columns of the factor matrices. This is implemented with the aid of a three-level hierarchy of priors. At the first level, Gaussian distributions are assigned to the factors, namely,

$$p(\mathbf{A}|\mathbf{t}_A,\beta) = \prod_{i=1}^{I}\mathcal{N}(\boldsymbol{a}_i\mid\mathbf{0},\beta^{-1}\mathbf{T}_A^{-1}), \tag{45}$$

$$p(\mathbf{B}|\mathbf{t}_B,\beta) = \prod_{j=1}^{J}\mathcal{N}(\boldsymbol{b}_j\mid\mathbf{0},\beta^{-1}\mathbf{T}_B^{-1}), \tag{46}$$

$$p(\mathbf{C}|\boldsymbol{\zeta},\beta) = \prod_{k=1}^{K}\mathcal{N}(\boldsymbol{c}_k\mid\mathbf{0},\beta^{-1}\mathbf{Z}^{-1}), \tag{47}$$

where $\mathbf{T}_A=\text{diag}(\mathbf{t}_A)$ and $\mathbf{T}_B=\text{diag}(\mathbf{t}_B)$ with $\mathbf{t}_A,\mathbf{t}_B\in\mathbb{R}^{L_{\text{ini}}R_{\text{ini}}\times 1}$ and $\mathbf{Z}=\text{diag}(\boldsymbol{\zeta})$, $\boldsymbol{\zeta}\in\mathbb{R}^{R_{\text{ini}}\times 1}$. Note that the key difference of this model with the one introduced in Section III and depicted in Fig. 2 is the use of different variables $\mathbf{t}_A,\mathbf{t}_B$ for enforcing column sparsity on $\mathbf{A}$ and $\mathbf{B}$. This is in contrast to the "coupling" of $\mathbf{A}_r,\mathbf{B}_r$ effected in BTD-HIRLS and the model of Fig. 2. Moreover, the parameters $\boldsymbol{\zeta}$ are now involved only in the prior of $\mathbf{C}$, which again "decouples" the third mode factor from the rest.

At the second level of the hierarchy, IG priors are placed over $\mathbf{t}_A,\mathbf{t}_B$ and $\boldsymbol{\zeta}$, namely

$$p(\mathbf{t}_A) = \prod_{r=1}^{R_{\text{ini}}}\prod_{l=1}^{L_{\text{ini}}}\mathcal{IG}\left(t_{A;r,l}\left|\frac{I+1}{2},\frac{\delta_{A;r,l}}{2}\right.\right), \tag{48}$$

$$p(\mathbf{t}_B) = \prod_{r=1}^{R_{\text{ini}}}\prod_{l=1}^{L_{\text{ini}}}\mathcal{IG}\left(t_{B;r,l}\left|\frac{J+1}{2},\frac{\delta_{B;r,l}}{2}\right.\right), \tag{49}$$

$$p(\boldsymbol{\zeta}) = \prod_{r=1}^{R_{\text{ini}}}\mathcal{IG}\left(\zeta_r\left|\frac{K+1}{2},\frac{\rho_r}{2}\right.\right), \tag{50}$$

where $\delta_{A;r,l},\delta_{B;r,l}$ and ρ_r are the scale parameters of the distributions over $t_{A;r,l},t_{B;r,l}$ and ζ_r, respectively. The third level involves Gamma priors over these variables, namely,

$$p(\delta_{A;r,l})=\mathcal{G}(\delta_{A;r,l}\mid\psi_A,\tau_A), \tag{51}$$

$$p(\rho_{A,r})=\mathcal{G}(\rho_{A,r}\mid\mu_A,\nu_A), \tag{52}$$

and similarly for $\delta_{B;r,l},\rho_{B,r}$.

This "decoupling" approach allows us to derive the MAP estimator for $\mathbf{A},\mathbf{B},\mathbf{C}$ and thus gain a deeper insight as to the regularization effect induced by the model. As explained earlier, a MAP-type problem cannot be derived for the main model presented in Section III due to the interrelation among different variables. Yet, the increased complexity of that model better captures the structure of BTD, as it is also empirically demonstrated in the experimental results. For Model II, the joint prior pdf of $\mathbf{A},\mathbf{B},\mathbf{C}$ can be computed from the following multiple integral

$$p(\mathbf{A},\mathbf{B},\mathbf{C}|\beta,\boldsymbol{\delta}_A,\boldsymbol{\delta}_B,\boldsymbol{\rho}) = \int p(\mathbf{A},\mathbf{B},\mathbf{C}|\beta,\mathbf{t}_A,\mathbf{t}_B,\boldsymbol{\zeta})\times p(\mathbf{t}_A|\boldsymbol{\delta}_A)p(\mathbf{t}_B|\boldsymbol{\delta}_B)p(\boldsymbol{\zeta}|\boldsymbol{\rho})\text{d}\mathbf{t}_A\text{d}\mathbf{t}_B\text{d}\boldsymbol{\zeta}, \tag{53}$$

where

$$p(\mathbf{A},\mathbf{B},\mathbf{C}|\beta,\mathbf{t}_A,\mathbf{t}_B,\boldsymbol{\zeta}) = \prod_{r=1}^{R_{\text{ini}}}\prod_{l=1}^{L_{\text{ini}}} p(\mathbf{a}_{r,l},\mathbf{b}_{r,l}|\beta,t_{r,l}) \times \prod_{r=1}^{R_{\text{ini}}} p(\mathbf{c}_r|\beta,\zeta_r). \tag{54}$$

