Bayesian Robust Tensor Factorization for Incomplete Multiway Data

Qibin Zhao, Guoxu Zhou, Liqing Zhang, Andrzej Cichocki, and Shun-ichi Amari

***Abstract*—We propose a generative model for robust tensor factorization in the presence of both missing data and outliers. The objective is to explicitly infer the underlying *low-CP-rank* tensor capturing the global information and a sparse tensor capturing the local information (also considered as outliers), thus providing the robust predictive distribution over missing entries. The *low-CP-rank* tensor is modeled by multilinear interactions between multiple latent factors on which the column sparsity is enforced by a hierarchical prior, while the sparse tensor is modeled by a hierarchical view of Student-t distribution that associates an individual hyperparameter with each element independently. For model learning, we develop an efficient variational inference under a fully Bayesian treatment, which can effectively prevent the overfitting problem and scales linearly with data size. In contrast to existing related works, our method can perform model selection automatically and implicitly without need of tuning parameters. More specifically, it can discover the groundtruth of *CP rank* and automatically adapt the sparsity inducing priors to various types of outliers. In addition, the tradeoff between the low-rank approximation and the sparse representation can be optimized in the sense of maximum model evidence. The extensive experiments and comparisons with many state-of-the-art algorithms on both synthetic and real-world datasets demonstrate the superiorities of our method from several perspectives.**

***Index Terms*—Tensor factorization, tensor completion, robust factorization, rank determination, variational Bayesian inference, video background modeling**

I. INTRODUCTION

TENSORS (i.e., multiway arrays) can provide an efficient and faithful representation of structural properties for multidimensional data. For instance, a facial image ensemble affected by multiple conditions can be represented as a higher order tensor with dimensionality of $pixel \times person \times pose \times illumination$. To model such data, tensor factorization has shown significant advantages in terms of capturing multiple interactions among a set of latent factors. Therefore its theory and algorithms have been an active area of study within the past decade, see e.g. [1], [2], and have been successfully applied to various fields of applications such as face recognition, social network analysis, image and video completion, and brain signal processing [3], [4], [5], [6]. The two most popular tensor factorization frameworks are Tucker [7] and CANDECOMP/PARAFAC (CP) [1], also known as canonical polyadic decomposition (CPD) [8], which naturally results in two different definitions of tensor rank, i.e., *multilinear rank* and *CP rank*.

When original data is only partially observed, tensor factorization can be applied for imputing the missing entries, known as tensor completion. CP factorization with missing values has been developed by employing CP weighted optimization (CPWOPT) [9] and nonlinear least squares (CPNLS) [10]. To naturally deal with missing data, the probabilistic framework for tensor factorization was exploited [3], [11], which has been extended to exponential family model [12] and nonparametric Bayesian model [13]. The main limitation of the existing tensor factorization scheme is that the tensor rank has to be specified manually, which tends to under- or over-fit the observations, resulting in severe deterioration of predictive performance. It is important to emphasize that our knowledge about the properties of tensor rank, especially *CP rank*, is surprisingly limited [14]. There is no straightforward algorithm to compute *CP rank* of an explicitly given tensor, and the problem has been shown to be NP-hard [15], [16]. In fact, determining or even bounding the tensor rank is quite difficult in contrast to matrix rank [17], [18]. In [19], ARD framework was applied to estimate the *multilinear rank*. However, the solution is based on MAP point estimation and is not applicable to incomplete tensor data. Recently, Bayesian low-rank decomposition of incomplete tensors has been proposed in [20], which can perform tensor completion while the CP rank can be also inferred by employing a multiplicative Gamma process prior. However, the missing data is handled by a heuristic way, i.e., estimating the missing data followed by the factorization on a whole tensor alternately. In addition, the inference is performed by Gibbs sampler which is generally shown slow convergence.

