Enhanced Tensor Rank Learning in Bayesian PARAFAC2 for Noisy Irregular Tensor Data

Zhongtao Chen
Department of EEE
The University of Hong Kong
ztchen@eee.hku.hk

Lei Cheng
ISEE College
Zhejiang University
lei_cheng@zju.edu.cn

Yik-Chung Wu
Department of EEE
The University of Hong Kong
ycwu@eee.hku.hk

***Abstract*—To analyze irregular multi-dimensional data with unaligned dimensions, which frequently appear in real-world signal processing and machine learning tasks, parallel factor analysis 2 (PARAFAC2) has become the state-of-the-art (SOTA) tensor model that yields interpretable learning results. Like other tensor decomposition models, tensor rank learning in PARAFAC2 is vital to overcome overfitting/underfitting, while the prevalent exhaustive searching scheme is computationally inefficient. To realize automatic tensor rank learning, a prior art applies sparsity-promoting Gaussian-gamma (GG) prior to one particular factor matrix. However, its tensor rank learning performance is not satisfactory in high-rank or low-signal-to-noise (SNR) regime. To achieve enhanced tensor rank learning, an advanced sparsity-promoting generalized hyperbolic (GH) prior is proposed to apply to all factor matrices. Theoretical analysis is presented to confirm the desirable sparsity property, and the conjugacy property of the prior is presented, which enables a tractable inference algorithm for enhanced tensor rank learning. Extensive experiments on synthetic data verify that the proposed method has more accurate rank estimates compared to the GG-based PARAFAC2 and its tensor de-noising performance is comparable to the direct fitting method with the rank being known.**

***Index Terms*—Tensor rank learning, PARAFAC2, Bayesian learning, Prior design and analysis**

I. Introduction

Irregular tensor data with different sizes along one dimension naturally arise in various signal processing and machine learning applications, ranging from audio/image tagging [1], [2], text signal processing [3], to electronic health records (EHR) analytics [4], [5]. The irregularity of such tensor data prohibits the straightforward utilization of conventional tensor decomposition models such as canonical polyadic decomposition (CPD) [6] and Tucker decomposition [7], thus urging for more advanced tensor data analytics tools.

Parallel factor analysis 2 (PARAFAC2), which was firstly introduced in [8], [9], has recently gained increasing interest due to its effectiveness in analyzing irregular tensor data [1]–[5]. In particular, given an irregular third-order tensor data $\mathcal{Y} = \{\mathbf{Y}_k \in \mathbb{R}^{I\times J_k}\}_{k=1}^K$, in which each tensor slice $\mathbf{Y}_k$ has a different column number J_k, PARAFAC2 seeks for rank-R factor matrices $\{\mathbf{U}^{(1)} \in \mathbb{R}^{I\times R}, \mathbf{U}^{(3)} \in \mathbb{R}^{K\times R}, \{\mathbf{F}_k \in \mathbb{R}^{J_k\times R}\}_{k=1}^K\}$ via solving the following problem [9]:

$$\min_{\mathbf{U}^{(1)},\mathbf{U}^{(3)},\{\mathbf{F}_k\}_{k=1}^K} \sum_{k=1}^K \left\| \mathbf{Y}_k - \mathbf{U}^{(1)}\mathrm{diag}(\mathbf{U}^{(3)}_{k,:})\mathbf{F}_k^\top \right\|_F^2,$$
$$\text{s.t.} \quad \mathbf{F}_i^\top \mathbf{F}_i = \mathbf{F}_j^\top \mathbf{F}_j, \quad \forall i,j \in \{1,\ldots,K\}. \quad (1)$$

To simplify the constraints of $\{\mathbf{F}_k\}_{k=1}^K$ in (1), a set of orthogonal matrices $\mathcal{P} = \{\mathbf{P}_k \in \mathbb{R}^{J_k\times R}\}$ and a rank-R factor matrix $\mathbf{U}^{(2)} \in \mathbb{R}^{R\times R}$ are introduced in [9], which transforms (1) to:

$$\min_{\{\mathbf{U}^{(n)}\}_{n=1}^3,\mathcal{P}} \sum_{k=1}^K \left\| \mathbf{Y}_k - \mathbf{U}^{(1)}\mathrm{diag}(\mathbf{U}^{(3)}_{k,:})\mathbf{U}^{(2)\top}\mathbf{P}_k^\top \right\|_F^2,$$
$$\text{s.t.} \quad \mathbf{P}_k^\top \mathbf{P}_k = \mathbf{I}_R, \ \forall k \in \{1,\ldots,K\}. \quad (2)$$

In [9], an alternating-optimization-based direct fitting (DF) algorithm was developed to solve problem (2).