After substituting (54) to (53) we get the expression for the joint prior distribution shown in (55) at the top of the next page. The integrals in (55) can be computed by working as in [28, Appendix B] whereby the joint prior pdf of $\mathbf{A},\mathbf{B},\mathbf{C}$ results as

$$p(\mathbf{A},\mathbf{B},\mathbf{C}|\beta,\boldsymbol{\delta},\boldsymbol{\rho}) \propto \exp\left[-\beta^{\frac{1}{2}}\sum_{r=1}^{R_{\text{ini}}}\sum_{l=1}^{L_{\text{ini}}}(\delta_{A;r,l}^{\frac{1}{2}}\|\mathbf{a}_{r,l}\|_2+\delta_{B;r,l}^{\frac{1}{2}}\|\mathbf{b}_{r,l}\|_2)\right] \times \exp\left(-\beta^{\frac{1}{2}}\sum_{r=1}^{R_{\text{ini}}}\rho_r^{\frac{1}{2}}\|\mathbf{c}_r\|_2\right), \tag{56}$$

which is a heavy-tailed multi-parameter multivariate Laplace distribution defined on the columns of $\mathbf{A},\mathbf{B}$ and $\mathbf{C}$. From (42) and (56), the MAP estimator of Model II is obtained from the solution of the following minimization problem

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \frac{\beta}{2}\left\|\boldsymbol{\mathcal{Y}}-\sum_{r=1}^{R_{\text{ini}}}\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}}\circ\mathbf{c}_r\right\|_{\mathrm{F}}^2 + \beta^{\frac{1}{2}}\left(\sum_{r=1}^{R_{\text{ini}}}\sum_{l=1}^{L_{\text{ini}}}(\delta_{A;r,l}^{\frac{1}{2}}\|\mathbf{a}_{r,l}\|_2+\delta_{B;r,l}^{\frac{1}{2}}\|\mathbf{b}_{r,l}\|_2)+\sum_{r=1}^{R_{\text{ini}}}\rho_r^{\frac{1}{2}}\|\mathbf{c}_r\|_2\right). \tag{57}$$

Remark: It should be noted that (57) bears a close resemblance to the deterministic criterion proposed in [33] and can thus be seen to offer a Bayesian interpretation thereof and suggest the corresponding Bayesian inference method as a probabilistic counterpart of the *Alternating Group Lasso (AGL)* algorithm developed in [33]. Given the correspondence of the main model, in Section III, with BTD-HIRLS, the comparison of these Bayesian methods presented in the next section complements in a way the comparative study of BTD-HIRLS and AGL previously reported [3].

VI. Simulation Results

In this section, we evaluate the effectiveness of the proposed Algorithm 1 in selecting and computing the appropriate BTD model for a given tensor, via simulations with both synthetic and real data. Its deterministic counterpart from [3] and the corresponding Bayesian inference methods resulting from Models I, II and referred to henceforth as BBTD model I and BBTD model II are included, for comparison purposes. In the last experiment, and in the hyperspectral image denoising problem, BBTD is also compared with the Bayesian CPD method that emanates from Model II as a special case with $L_r = 1$, $r = 1, 2, \ldots, R$.

A. Synthetic data experiments

In this part, we first test the BBTD method stated as Algorithm 1 in comparison to its alternatives. We also demonstrate its robustness to initialization and compare its ability to recover the correct ranks of the BTD model against BTD-HIRLS. The adopted figure of merit is the *Normalized Mean Squared Error (NMSE)* over block terms, defined as $\text{NMSE} = \sum_{r=1}^{R}\frac{\|\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}}\circ\mathbf{c}_r-\hat{\mathbf{A}}_r\hat{\mathbf{B}}_r^{\mathrm{T}}\circ\hat{\mathbf{c}}_r\|_{\mathrm{F}}^2}{\|\mathbf{A}_r\mathbf{B}_r^{\mathrm{T}}\circ\mathbf{c}_r\|_{\mathrm{F}}^2}$. As in [3], the Hungarian algorithm is employed to match the $\hat{R}$ estimated non-zero block terms with the true ones.

a) Performance comparison between the main model and Models I and II: In this experiment, we generate $18\times18\times10$ tensors $\boldsymbol{\mathcal{Y}}$ as in (4), with $R = 3$ and the L_rs set as $L_1 = 8, L_2 = 6$ and $L_3 = 4$. 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 $\text{SNR} = 10\log_{10}\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}^2/(\sigma^2\|\boldsymbol{\mathcal{N}}\|_{\mathrm{F}}^2)$ of 5 and 15 dB. Both R and all L_rs are overestimated as $R_{\text{ini}} = L_{\text{ini}} = 10$. Fig. 4 illustrates the best run in terms of the NMSE, obtained out of 10 random initializations of the algorithms. As it can be observed, the main model described in Section III performs comparably to Model II, which can be viewed as a relaxed version thereof. Notably, Algorithm 1 converges somewhat faster. It is worth noting that the algorithm associated with Model I exhibits a poorer performance – especially at low SNR – due to its inaptitude in dealing effectively with the over-parameterized regime when it comes to the BTD ranks. As opposed to Model I, both the main model and Model II, which use heavy-tailed priors on the latent BTD factors, show their efficacy in addressing the challenges incurred by the unawareness of BTD ranks and successfully model the tensors at both SNR values examined.