The convex optimization of nuclear norm has gained considerable attention in matrix completion, which essentially seeks the minimum rank under the condition of limited observations. Since *multilinear rank* of a tensor is defined as the rank of its mode-n matricizations, it can be optimized by simply applying nuclear norm based framework, yielding an extension to tensor completion [21], which thus attracted many studies on low multilinear rank approximations [22], [23], [24]. In addition, the auxiliary information can be exploited to improve completion accuracy [6], [25], which, however, is only suitable

Q. Zhao is with Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Japan and Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

G. Zhou is with Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Japan.

L. Zhang is with Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

A. Cichocki is with Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Japan and Systems Research Institute in Polish Academy of Science, Warsaw, Poland.

S. Amari is with Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Japan.

to some specific applications. It is worth noting that the convex optimization of nuclear norm requires several tuning parameters, which is prone to over- or under-estimate the tensor rank. In addition, since *CP rank*, the standard definition of tensor rank, cannot be optimized by applying matrix techniques straightforwardly, its determination still remains challenging so far.

On the other hand, non-Gaussian noises or outliers may occur frequently in image and video data. To handle this problem, many robust techniques have been developed such as robust PCA [26], [27], [28], [29] and robust matrix factorization [30], [31], [32], [33], [34]. For robust tensor factorization, 2DSVD using R_1-norm as the objective function was proposed by Huang et al. [35] and robust Tucker decompositions were studied in [36], [37]. To handle both missing data and outliers, the nuclear norm regularization has been combined with L_1-norm loss function, which leads to a robust tensor completion [38]. However, the main limitation of above mentioned approaches is that the performance is quite sensitive to tuning parameters whose optimal selection is unrealistic or prohibitively expensive. For example, the parameter balancing model capacity between a low-rank term and a sparse term is generally tuned by performance evaluated on the groundtruth of missing data that is unknown in practice, implying that most existing approaches are impractical to obtain the optimal results. Therefore, an automatic model and parameter selection based solely on observed data, which can achieve an optimal predictive performance, is appealing. Another limitation of existing robust tensor factorizations is that optimizations of latent factors are mainly based on point estimation, which is prone to overfitting especially when a large amount of missing data is present and not able to provide uncertainty information of predictions.

To address all these issues under a unified framework, we propose a probabilistic model with aim to recover the underlying low-rank tensor, modeled by multiplicative interactions among multiple groups of latent factors, and an additive sparse tensor modeling outliers, from partially observed data represented by a tensor of any order. More specifically, for the low-rank term, we specify a hierarchical sparsity-inducing prior shared by multiple groups of latent factors, which gains a column sparsity along the latent components, resulting in an automatic determination of *CP rank*. For the sparse term, a hierarchical view of Student-t prior is placed independently on each element associated with an individual hyperparameter. The top-level hyperparameters can be learned by maximizing a lower-bound of the model evidence, resulting in that the sparsity constraint can be automatically adapted to varying fractions of outliers. To learn the model under a fully Bayesian framework, we derive a variational Bayesian algorithm for posterior inference, which is of high efficiency. Our method can be used for robust tensor factorization, robust tensor completion and anomaly detection with a significant advantage of automatic model and parameter selections without requiring any tuning parameters. Empirical results on both synthetic and real-world datasets demonstrate that the proposed method outperforms state-of-the-art methods in terms of predictive performance and robustness to outliers, even though the

groundtruth is allowed to be used to tune the parameters in competing methods.

The rest of this paper is organized as follows. Section II discusses the related work. Section III introduces preliminary multilinear operations and notations. Section IV presents model specification and approximate Bayesian inference for robust tensor factorization, whose advantages are summarized in Section V. Section VI shows extensive experimental results, followed by the conclusion in Section VII.

II. RELATED WORK

Our work is somewhat related to probabilistic approaches for robust PCA [39], [40] and for robust matrix factorization [41], [42], [43]. In [39], Beta-Bernoulli distribution is exploited to model outliers and the low-rank matrix exclusively, which, however, results in high model complexity and slow inference. Missing values are considered in [40], where the number of latent components needs to be specified in advance. In [43], Jeffreys prior is adopted to model both noise and outliers. However, it cannot handle missing values. PRMF [41] uses Laplace distribution to model the residuals, while the rank of the underlying model should be given in advance. A fully Bayesian treatment of PRMF [42] employs a hierarchical view of Laplace distribution as the noise model, and applies MCMC sampling for model inference. However, the rank also needs to be tuned manually and missing values are not considered. Finally, all these matrix based approaches cannot handle interactions of multiple factors, which is crucial for higher-order tensors.