While the DF algorithm [9] has demonstrated remarkable performance in PARAFAC2-related applications, practitioners need to carefully tune the column number R of factor matrices $\{\mathbf{U}^{(n)}\}_{n=1}^3$, which is known as tensor rank in tensor-related literature [10]. In particular, the tensor rank R determines the number of unknown model parameters to be estimated, thus controlling the model order. If the tensor rank is overestimated/underestimated, the issue of overfitting/underfitting arises for noisy tensor data, thereby failing to yield interpretable learning results.

A common practice to tune the tensor rank R is via trial-and-error experiments or model selection methods, e.g., the test method based on core consistency diagnostic [11]. However, such approaches demand multiple runs of the PARAFAC2 algorithm, which is computationally inefficient. To alleviate the computational burden, Bayesian learning, which has been widely adopted in various tensor decompositions such as CP [12]–[15], Tucker [16], tensor train [17], block-term tensor [18], has also been applied in probabilistic PARAFAC2 [19].

While probabilistic PARAFAC2 [19] successfully applies the principles of sparse Bayesian learning and achieves automatic tensor rank identification, its tensor rank learning performance deteriorates significantly when the true tensor rank is close to the dimension of the irregular tensor data, or the signal-to-noise ratio (SNR) is low. Such deterioration stems from its prior design. In particular, probabilistic PARAFAC2 [19] adopts Gaussian-gamma (GG) prior as the sparsity-promoting prior that is hierarchically constructed with Gaussian distribution and inverse gamma (IG) distribution.

ISBN: 978-9-4645-9360-0

However, it is known that the GG prior does not yield satisfactory tensor rank learning performance in high-rank or low-SNR regime [15].

To achieve enhanced tensor rank learning performance, we propose to replace the GG prior with advanced sparsity-promoting priors that are more flexible in their functional form, in order to adapt to more diverse levels of sparsity compared with the GG prior. In this paper, we choose the GH prior as the sparsity-promoting prior, which has already been applied to Bayesian CP [15] and block-term tensor [18]. The GH prior includes the GG prior as its special case, and is thus expected to provide more flexibility in modeling different levels of sparsity and more accurate tensor rank learning results. Additionally, to impose a common sparsity structure across factor matrices, we apply the GH prior to all factor matrices to further benefit tensor rank learning [20]. Since the proposed prior has the appealing conjugacy property, a variational inference algorithm with closed-form update expressions is developed to learn the factor matrices and tensor rank.

Extensive numerical experiments on synthetic data demonstrate the superior rank learning and data de-noising performance of the proposed method, especially in high-rank or low-SNR regime.

Notation: Boldface lowercase and uppercase letters will be used for vectors and matrices, respectively. Tensors are written as calligraphic letters. Superscript $\top$ denotes transpose, and the operator $\mathrm{Tr}[\cdot]$ denotes the trace of the argument. $\|\cdot\|_F$ represents the Frobenius norm. The $N \times N$ diagonal matrix with diagonal elements through a_1 to a_N is represented as $\mathrm{diag}\{a_1, \dots, a_N\}$, while $\mathbf{I}_M$ denotes the identity matrix with size $M \times M$. $\mathbb{E}[\cdot]$ represents the expectation of the argument. The operator $\otimes$ denotes the Kronecker product, and $\diamond$ denotes the Khatri-Rao product.

II. A Brief Review of Probabilistic PARAFAC2 [19]

In this section, we briefly review probabilistic PARAFAC2 [19]. It starts with $L(L \geq R)$ columns in all factor matrices and places sparsity-promoting Gaussian-gamma (GG) prior on factor matrix $\mathbf{U}^{(3)}$ to encode sparsity information:

$$p(\mathbf{U}^{(1)}) = \mathcal{MN}(\mathbf{U}^{(1)}|\mathbf{0}_{I\times L}, \mathbf{I}_I, \mathbf{I}_L), \tag{3}$$

$$p(\mathbf{U}^{(2)}) = \mathcal{MN}(\mathbf{U}^{(2)}|\mathbf{0}_{L\times L}, \mathbf{I}_L, \mathbf{I}_L), \tag{4}$$

$$p(\mathbf{U}^{(3)}|\mathbf{\Gamma}) = \mathcal{MN}(\mathbf{U}^{(3)}|\mathbf{0}_{K\times L}, \mathbf{I}_K, \mathbf{\Gamma}), \tag{5}$$