In the following, we focus on Algorithm 1, which better captures the structure of the BTD model. At the same time it makes use of fewer latent variables at the second and third level of hierarchy than those in Model II and hence we consider it as a more compact version of the latter.

b) Robustness to initialization: In an effort to see how robust the proposed BBTD algorithm is to initialization, we set SNR=15 dB and generate tensors as previously. We run 500 realizations of the experiment. For each, we apply the proposed BBTD and the BTD-HIRLS algorithms with 12 different random initializations. Fig. 5 shows the Empirical Cumulative Distribution Function (ECDF) of the obtained NMSE, where the ith curve from bottom to top corresponds to selecting the best out of i initializations, for $i = 1, 2, \ldots, 12$. It can be observed that BBTD is rather insensitive to initialization. Surprisingly, its performance is affected by random initialization even less than BTD-HIRLS, whose robustness has also been verified [3]. We thus have empirical evidence that only a small number of initializations suffices to estimate an accurate BTD model with the proposed BBTD algorithm.

c) Rank recovery: Here we use the same generative model described above for building data tensors $\boldsymbol{\mathcal{Y}}$ of dimensions $30\times30\times30$. Our objective is to assess the ability of BBTD to select the correct BTD model. For comparison purposes, we also employ the BTD-HIRLS algorithm, which

$$p(\mathbf{A},\mathbf{B},\mathbf{C}|\beta,\boldsymbol{\delta}_A,\boldsymbol{\delta}_B,\boldsymbol{\rho}) = \prod_{r=1}^{R_{\text{ini}}}\prod_{l=1}^{L_{\text{ini}}}\int_0^\infty p(\mathbf{a}_{r,l},\mathbf{b}_{r,l}|\beta,t_{A;r,l},t_{B;r,l})p(t_{A;r,l}|\delta_{A;r,l})p(t_{B;r,l}|\delta_{B;r,l})\mathrm{d}t_{A;r,l}\mathrm{d}t_{B;r,l}$$
$$\times\prod_{r=1}^{R_{\text{ini}}}\int_0^\infty p(\mathbf{c}_r|\beta,\zeta_r)p(\zeta_r|\rho_r)\mathrm{d}\zeta_r. \tag{55}$$

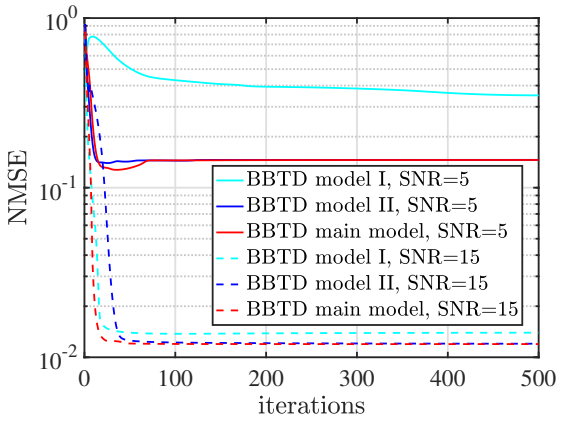


Fig. 4. NMSE vs. iterations for the proposed BBTD algorithm ('BBTD main model') and its variants ('BBTD model I', 'BBTD model II') at two SNR values.

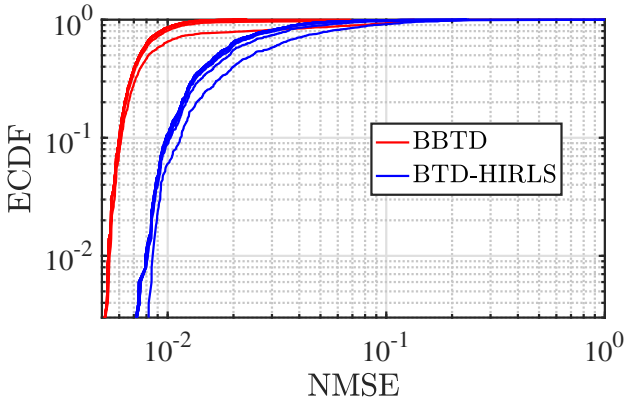


Fig. 5. Empirical Cumulative Distribution Function (ECDF) of NMSE obtained by BBTD and BTD-HIRLS from 500 independent runs. The ith curve from bottom to top corresponds to the result of selecting the best out of $i = 1, 2, \ldots, 12$ different initializations. SNR=15 dB.