Higher-order robust PCA (HORPCA), proposed very recently in [38], is the only existing tensor method that can handle both missing data and outliers. It formulates the problem by a convex optimization framework in which nuclear norm and L_1-norm are exploited as regularization terms on the low-rank tensor and residual errors, respectively. However, it essentially optimizes the *multilinear rank* and the predictive performance is sensitive to tuning parameters. To our best knowledge, our paper is the first to present a fully Bayesian model for robust tensor factorization dealing with both missing data and outliers within one framework.

III. PRELIMINARIES AND NOTATIONS

The order of a tensor is the number of dimensions, also known as ways or modes. Vectors are denoted by boldface lowercase letters, e.g., $\mathbf{a}$. Matrices are denoted by boldface capital letters, e.g., $\mathbf{A}$. Higher-order tensors (order ≥ 3) are denoted by boldface calligraphic letters, e.g., $\boldsymbol{\mathcal{A}}$. Given an Nth order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, its $(i_1, i_2, \ldots, i_N)$th entry is denoted by $\mathcal{X}_{i_1 i_2 \ldots i_N}$ where the indices typically range from 1 to their capital version, e.g., $i_n = 1, 2, \ldots, I_n, n = 1, \ldots, N$.

The inner product of two tensors is defined by $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle = \sum_{i_1 i_2 \ldots i_N} \mathcal{A}_{i_1 i_2 \ldots i_N} \mathcal{B}_{i_1 i_2 \ldots i_N}$, and the squared Frobenius norm by $\|\boldsymbol{\mathcal{A}}\|_F^2 = \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{A}} \rangle$.
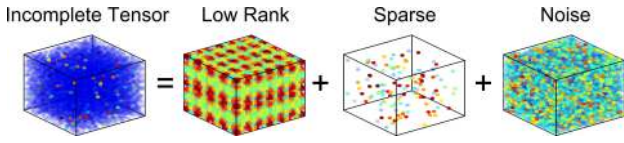
Fig. 1. Bayesian robust tensor factorization.

Definition III.1. The *generalized inner product* of $N \geq 3$ vectors, matrices, or tensors is defined as a sum of element-wise products. For example,

$$\left\langle \mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)} \right\rangle = \sum_{i,j} \prod_n A_{ij}^{(n)}. \tag{1}$$

The *Hadamard product* is an entrywise product of two vectors, matrices or tensors of the same sizes. For instance, $\mathbf{A} \in \mathbb{R}^{I \times J}$ and $\mathbf{B} \in \mathbb{R}^{I \times J}$, their Hadamard product, denoted by $\mathbf{A} \circledast \mathbf{B}$, is a matrix of size $I \times J$. Without loss of generality, the Hadamard product of a set of matrices $\{\mathbf{A}^{(n)}\}_{n=1}^N$ is simply denoted by

$$\circledast_n \mathbf{A}^{(n)} = \mathbf{A}^{(1)} \circledast \mathbf{A}^{(2)} \circledast \cdots \circledast \mathbf{A}^{(N)}. \tag{2}$$