$$p(\mathbf{\Gamma}) = \prod_{l=1}^{L} \mathrm{IG}(\gamma_l|e_l^0, f_l^0), \tag{6}$$

where $\mathcal{MN}(\mathbf{X}|\mathbf{M}, \mathbf{U}, \mathbf{V})$ denotes a matrix normal distribution on random matrix $\mathbf{X}$ parametrized by mean matrix $\mathbf{M}$, row covariance matrix $\mathbf{U}$, and column covariance matrix $\mathbf{V}$, with the probability density function (pdf) $\mathcal{MN}(\mathbf{X}|\mathbf{M}, \mathbf{U}, \mathbf{V}) \propto \exp(-1/2\mathrm{Tr}[\mathbf{V}^{-1}(\mathbf{X}-\mathbf{M})^\top\mathbf{U}^{-1}(\mathbf{X}-\mathbf{M})])$.

In the GG prior (5)-(6), the column covariance matrix $\mathbf{\Gamma}$ is a diagonal matrix with elements being $\{\gamma_1, \gamma_2, \dots, \gamma_L\}$, where γ_l represents the variance of column $\mathbf{U}^{(3)}_{:,l}$. Each variance γ_l is assigned with an inverse gamma (IG) prior, where e_l^0, f_l^0 are pre-determined hyper-parameters. The prior of other factor matrices (3)-(4) follow standard matrix normal distribution, where no sparsity information is embedded.

On the other hand, the likelihood function in probabilistic PARAFAC2 [19] is obtained from the objective function of (2), by assuming each observation in $\mathcal{Y}$ is subject to independent Gaussian noise perturbation. This results in

$$p(\mathcal{Y}|\{\mathbf{U}^{(n)}\}_{n=1}^3, \tau; \mathcal{P}) \propto \exp\left\{-\frac{\tau}{2}\sum_{k=1}^{K}\left\|\mathbf{Y}_k - \mathbf{U}^{(1)}\mathrm{diag}(\mathbf{U}^{(3)}_{k,:})\mathbf{U}^{(2)\top}\mathbf{P}_k^\top\right\|_F^2\right\}. \tag{7}$$

For the noise precision τ, a gamma prior $\mathrm{Ga}(\tau|c^0, d^0)$ is assigned, where c^0, d^0 are pre-determined hyper-parameters.

Based on the prior and likelihood functions in (3)-(7), an inference algorithm under the variational inference framework was developed in [19], which will drive irrelevant columns to zero-value and thereby achieve automatic tensor rank learning.

However, as will be shown in Section V, numerical results reveal that the tensor rank learning performance of probabilistic PARAFAC2 [19] is not satisfactory in high-rank or low-SNR regime. This is due to the rigidity of GG prior that it cannot adapt to different levels of sparsity [15]. Furthermore, probabilistic PARAFAC2 [19] does not apply the sparsity-promoting prior to all factor matrices, which fails to capture the common sparsity structure across factor matrices [20].

III. Novel Prior for Bayesian PARAFAC2

To achieve enhanced tensor rank learning, we propose to apply advanced generalized hyperbolic (GH) prior to all factor matrices. Since the GH prior can be expressed as a Gaussian scale mixture where the mixing distribution is generalized inverse Gaussian (GIG) distribution, it allows a hierarchical construction of the GH prior:

$$p(\mathbf{U}^{(1)}|\mathbf{Z}) = \mathcal{MN}(\mathbf{U}^{(1)}|\mathbf{0}_{I\times L}, \mathbf{I}_I, \mathbf{Z}), \tag{8}$$

$$p(\mathbf{U}^{(2)}|\mathbf{Z}) = \mathcal{MN}(\mathbf{U}^{(2)}|\mathbf{0}_{L\times L}, \mathbf{I}_L, \mathbf{Z}), \tag{9}$$

$$p(\mathbf{U}^{(3)}|\mathbf{Z}) = \mathcal{MN}(\mathbf{U}^{(3)}|\mathbf{0}_{K\times L}, \mathbf{I}_K, \mathbf{Z}), \tag{10}$$

$$p(\mathbf{Z}) = \prod_{l=1}^{L} \mathrm{GIG}(z_l|a_l^0, b_l^0, \lambda_l^0), \tag{11}$$

where the column covariance matrix $\mathbf{Z} = \mathrm{diag}\{z_1, \dots, z_L\}$, and the pdf of the GIG distribution is

$$\mathrm{GIG}(z_l|a_l^0, b_l^0, \lambda_l^0) = \frac{(a_l^0/b_l^0)^{\lambda_l^0/2}}{2K_{\lambda_l^0}(\sqrt{a_l^0 b_l^0})} z_l^{\lambda_l^0-1} \exp\left(-\frac{1}{2}(a_l^0 z_l + b_l^0 z_l^{-1})\right). \tag{12}$$

The GH prior (8)-(11) is more flexible in its functional form compared to the GG prior (5)-(6), as it can be reduced to the GG prior by setting its parameters to certain values, which is formally presented in the property below.