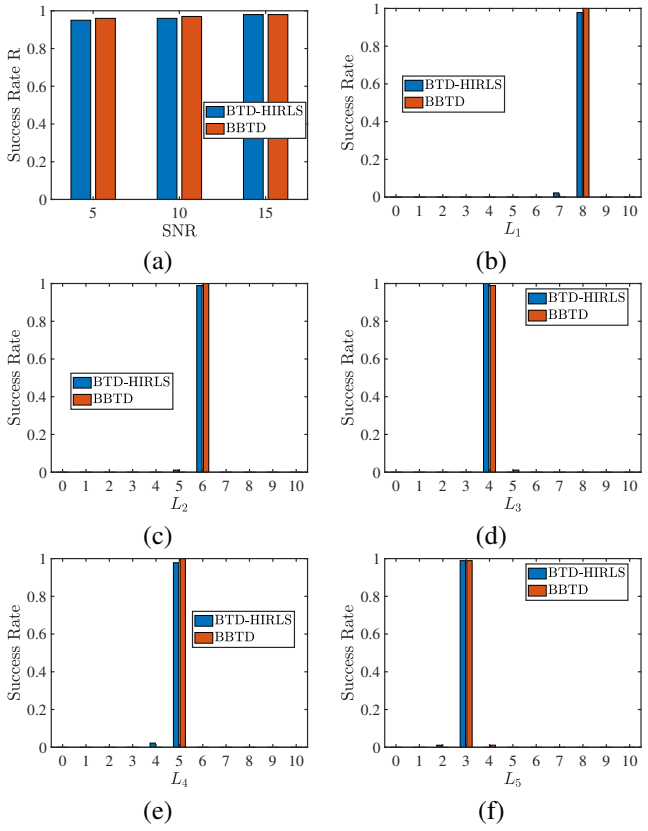


Fig. 6. Success rates of recovering (a) R and (b)–(f) L_rs for SNR=10 dB with the aid of the BBTD and BTD-HIRLS algorithms. Scenario A: $\min(I,J) > \sum_{r=1}^{R} L_r$.

has demonstrated high model selection ability in [3]. Two different scenarios are considered, that differ in the validity of the well-known sufficient BTD uniqueness condition of having full column rank $\mathbf{A},\mathbf{B}$ matrices and a $\mathbf{C}$ matrix with non-collinear columns [1].

Scenario A: In this scenario, we set $R = 5$ and the true L_rs are set to $L_1 = 8, L_2 = 6, L_3 = 4, L_4 = 5$ and $L_5 = 3$. This setting is favorable w.r.t. the above condition since $\min(I,J) > \sum_{r=1}^{R} L_r$. Fig. 6(a) shows the success rates of the recovery of R for SNR equal to 5, 10, and 15 dB. BBTD performs slightly better than BTD-HIRLS at 5 and 10 dB and has similar performance at 15 dB. Note that BTD-HIRLS required its regularization parameter to be finely tuned, whereas this is automatically performed in the BBTD algorithm, in a data-driven way. In Figs. 6(b)–(f), and restricting attention to those realizations where both algorithms have succeeded in recovering the true value of R, the success rates of recovering the block ranks L_r are depicted, at an SNR value of 10 dB. Observe that there is an almost 100% success for all R terms. These results provide empirical evidence of the competence of BBTD in this challenging, yet critical, task of inferring the correct model structure.

Scenario B: We now choose the following values for the block ranks, $L_1 = 8, L_2 = 6, L_3 = 8, L_4 = 6$ and $L_5 = 7$, for which $\min(I,J) < \sum_{r=1}^{R} L_r$ and hence the sufficient uniqueness condition is no longer satisfied. This experimental setting is therefore considered to be even more challenging than Scenario A. As shown in Fig. 7(a), BBTD performs comparably to the BTD-HIRLS in estimating R, with the latter being somewhat better at SNR=5 dB. It should, however, be reminded that this is the result of tuning the regularization parameter, a task that can be far from being easy in real-world applications. Moreover, a behavior similar to that in