The *Kronecker product* [1] of matrices $\mathbf{A} \in \mathbb{R}^{I \times J}$ and $\mathbf{B} \in \mathbb{R}^{K \times L}$ is a matrix of size $IK \times JL$, denoted by $\mathbf{A} \otimes \mathbf{B}$. The *Khatri-Rao product* of matrices $\mathbf{A} \in \mathbb{R}^{I \times K}$ and $\mathbf{B} \in \mathbb{R}^{J \times K}$ is a matrix of size $IJ \times K$ defined by a columnwise Kronecker product, and denoted by $\mathbf{A} \odot \mathbf{B}$. In particular, the Khatri-Rao product of a set of matrices in a reverse order is denoted by

$$\bigodot_n \mathbf{A}^{(n)} = \mathbf{A}^{(N)} \odot \mathbf{A}^{(N-1)} \odot \cdots \odot \mathbf{A}^{(1)}, \tag{3}$$

while the Khatri-Rao product of $\{\mathbf{A}^{(n)}\}_{n=1}^N$ except the nth matrix, denoted by $\mathbf{A}^{(\backslash n)}$, is defined by

$$\bigodot_{k \neq n} \mathbf{A}^{(k)} = \mathbf{A}^{(N)} \odot \cdots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \cdots \odot \mathbf{A}^{(1)}. \tag{4}$$

IV. BAYESIAN ROBUST CP FACTORIZATION

A. Model Specification

Let $\mathcal{Y}$ be an incomplete Nth-order tensor of size $I_1 \times I_2 \times \cdots \times I_N$ with missing entries. $\mathcal{Y}_\Omega$ denotes the observed entries $\{\mathcal{Y}_{i_1 i_2 \ldots i_N} | (i_1, i_2, \cdots, i_N) \in \Omega\}$ where Ω denotes a set of indices. We also define an indicator tensor $\mathcal{O}$, whose entry $\mathcal{O}_{i_1 i_2 \cdots i_N}$ is equal to 1 if $(i_1, i_2, \cdots, i_N) \in \Omega$, otherwise is equal to 0. We assume $\mathcal{Y}$ is a noisy measurement of the true latent tensor $\mathcal{X}$, and is corrupted by outliers $\mathcal{S}$, i.e., $\mathcal{Y} = \mathcal{X} + \mathcal{S} + \varepsilon$, where $\mathcal{X}$ is generated by tensor factorization with a *low-CP-rank*, representing the global information, $\mathcal{S}$ is enforced to be sparse, representing the local information, and ε is isotropic Gaussian noise (see Fig. 1).

The standard CP factorization [1] is expressed by

$$\mathcal{X} = \sum_{r=1}^R \mathbf{a}_{\cdot r}^{(1)} \circ \cdots \circ \mathbf{a}_{\cdot r}^{(N)} = [\![\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!], \tag{5}$$

where $\circ$ denotes the outer product of vectors and $[\![\cdots]\!]$ is a shorthand notation of CP factorization. $\{\mathbf{A}^{(n)} | n = 1, \ldots, N\}$ are latent factor matrices corresponding to each of N modes, respectively. CP model can be interpreted as a sum of R

rank-one tensors, which is related to the definition of *CP rank* that is the smallest integer R for which the above representation holds. The mode-n factor matrix of size $I_n \times R$ can be denoted by row-wise or column-wise vectors, that is, $\mathbf{A}^{(n)} = \left[\mathbf{a}_1^{(n)}, \ldots, \mathbf{a}_{I_n}^{(n)} \right]^T = \left[\mathbf{a}_{\cdot 1}^{(n)}, \ldots, \mathbf{a}_{\cdot R}^{(n)} \right]$.

To formulate robust CP factorization under the probabilistic framework, a generative model is introduced based on model assumptions. Specifically, the observation model is expressed by

$$p\left(\mathcal{Y}_\Omega \big| \{\mathbf{A}^{(n)}\}_{n=1}^N, \mathcal{S}_\Omega, \tau \right) = \prod_{i_1=1}^{I_1} \cdots \prod_{i_N=1}^{I_N}$$

$$\mathcal{N}\left(\mathcal{Y}_{i_1 \ldots i_N} \Big| \left\langle \mathbf{a}_{i_1}^{(1)}, \cdots, \mathbf{a}_{i_N}^{(N)} \right\rangle + \mathcal{S}_{i_1 \ldots i_N}, \tau^{-1} \right)^{\mathcal{O}_{i_1 \cdots i_N}}, \tag{6}$$