Property 1. *When $a_l^0 \to 0$ with $\lambda_l^0 < 0$, GIG distribution reduces to inverse gamma (IG) distribution* $\mathrm{IG}(z_l|b_l^0/2,-\lambda_l^0)$,

$$\mathrm{GIG}(z_l|a_l^0 \to 0, b_l^0, \lambda_l^0 < 0) = \frac{(b_l^0/2)^{-\lambda_l^0}}{\Gamma(-\lambda_l^0)} z_l^{\lambda_l^0 - 1} \exp\left(-\frac{1}{2} b_l^0 z_l^{-1}\right). \tag{13}$$

The GH prior generalizes not only the GG prior, but also many other prevalent sparsity-promoting prior [15], [21], including Laplacian and Mcakay's Bessel. Therefore, the flexibility in its functional form is expected to adapt to different levels of sparsity, and thus to improve tensor rank learning performance.

Furthermore, by comparing the proposed prior (8)-(11) with probabilistic PARAFAC2 [19] (3)-(6), another significant difference is that the column covariance matrix $\mathbf{Z}$ is shared across all factor matrices to impose a common sparsity pattern, which was shown to be beneficial for rank selection [20], and widely practiced in Bayesian tensor modelings [12]–[18].

Aside from the flexibility adapting to different levels of sparsity and the shared sparsity pattern, the conjugacy property of the GIG prior (11) presented below facilitates the derivation of the inference algorithm in Section IV.

Property 2. *GIG distribution* (11) *is conjugate to the product of pdfs* (8)-(10).

IV. Bayesian PARAFAC2 with Enhanced Tensor Rank Learning

The proposed Bayesian PARAFAC2 model is composed of the prior introduced in the previous section, and the likelihood function in (7). The complete model is represented by the joint probability of the noisy irregular tensor data $\mathcal{Y}$ and $\boldsymbol{\Theta} = \{\{\mathbf{U}^{(n)}\}_{n=1}^{3}, \{z_l\}_{l=1}^{L}, \tau\}$, parametrized by orthogonal matrices in $\mathcal{P}$,

$$p(\mathcal{Y}, \boldsymbol{\Theta}; \mathcal{P}) = p(\mathcal{Y}|\{\mathbf{U}^{(n)}\}_{n=1}^{3}, \tau; \mathcal{P}) \prod_{n=1}^{3} p(\mathbf{U}^{(n)}|\mathbf{Z}) p(\mathbf{Z}) p(\tau). \tag{14}$$

A. General Philosophy of Variational EM

To learn parameters $\boldsymbol{\Theta}$ and $\mathcal{P}$, Bayesian statistics suggests to derive posterior distribution and evidence function respectively, which however are analytically intractable. To provide a viable solution, variational expectation-maximization (EM) framework is employed in this paper, which maximizes the evidence lower bound (ELBO) [22]:

$$\max_{Q(\boldsymbol{\Theta}), \mathcal{P}} \mathrm{ELBO}(Q(\boldsymbol{\Theta}), \mathcal{P}) \triangleq \mathbb{E}_{Q(\boldsymbol{\Theta})} \left\{ \ln \frac{p(\mathcal{Y}, \boldsymbol{\Theta}; \mathcal{P})}{Q(\boldsymbol{\Theta})} \right\}, \tag{15}$$

where $Q(\boldsymbol{\Theta})$ is known as variational pdf that aims to approach the posterior distribution. Particularly, variational EM algorithm alternatingly maximizes ELBO with respect to $Q(\boldsymbol{\Theta})$ with $\mathcal{P}$ fixed in the E-step and optimizes over $\mathcal{P}$ while keeping $Q(\boldsymbol{\Theta})$ fixed in the M-step.