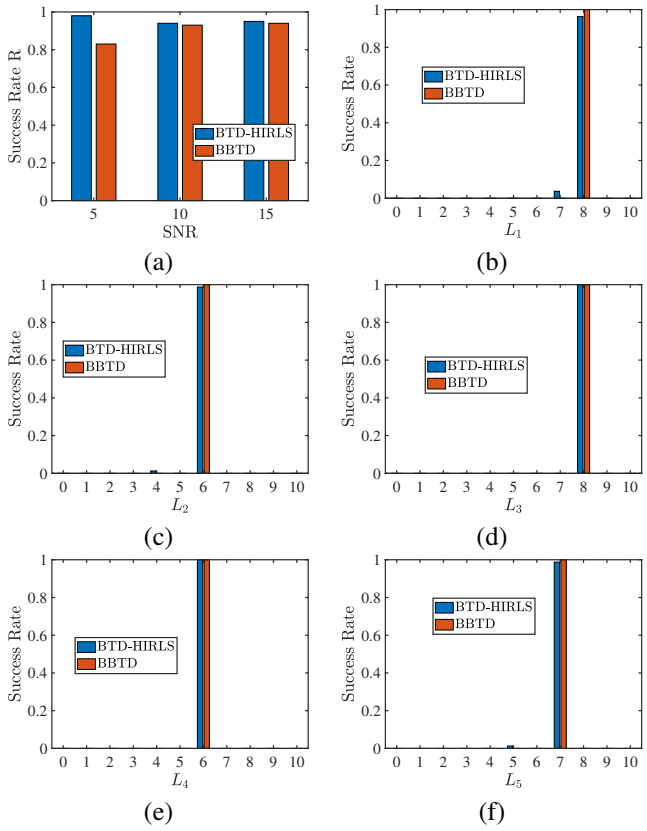


Fig. 7. Success rates of recovering (a) R and (b)–(f) L_rs for SNR=10 dB with the aid of the BBTD and BTD-HIRLS algorithms. Scenario B: $\min(I, J) < \sum_{r=1}^{R} L_r$.

Scenario A is observed when it comes to the success rates of recovering the L_rs at SNR=10 dB; see Figs. 7(b)–(f). BBTD is again slightly superior to BTD-HIRLS in carrying out this intricate task while enjoying the advantage of being completely automatic. Additional cases, for varying SNR values, ranks and tensor sizes, have been tested and the results were similar to those obtained above, showing the applicability of Algorithm 1 in a wide range of scenarios.

B. Real data experiment: Hyperspectral image denoising

Hyperspectral imagery (HSI) can be represented with the aid of 3-way tensors whose first two modes correspond to the spatial domain and the third one to the spectral domain. It is known that there is inherent correlation in both domains, which explains the fact that low-rank matrix and tensor representations have been widely adopted for numerous HSI processing tasks such as unmixing [5], [34] and restoration [35]. It should be emphasized that the very nature of HSI, accurately described by a linear mixing model [5], points to BTD as the most suitable choice of a decomposition model as compared to classical CPD. Indeed, the model structure and parameters are in a direct correspondence with the HSI constituents: the R matrices $\mathbf{E}_r$ can be interpreted as the abundance maps while $\mathbf{C}$ contains the endmember spectral signatures in its columns.

As an example of the application of our method in this context, we consider the problem of denoising hyperspectral images, and compare with the results of BTD-HIRLS and

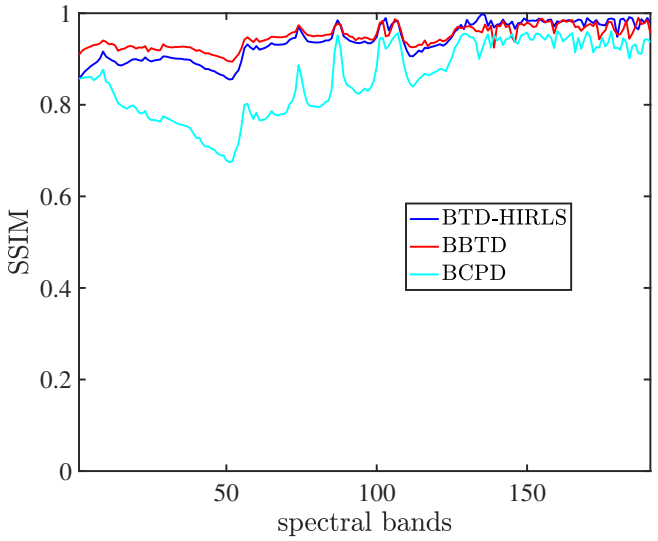
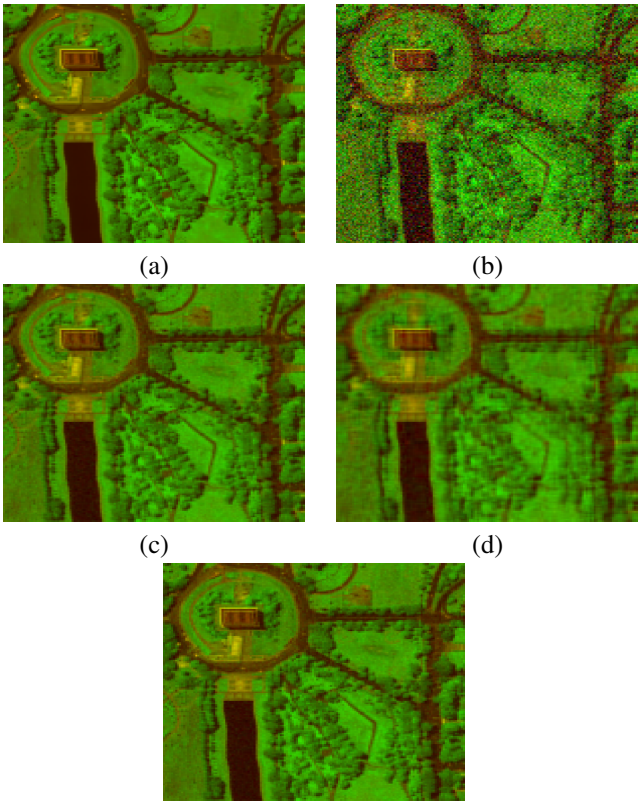


Fig. 8. SSIM of the hyperspectral images recovered by BBTD, BTD-HIRLS and BCPD.

a Bayesian CPD method resulting from BTD model II as a special case and referred to here as BCPD. We carry out denoising experiments on three different HSI datasets, namely a) the Washington DC Mall AVIRIS, b) the Salinas Valley, and c) the Indian Pines images.