where τ denotes the noise precision, $\mathbf{a}_{i_n}^{(n)}$ denotes the i_nth row vector of $\mathbf{A}^{(n)}$, and $\mathcal{S}$ only has values corresponding to observed locations. The likelihood model in (6) indicates that $\mathcal{Y}_{i_1 \cdots i_N}$ is generated by multiple R-dimensional latent vectors $\left\{ \mathbf{a}_{i_n}^{(n)} \big| n = 1, \ldots, N \right\}$, whereas each $\mathbf{a}_{i_n}^{(n)}$ affects a set of observations, i.e., a subtensor whose mode-n index is i_n. The essential difference between matrix factorization and tensor factorization is that the generalized inner product of $N(\geq 3)$ latent vectors allows us to capture multilinear interactions reflecting the intrinsic structural property of data, which however leads to much more difficulties in model learning.

In practice, *CP rank*, i.e., the dimensionality of latent space denoted by R, is unknown and considered as a tuning parameter whose optimal selection is quite challenging especially in the presence of missing data. Since R controls the model complexity, we actually seek an automatic model selection strategy that can infer the true *CP rank* from partially observed data. To achieve this, in contrast to rank minimization on $\mathcal{X}$, we attempt to minimize the dimensionality of latent space, which corresponds to column-wise sparsity of factor matrices. Hence, we employ a sparsity inducing prior over factor matrices by associating an individual hyperparameter to each latent dimension. More specifically, a hierarchical prior is equally specified over N factor matrices, which is expressed by

$$p\left(\mathbf{A}^{(n)} \big| \boldsymbol{\lambda} \right) = \prod_{i_n=1}^{I_n} \mathcal{N}\left(\mathbf{a}_{i_n}^{(n)} \big| \mathbf{0}, \mathbf{\Lambda}^{-1} \right), \forall n \in [1, N]$$

$$p(\boldsymbol{\lambda}) = \prod_{r=1}^R \text{Ga}(\lambda_r | c_0, d_0), \tag{7}$$

where $\mathbf{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ denotes an inverse covariance matrix and is shared by latent factor matrices in all modes. The hyperprior over $\boldsymbol{\lambda}$ is an i.i.d. Gamma distribution $\text{Ga}(x|a, b) = \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}$ where $\Gamma(a)$ is the Gamma function.

Due to the sparsity property, the initialization of R is usually set to its maximum possible value, while the effective dimensionality can be inferred automatically under Bayesian inference framework. For instance, if a particular λ_r has a posterior distribution concentrated at large values, $\{\mathbf{a}_r^{(n)} | \forall n \in$
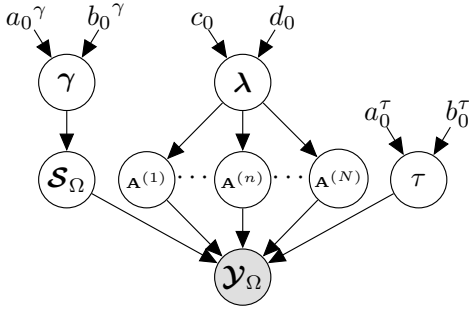
Fig. 2. The probabilistic graphical model of Bayesian robust CP factorization of an incomplete tensor.

$[1, N]\}$ will tend to be zero. Since the priors are shared by N factor matrices, our framework can learn the same sparsity pattern for all factor matrices, yielding the minimum number of rank-one tensors.

The sparse term $\boldsymbol{S}$ is modeled also by a hierarchical sparsity inducing prior. More specifically, Gaussian priors are placed on each data entry associated with an individual precision hyperparameter on which an i.i.d. Gamma hyperprior is placed, that is

$$
\begin{aligned}
p(\boldsymbol{S}_\Omega|\boldsymbol{\gamma}) &= \prod_{i_1,\ldots,i_N} \mathcal{N}(\mathcal{S}_{i_1\ldots i_N}|0, \gamma_{i_1\ldots i_N}^{-1})^{\mathcal{O}_{i_1\ldots i_N}}, \\
p(\boldsymbol{\gamma}) &= \prod_{i_1,\ldots,i_N} \mathrm{Ga}(\gamma_{i_1\ldots i_N}|a_0^\gamma, b_0^\gamma).
\end{aligned} \tag{8}
$$

Note that when an individual parameter $\gamma_{i_1\ldots i_N}$ goes to infinity, the corresponding element in $\boldsymbol{S}$ is enforced to be exact zero.