In the E-step, given a fixed $\mathcal{P}$, problem (15) is equivalent to minimizing the Kullback-Leiber (KL) divergence between $Q(\boldsymbol{\Theta})$ and posterior distribution $p(\boldsymbol{\Theta}|\mathcal{Y}; \mathcal{P})$. Under no additional assumption, the optimal solution is the exact posterior distribution, which is intractable due to the high-dimensional integration involved. To address this issue, we restrict $Q(\boldsymbol{\Theta})$ to be in the mean-filed family. More specifically, the mean-field family collects pdfs with the following form: $Q(\boldsymbol{\Theta}) = \prod_{k=1}^{K} Q(\boldsymbol{\Theta}_k)$, where $\boldsymbol{\Theta}$ is partitioned into mutually disjoint non-empty subsets $\boldsymbol{\Theta}_k$. Then each optimal $Q^*(\boldsymbol{\Theta}_k)$ was shown to be [22]:

$$Q^*(\boldsymbol{\Theta}_k) = \frac{\exp\left(\mathbb{E}_{\prod_{j \neq k} Q(\boldsymbol{\Theta}_j)} \left[\ln p(\mathcal{Y}, \boldsymbol{\Theta}; \mathcal{P})\right]\right)}{\int \exp\left(\mathbb{E}_{\prod_{j \neq k} Q(\boldsymbol{\Theta}_j)} \left[\ln p(\mathcal{Y}, \boldsymbol{\Theta}; \mathcal{P})\right]\right) d\boldsymbol{\Theta}_k}. \tag{16}$$

For the proposed Bayesian PARAFAC2 model, the mean-field assumption is designed as $Q(\boldsymbol{\Theta}) = \prod_{n=1}^{3} Q(\mathbf{U}^{(n)}) \prod_{l=1}^{L} Q(z_l) Q(\tau)$.

On the other hand, in the M-step, with $Q(\boldsymbol{\Theta})$ fixed, it can be seen that the optimal parameter $\mathcal{P}$ are sought via the maximization of the expectation of the complete log likelihood. In particular, by substituting (14) into (15) and extracting only the terms related to $\mathcal{P}$, the optimization problem becomes

$$\begin{aligned} &\max_{\mathcal{P}} \ \mathbb{E}_{Q(\boldsymbol{\Theta})} \left[-\frac{\tau}{2} \sum_{k=1}^{K} \left\| \mathbf{Y}_k - \mathbf{U}^{(1)} \mathrm{diag}(\mathbf{U}_{k,:}^{(3)}) \mathbf{U}^{(2)\top} \mathbf{P}_k^{\top} \right\|_F^2 \right], \\ &\text{s.t.} \quad \mathbf{P}_k^{\top} \mathbf{P}_k = \mathbf{I}_L, \ \ \forall k \in \{1, \ldots, K\}. \end{aligned} \tag{17}$$

By substituting the mean-filed assumption in (17), the solution to the maximization problem is given as [9]

$$\mathbf{P}_k = \boldsymbol{\Psi}_k \boldsymbol{\Xi}_k^{\top}, \ \ \forall k \in \{1, \ldots, K\}, \tag{18}$$

where $\boldsymbol{\Xi}_k$ and $\boldsymbol{\Psi}_k$ are orthogonal matrices obtained from the following singular value decomposition (SVD):

$$\mathbb{E}[\mathbf{U}^{(2)}] \mathrm{diag}(\mathbb{E}[\mathbf{U}_{k,:}^{(3)}]) \left[\mathbb{E}[\mathbf{U}^{(1)}]\right]^{\top} \mathbf{Y}_k = \boldsymbol{\Xi}_k \boldsymbol{\Upsilon}_k \boldsymbol{\Psi}_k^{\top}. \tag{19}$$

B. Deriving Optimal Variational Pdfs

By substituting (14) into (16), the optimal variational pdfs for various variables can be derived. Due to the page limit, the lengthy derivations are omitted and we directly present the optimal variational pdfs below.

Update of $Q(\mathbf{U}^{(n)})$

The optimal variational pdfs $Q^*(\mathbf{U}^{(n)})$ are derived to be a matrix normal distributions, with identity row covariance matrix, column covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma}^{(n)} = \Bigg[\mathbb{E}[\tau] \, \mathbb{E}\Bigg[&\Bigg(\mathop{\diamond}_{k=1, k \neq n}^{3} \mathbf{U}^{(k)} \Bigg)^{\top} \\ &\times \Bigg(\mathop{\diamond}_{k=1, k \neq n}^{3} \mathbf{U}^{(k)} \Bigg) \Bigg] + \mathbb{E}[\mathbf{Z}] \Bigg]^{-1}, \end{aligned} \tag{20}$$

and mean matrix

$$\mathbf{M}^{(n)} = \mathbb{E}[\tau] \, \mathcal{W}_{(n)}^{\top} \mathbf{E} \left[\mathop{\diamond}_{k=1, k \neq n}^{3} \mathbf{U}^{(k)} \right] \boldsymbol{\Sigma}^{(n)}, \tag{21}$$

where tensor $\mathcal{W} \in \mathbb{R}^{I\times L\times K}$ is constructed with each tensor slice being $\mathbf{W}_k = \mathbf{Y}_k\mathbf{P}_k$ and $\mathcal{W}_{(n)}$ denotes the matrix obtained by unfolding the tensor $\mathcal{W}$ along its n-th dimension.