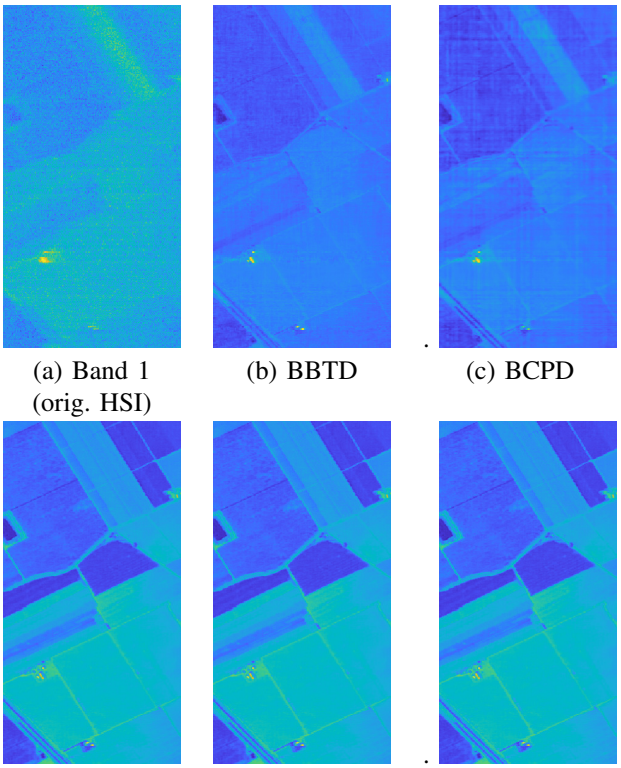
1) Washington DC Mall AVIRIS: We generate a noisy version of the Washington DC Mall AVIRIS image captured at $K = 191$ contiguous spectral bands in the 0.4 to 2.4μm region of the visible infrared spectrum [36]. In all three HSIs we add i.i.d. Gaussian noise and choosing its power so as to get SNR=5 dB. The size of the image at each spectral band is 150×150 pixels and hence the HSI cube can be seen as a $150 \times 150 \times 191$ tensor. Our objective is to suppress the noise by fitting a decomposition model to this tensor. Of course, the correct R and L_rs must also be estimated. To this end, they are overestimated as $R_{\text{ini}} = 50$ and $L_{\text{ini}} = 10$. Finally, we initialize the tensor rank of BCPD to $R_{\text{ini}}^{\text{BCPD}} = R_{\text{ini}} L_{\text{ini}} = 500$.

We compare the performance of the three methods both visually and in terms of the *Structural Similarity Index Measure (SSIM)*, a popular perceptual metric of the degradation of an image as perceived change in structural information. SSIM is defined for two image windows x, y as $\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$, where $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2$ are their mean averages and variances, respectively and σ_{xy} is their covariance. c_1, c_2 are small constants that are used for averting division by zero. BTD-HIRLS is again used for comparison purposes, with its regularization parameter being finely tuned in accordance with SSIM. As it can be seen in Fig. 8, the BBTD algorithm outperforms BTD-HIRLS, exhibiting higher or similar SSIM values over a wide range of spectral bands. The CPD model, resulting from the BCPD method, does not capture the low-rank structure of the HSI tensor equally well. The superior performance of the BTD-based algorithms as compared to the CPD one can be explained by the estimated ranks in each case. Specifically, the



(a) (b) (c) (d) (e)

Fig. 9. False color RGB images (made from bands 34, 64, and 135) of the Washington DC Mall AVIRIS hyperspectral image. (a) Original; (b) Noisy; Denoised with (c) BTD-HIRLS, (d) BCPD, and (e) the proposed BBTD algorithm.



(a) Band 1 (orig. HSI) (b) BBTD (c) BCPD

(d) Band 190 (orig. HSI) (e) BBTD (f) BCPD

Fig. 10. Reconstructed by BBTD and BCPD bands no. 1 and 190 of the Indian Pines hyperspectral image.

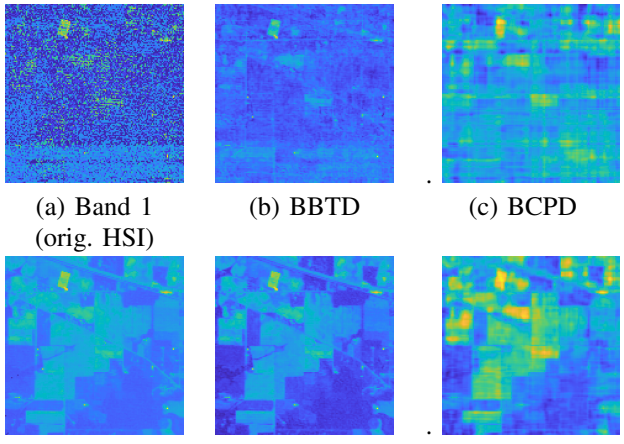
estimated $\hat{R}$ by both BBTD and BTD-HIRLS is 9. Based on the compelling interpretation that the BTD model offers when it comes to HSI decomposition [5], $\hat{R}$ corresponds to the number of endmembers (distinct materials) that exist in the depicted scene. That said, $\hat{R} = 9$ turns out to be in a good agreement with what is known in the HSI literature for the number of endmembers existing in the scene depicted by the Washington DC Mall AVIRIS HSI [36]. On the other hand, the BCPD estimate of the CPD rank is $\hat{R} = 32$, that is, it largely overestimates the number of endmembers in the scene. It should not be surprising that the CPD model is not able to provide an accurate tensor representation of the image, manifesting the limitations of the CPD representation in capturing the inherent structure of HSI.