The priors in (7) and (8) are related to the framework of sparse Bayesian learning (SBL) [44] which is usually employed for variable selections. Since Laplacian and Student-t distributions are commonly applied to enforcing sparsity, we may question why the choice of a Gaussian prior should express any preference for sparsity. In fact, (8) can be interpreted as an infinite zero-mean Gaussian mixture with mixture coefficients drawn from a Gamma distribution, which is thus a hierarchical view of Student-t distribution. In other words, the marginal prior of $\boldsymbol{S}_\Omega$ is an i.i.d. Student-t distribution with the sparsity controlled by (a_0^γ, b_0^γ) to some extent. For the case of noninformative hyperprior with $a_0^\gamma = b_0^\gamma = 0$, we obtain the improper marginal prior $p(\mathcal{S}_{i_1\ldots i_N}) \propto 1/|\mathcal{S}_{i_1\ldots i_N}|$. Note that if $a_0^\gamma = 1$, the hyperprior becomes an exponential distribution, such that the marginal prior over $\boldsymbol{S}_\Omega$ is a Laplacian distribution. The elegance of this strategy therefore lies in the use of hierarchical modeling to obtain a prior which encourages sparsity while keeping fully conjugate exponential-family distributions throughout, which leads to the possibility of the fully Bayesian treatment. Although our setting is related to SBL, the crucial difference lies in that our model specification can achieve column-wise sparsity, and the statistical property is shared by a set of factor matrices $\{\mathbf{A}^{(n)}\}_{n=1}^N$.

To complete the model, we also place a hyperprior over the noise precision τ, that is

$$
p(\tau) = \mathrm{Ga}(\tau|a_0^\tau, b_0^\tau). \tag{9}
$$

Finally, the probabilistic graphical model of robust tensor factorization is illustrated in Fig. 2. For simplicity of notations, all unknowns including both latent factor matrices and hyperparameters are collected and denoted together by $\Theta = \{\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}, \boldsymbol{\lambda}, \boldsymbol{S}_\Omega, \boldsymbol{\gamma}, \tau\}$. Therefore, the joint distribution of the model, i.e., $p(\boldsymbol{\mathcal{Y}}_\Omega, \Theta)$, can be expressed by

$$
p\left(\boldsymbol{\mathcal{Y}}_\Omega \Big| \{\mathbf{A}^{(n)}\}_{n=1}^N, \boldsymbol{S}_\Omega, \tau\right) \prod_{n=1}^N p(\mathbf{A}^{(n)}|\boldsymbol{\lambda})p(\boldsymbol{S}_\Omega|\boldsymbol{\gamma})p(\boldsymbol{\lambda})p(\boldsymbol{\gamma})p(\tau).
$$

In general, we can simply perform MAP estimation of Θ from the log-joint distribution (see Sec. 1 of Appendix) and most existing tensor factorization based on optimization approaches can be interpreted as point estimation by either maximum likelihood or MAP principles. However, in this study, we aim to provide a fully Bayesian treatment of the model by inferring the posterior distribution of Θ, expressed by $p(\Theta|\boldsymbol{\mathcal{Y}}_\Omega) = \frac{p(\Theta, \boldsymbol{\mathcal{Y}}_\Omega)}{\int p(\Theta, \boldsymbol{\mathcal{Y}}_\Omega) d\Theta}$. Thus the predictive distribution over missing entries $\boldsymbol{\mathcal{Y}}_{\backslash\Omega}$ can be also inferred by $p(\boldsymbol{\mathcal{Y}}_{\backslash\Omega}|\boldsymbol{\mathcal{Y}}_\Omega) = \int p(\boldsymbol{\mathcal{Y}}_{\backslash\Omega}|\Theta)p(\Theta|\boldsymbol{\mathcal{Y}}_\Omega) d\Theta$.