Update of $Q(z_l)$

By Property 2, the optimal variational pdf of $Q(z_l)$ can be derived to be a GIG distribution $\mathrm{GIG}(z_l|a_l, b_l\lambda_l)$, with parameters

$$a_l = a_l^0, \tag{22}$$

$$b_l = b_l^0 + \sum_{n=1}^{N} \mathbb{E}\left[\left[\mathbf{U}_{:,l}^{(n)}\right]^\top \mathbf{U}_{:,l}^{(n)}\right], \tag{23}$$

$$\lambda_l = \lambda_l^0 - (I + L + K)/2. \tag{24}$$

Update of $Q(\tau)$

Finally, for the noise precision τ, the optimal variational pdf $Q^*(\tau)$ is derived to be a gamma distribution $\mathrm{Ga}(c, d)$, with

$$c = c^0 + (I \times \sum_{k=1}^{K} J_k)/2, \tag{25}$$

$$d = d^0 + \frac{1}{2}\mathbb{E}\left[\sum_{k=1}^{K}\left\|\mathbf{Y}_k - \mathbf{U}^{(1)}\mathrm{diag}(\mathbf{U}_{k,:}^{(3)})\mathbf{U}^{(2)\top}\mathbf{P}_k^\top\right\|_F^2\right]. \tag{26}$$

C. Algorithm Summary and Insights

To summarize, the variational EM algorithm for the proposed Bayesian model estimates $Q(\boldsymbol{\Theta})$ and $\mathcal{P}$ by iteratively updating the parameters via (18)-(26) until convergence. The means of the initial factor matrices $\{\mathbf{M}^{(n)}\}_{n=1}^3$ of $\{Q(\mathbf{U}^{(n)})\}_{n=1}^3$ are obtained via a DF algorithm [9], while the initial column covariance matrices $\{\boldsymbol{\Sigma}^{(n)}\}_{n=1}^3$ are set as identity matrices. The hyper-parameters $\{\{a_l^0, b_l^0, \lambda_l^0\}_{l=1}^L, c^0, d^0\}$ are set to be 10^{-6} to yield non-informative prior. The computational complexity of the variational EM algorithm is $O(L^3IK)$, where L is the initial tensor rank.

V. Numerical Experiments

To validate the performance of the proposed algorithm, experimental results on synthetic data are reported in this section. We compare the performance of the following algorithms: 1) the proposed enhanced Bayesian PARAFAC2 algorithm (EB-PARAFAC2) that can handle both regular and irregular tensor data; 2) probabilistic PARAFAC2 algorithm [19] (B-PARAFAC2) that was originally derived for regular tensor data, but we extend it to irregular tensor data; 3) the direct fitting (DF) algorithm [9] which assumes the true rank is known. Two tensor sizes are considered: 1) irregular tensor with $I = 30$, $K = 30$, $J_k = 30$ for $1 \le k \le 20$, $J_k = 40$ for $21 \le k \le 30$; 2) regular tensor with $I = 30$,$K = 30$, $J_k = 30$ for $1 \le k \le 30$. All simulation results in this section are obtained by averaging 100 Monte Carlo runs.

Noise-free rank-R irregular tensor $\mathcal{X} = \{\mathbf{X}_k = \mathbf{U}^{(1)}\mathrm{diag}(\mathbf{U}_{k,:}^{(3)})[\mathbf{U}^{(2)}]^\top\mathbf{P}_k^\top \in \mathbb{R}^{I\times J_k}\}_{k=1}^K$ is generated following the procedure specified by [9]. Observation tensor $\mathcal{Y}$ is constructed by adding noise sampled from $\mathcal{N}(0, \sigma^2)$ on each entry of tensor $\mathcal{X}$. The noise level is measured by SNR, which is defined as $10\log(\mathrm{var}(\mathcal{X})/\sigma^2)$ [12], where the variance $\mathrm{var}(\mathcal{X})$ of tensor $\mathcal{X}$ is calculated by treating all entries in tensor $\mathcal{X}$ as statistically independent.