For a visual comparison of the results of the three algorithms, Figs. 9(a) and (b) depict false color images of the true and the noisy image, respectively, while the BTD-HIRLS, BCPD and BBTD reconstruction results are respectively given in parts (c), (d) and (e) of the figure. The comparable performance of the two BTD methods observed in Fig. 8 is confirmed here by visual inspection. Moreover, as expected, BCPD provides clearly poorer results, with a blurring effect being clearly visible in the corresponding false color image.

2) Salinas Valley and Indian Pines: In the second real hyperspectral image denoising experiment, we use the Salinas Valley and Indian Pines datasets [16].

Salinas Valley is captured at 224 spectral bands ranging and its spatial resolution is 3.7 meters. The scene depicts an agricultural area with different vegetation species. In Fig. 10(a), we can see the noisy band 1 of the original HSI and the reconstructed images from BBTD (Fig. 10(b)) and BCPD (Fig. 10(c)). Clearly, BBTD is shown to be able to denoise the first band of the HSI, outperforming the BCPD, which is based on the CPD tensor decomposition. Moreover, both BBTD and BCPD reconstruct quite reliably the 190th band (Fig. 10(c)–(d)). It should be noted that the Salinas image is highly structured. Thus, BBTD, though initialized with $R_{\text{ini}} = 50$ and $L_{\text{ini}} = 10$, converges to a $\hat{R} = 25$. A low tensor rank is also imposed by BCPD, which is initialized with a CP rank $R_{\text{ini}}^{\text{BCPD}} = 500$ but converges to a tensor of CP rank equal to 84.

The Indian Pines HSI is captured at 145×145 pixels and 224 spectral reflectance bands in the wavelength range 0.4—2.5 μm. The scene contains agriculture, forest, as well as vegetation areas, highways, etc. Similarly to the Salinas Valley, a few of its bands are noisy and are usually removed in a preprocessing step before performing downstream tasks such as classification, clustering, etc. In this experiment, BBTD and BCPD are initialized as for the case of the Salinas Valley HSI, i.e., for BBTD $R_{\text{ini}} = 50$ and $L_{\text{ini}} = 10$, and for BCPD $R_{\text{ini}} = 500$. Again, both algorithms eliminate redundant components of the respective tensor models. Namely, BBTD finds a BTD model of the HSI with rank $\hat{R} = 20$ block terms. Moreover, BCPD converges to a CPD decomposition with



(a) Band 1 (orig. HSI) (b) BBTD (c) BCPD

(d) Band 10 (orig. HSI) (e) BBTD (f) BCPD

Fig. 11. Reconstructed by BBTD and BCPD bands 1 and 10 of the Indian Pines HSI hyperspectral image.

$\hat{R} = 80$ rank-1 terms. As is shown in Fig. 11, the BBTD reconstructs the 1st and 10th bands of the HSI providing a reliable denoised image. At the same time, it is shown that BBTD performs better that BCPD thus providing additional evidence when it comes to the superiority of BTD over CPD when it comes to modelling highly structured hyperspectral images.

VII. Conclusions

As a follow-up to our earlier work on BTD model selection and computation based on $\ell_{1,2}$ norm-based regularization, we developed *for the first time* in this paper a Bayesian method for the same problem, which completely relieves its user from having to tune a regularization parameter. The proposed fully-automatic variational inference scheme originates from a Bayesian probabilistic model designed to match perfectly with the BTD model structure and promote model selection through heavy-tailed prior distributions assigned to the BTD factors in the spirit of ARD and SBL. Two alternative simplified Bayesian models were also presented and their model selection properties were investigated by way of their individual joint posterior distribution maximization tasks. Extensive empirical results showed that the proposed algorithm is extremely robust to initialization, converges fast and its model selection ability is comparable to that of its regularization-based counterpart, which however requires parameter fine–tuning. Finally, the appropriateness of the BTD model in approximating hyperspectral imaging data was demonstrated in a HSI denoising experiment, where the proposed algorithm was favorably compared to a Bayesian rank-revealing CPD algorithm emanating from one of the simplified models mentioned above.

Future work will focus on the development of constrained (e.g., to ensure nonnegativity) and online variants of the proposed method.

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.*, vol. 65, no. 13, pp. 3551–3582, Jul. 2017.

[3] A. A. Rontogiannis, E. Kofidis, and P. V. Giampouras, "Block-term tensor decomposition: Model selection and computation," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 464–475, Apr. 2021.