We firstly examine the tensor rank learning performance. The tensor rank upper bound L for both EB-PARAFAC2 and B-PARAFAC2 algorithm are set to be the $\min\{J_k\}_{k=1}^K$. The percentages of accurate tensor rank estimates are reported in Fig. 1. For irregular tensor data, when SNR is 15 dB, as seen in Figure 1(a), GG-based B-PARAFAC2 does not yield satisfactory tensor rank learning performance when the tensor rank is close to the dimension of the irregular tensor data (R is in $\{21, 24, 27\}$ in this case), while the proposed EB-PARAFAC2 demonstrates more robust tensor rank learning performance, since the proposed prior is more flexible to different levels of sparsity and imposes a common sparsity structure. Similarly, when SNR is 5 dB, as shown in 1(b), while the percentage of accurate tensor rank learning estimates decreases for both algorithms when tensor rank increases, the proposed EB-PARAFAC2 algorithm gives more accurate tensor rank estimates compared to B-PARAFAC2. For regular tensor data, as reported in Fig. 1(c) and 1(d), the proposed EB-PARAFAC2 algorithm also yields better tensor rank learning results than B-PARAFAC2 in all cases.

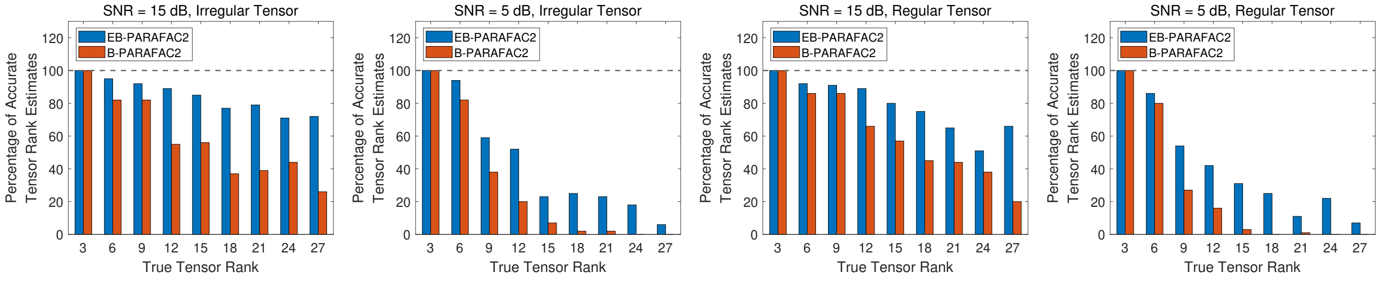
The more accurate tensor rank learning of the proposed method leads to better tensor de-noising performance. To see this, we examine the accuracy of tensor recovery by measuring the root mean squared error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{I \times \sum_{k=1}^{K} J_k}\left\|\mathrm{vec}(\mathcal{X}) - \mathrm{vec}(\hat{\mathcal{X}})\right\|_2^2}, \tag{27}$$

where $\hat{\mathcal{X}}$ represents the reconstructed clean tensor. For irregular tensor data, when SNR is 15 dB, as seen from Fig. 2(a), the proposed EB-PARAFAC2 algorithm achieves similar performance to the genie-aided DF algorithm (which has the knowledge of the true tensor rank), and they perform better than B-PARAFAC2, especially in high-rank cases (R is in $\{21, 24, 27\}$). Furthermore, when the SNR is 5 dB, as shown in Fig. 2(b), with the enhanced tensor rank learning capability, the tensor data recovery performance of EB-PARAFAC2 is better than that of B-PARAFAC2. However, due to the inaccurate tensor rank learning, they are both worse than the genie-aided DF algorithm. For regular tensor data, as demonstrated in Fig. 2(c) and 2(d), the proposed EB-PARAFAC2 algorithm, even without the knowledge of tensor rank, attains comparable performance with the genie-aided DF algorithm, and performs better than B-PARAFAC2.

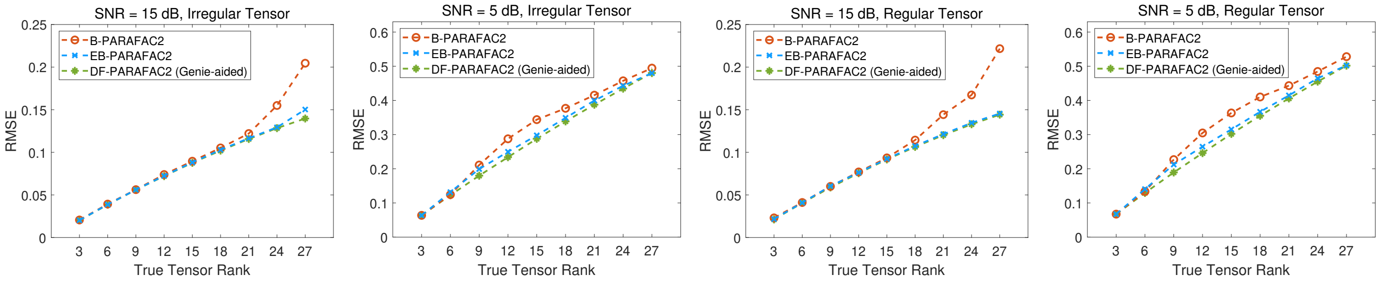
VI. Conclusions

In this paper, a novel Bayesian learning algorithm for PARAFAC2 model with enhanced tensor rank learning was proposed. The proposed Bayesian framework adopts the advanced GH prior as the sparsity-promoting prior and imposes a common sparsity structure to improve tensor rank learning. Extensive experiments demonstrated the enhanced performance of the proposed algorithm in terms of tensor rank learning and data recovery accuracy.



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

Fig. 1: Performance of tensor rank learning under different conditions.



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

Fig. 2: Performance of tensor recovery under different conditions.

References

[1] Y. Panagakis and C. Kotropoulos, "Automatic music tagging via parafac2," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 481–484.

[2] E. Pantraki and C. Kotropoulos, "Automatic image tagging and recommendation via parafac2," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[3] P. A. Chew, B. W. Bader, T. G. Kolda, and A. Abdelali, "Cross-language information retrieval using parafac2," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 143–152.

[4] K. Yin, A. Afshar, J. C. Ho, W. K. Cheung, C. Zhang, and J. Sun, "Logpar: Logistic parafac2 factorization for temporal binary data with missing values," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1625–1635.

[5] Y. Ren, J. Lou, L. Xiong, and J. C. Ho, "Robust irregular tensor factorization and completion for temporal health data analysis," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1295–1304.

[6] R. A. Harshman *et al.*, "Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis," 1970.

[7] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[8] R. A. Harshman, "Parafac2: Mathematical and technical notes," *UCLA Working Papers in Phonetics*, vol. 22, no. 10, pp. 30–44, 1972.

[9] H. A. Kiers, J. M. Ten Berge, and R. Bro, "Parafac2—part i. a direct fitting algorithm for the parafac2 model," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 13, no. 3-4, pp. 275–294, 1999.

[10] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[11] M. H. Kamstrup-Nielsen, L. G. Johnsen, and R. Bro, "Core consistency diagnostic in parafac2," *Journal of Chemometrics*, vol. 27, no. 5, pp. 99–105, 2013.

[12] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian cp factorization of incomplete tensors with automatic rank determination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1751–1763, 2015.

[13] L. Cheng, Y.-C. Wu, and H. V. Poor, "Probabilistic tensor canonical polyadic decomposition with orthogonal factors," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 663–676, 2016.

[14] L. Cheng, X. Tong, S. Wang, Y.-C. Wu, and H. V. Poor, "Learning nonnegative factors from tensor data: Probabilistic modeling and inference algorithm," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1792–1806, 2020.

[15] L. Cheng, Z. Chen, Q. Shi, Y.-C. Wu, and S. Theodoridis, "Towards flexible sparsity-aware modeling: Automatic tensor rank learning using the generalized hyperbolic prior," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1834–1849, 2022.

[16] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian sparse tucker models for dimension reduction and tensor completion," *arXiv preprint arXiv:1505.02343*, 2015.

[17] L. Xu, L. Cheng, N. Wong, and Y.-C. Wu, "Overfitting avoidance in tensor train factorization and completion: Prior analysis and inference," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1439–1444.

[18] P. V. Giampouras, A. A. Rontogiannis, and E. Kofidis, "Block-term tensor decomposition model selection and computation: The bayesian way," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1704–1717, 2022.

[19] P. J. Jørgensen, S. F. Nielsen, J. L. Hinrich, M. N. Schmidt, K. H. Madsen, and M. Mørup, "Analysis of chromatographic data using the probabilistic parafac2," in *33rd Conf. on Neural Information Processing Systems*, 2019.

[20] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse bayesian methods for low-rank matrix estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 3964–3977, 2012.

[21] S. D. Babacan, S. Nakajima, and M. N. Do, "Bayesian group-sparse modeling and variational inference," *IEEE transactions on signal processing*, vol. 62, no. 11, pp. 2906–2921, 2014.

[22] C. M. Bishop, "Pattern recognition," *Machine learning*, vol. 128, no. 9, 2006.