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

[5] Y. Qian, F. Xiong, S. Zeng, J. Zhou, and Y. Y. Tang, "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.

[6] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *J. ACM*, vol. 60, no. 6, Nov. 2013, article 45.

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

[8] T. Park and G. Casella, "The Bayesian lasso," *J. Amer. Stat. Assoc.*, vol. 103, no. 482, pp. 681–686, Jun. 2008.

[9] R. M. Neal, *Bayesian Learning for Neural Networks*. Springer, 1996.

[10] M. Mørup and L. K. Hansen, "Automatic relevance determination for multi-way models," *J. Chemometrics*, vol. 23, pp. 352–363, 2009.

[11] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian CP factorization of incomplete tensors with automatic rank determination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1751–1763, Sep. 2015.

[12] 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.

[13] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.

[14] Q. Zhao, G. Zhou, L. Zhang, A. Cichocki, and S.-I. Amari, "Bayesian robust tensor factorization for incomplete multiway data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 736–748, Apr. 2016.

[15] Y. Du, Y. Zheng, K.-C. Lee, and S. Zhe, "Probabilistic streaming tensor decomposition," in *Proc. ICDM-2018*, Singapore, Nov. 2018.

[16] L. Cheng, Z. Chen, Q. Shi, Y.-C. Wu, and S. Theodoridis, "Towards probabilistic tensor canonical polyadic decomposition 2.0: Automatic tensor rank learning using generalized hyperbolic prior," arXiv:2009.02472v1 [cs.LG], Sep. 2020.

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

[18] D. Spencer, "Inference and uncertainty quantification for high-dimensional tensor regression with tensor decompositions and Bayesian methods," Ph.D. dissertation, Univ. California Santa Cruz, Jun. 2020.

[19] L. Xu, L. Cheng, N. Wong, and Y.-C. Wu, "Learning tensor train representation with automatic rank determination from incomplete noisy data," arXiv:2010.06564v1 [eess.SP], Oct. 2020.

[20] F. Sedighin, A. Cichocki, and A.-H. Phan, "Adaptive rank selection for tensor ring decomposition," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 454–463, Apr. 2021.

[21] Z. Long, C. Zhu, J. Liu, and Y. Liu, "Bayesian low rank tensor ring for image recovery," *IEEE Trans. Image Process.*, vol. 30, pp. 3568–3580, Mar. 2021.

[22] Y. Zhou and Y.-M. Cheung, "Bayesian low-tubal-rank robust tensor factorization with multi-rank determination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 62–76, Jan. 2021.

[23] C. Hawkins and Z. Zhang, "Bayesian tensorized neural networks with automatic rank selection," arXiv:1905.10478v1 [cs.LG], May 2019.

[24] M. Zhou, Y. Liu, Z. Long, L. Chen, and C. Zhu, "Tensor rank learning in CP decomposition via convolutional neural network," *Signal Process.: Image Commun.*, vol. 73, pp. 12–21, Apr. 2019.

[25] S. D. Babacan, S. Nakajima, and M. N. Do, "Bayesian group-sparse modeling and variational inference," *IEEE Trans. Signal Process.*, vol. 62, no. 11, pp. 2906–2921, Jun. 2014.

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

[27] Z. Zhang, S. Wang, D. Liu, and M. I. Jordan, "EP-GIG priors and applications in Bayesian sparse learning," *J. Mach. Learn. Res.*, vol. 13, pp. 2031–2061, Jun. 2012.

[28] P. V. Giampouras, A. A. Rontogiannis, K. E. Themelis, and K. D. Koutroumbas, "Online sparse and low-rank subspace learning from incomplete data: A Bayesian view," *Signal Process.*, vol. 137, pp. 199–212, 2017.

[29] A. J. Brockmeier, J. C. Principe, A.-H. Phan, and A. Cichocki, "Greedy algorithm for model selection of tensor decompositions," in *Proc. ICASSP-2013*, Vancouver, Canada, May 2013.

[30] 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.
[31] ——, "A Bayesian approach to block-term tensor decomposition model selection and computation," in *Proc. EUSIPCO-2021*, Dublin, Ireland, Aug. 2021.
[32] A. J. Lemonte and G. M. Cordeiro, "The exponentiated generalized inverse Gaussian distribution," *Stat. Prob. Lett.*, vol. 81, pp. 506—-517, 2011.
[33] J. H. de M. Goulart, P. M. R. de Oliveira, R. C. Farias, V. Zarzoso, and P. Comon, "Alternating group lasso for block-term tensor decomposition and application to ECG source separation," *IEEE Trans. Signal Process.*, vol. 68, pp. 2682–2696, Apr. 2020.
[34] P. V. Giampouras, K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "Simultaneously sparse and low-rank abundance matrix estimation for hyperspectral image unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4775–4789, 2016.
[35] F. Xiong, J. Zhou, and Y. Qian, "Hyperspectral restoration via L_0 gradient regularized low-rank tensor factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10 410–10 425, Dec. 2019.
[36] 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.*, vol. 67, no. 2, pp. 490–503, Jan. 2019.