B. Model Learning via Bayesian Inference

Since exact Bayesian inference of our model is analytically intractable, we must resort to the approximate inference. Although variational Bayesian (VB) inference [45] is difficult for derivations, especially when multiple interactions of latent factors are involved, it has advantages of closed-form posterior approximations and high efficiency as compared to sampling based inference methods. Hence, we employ VB inference to learn our model and present only the main results, while the detailed derivations and proofs are provided in Appendix[1].

We therefore seek a distribution $q(\Theta)$ to approximate the true posterior distribution $p(\Theta|\boldsymbol{\mathcal{Y}}_\Omega)$ in the sense of minimizing the KL divergence, that is

$$
\begin{aligned}
\mathrm{KL}\big(q(\Theta)\big|\big|p(\Theta|\boldsymbol{\mathcal{Y}}_\Omega)\big) &= \ln p(\boldsymbol{\mathcal{Y}}_\Omega) - \mathcal{L}(q), \\
\text{where} \quad \mathcal{L}(q) &= \int q(\Theta) \ln \left\{ \frac{p(\boldsymbol{\mathcal{Y}}_\Omega, \Theta)}{q(\Theta)} \right\} d\Theta.
\end{aligned} \tag{10}
$$

$\ln p(\boldsymbol{\mathcal{Y}}_\Omega)$ denotes the model evidence that is a constant, and its *lower bound* is denoted by $\mathcal{L}(q)$. Thus, minimum of KL divergence implies the maximum of $\mathcal{L}(q)$. By applying mean field approximation, we assume that the posteriors can be factorized as

$$
q(\Theta) = \prod_{n=1}^N q\left(\mathbf{A}^{(n)}\right) q(\boldsymbol{S}_\Omega)q(\boldsymbol{\lambda})q(\boldsymbol{\gamma})q(\tau). \tag{11}
$$

Note that this is the only assumption about the distribution, while the particular functional forms of the individual factors can be explicitly derived in turn by virtue of conjugate exponential family in our hierarchical model.

1) Posterior distribution of factor matrices: From the graphical model shown in Fig. 2, the inference of mode-n factor matrix $\mathbf{A}^{(n)}$ can be performed by receiving the messages from observed data, which are expressed by the likelihood term (6), and incorporating the messages from their parents, which are expressed by the prior term (7). The posteriors are

---

[1]The Appendix is provided in supplementary materials.

shown to be factorized as independent distributions of their rows which are also Gaussian (see Sec. 2 of Appendix for details), i.e.,

$$q(\mathbf{A}^{(n)}) = \prod_{i_n=1}^{I_n} \mathcal{N}\left(\mathbf{a}_{i_n}^{(n)}\Big|\tilde{\mathbf{a}}_{i_n}^{(n)}, \mathbf{V}_{i_n}^{(n)}\right), \ \forall n \in [1, N], \quad (12)$$

where the posterior parameters can be updated by

$$\tilde{\mathbf{a}}_{i_n}^{(n)} = \mathbb{E}_q[\tau]\mathbf{V}_{i_n}^{(n)}\mathbb{E}_q\big[\mathbf{A}_{i_n}^{(\backslash n)T}\big]\mathrm{vec}\left(\boldsymbol{\mathcal{Y}} - \mathbb{E}_q[\boldsymbol{\mathcal{S}}]\right)_{\mathbb{I}(\mathcal{O}_{i_n}=1)}$$
$$\mathbf{V}_{i_n}^{(n)} = \left(\mathbb{E}_q[\tau]\mathbb{E}_q\big[\mathbf{A}_{i_n}^{(\backslash n)T}\mathbf{A}_{i_n}^{(\backslash n)}\big] + \mathbb{E}_q[\boldsymbol{\Lambda}]\right)^{-1}. \quad (13)$$

$\mathbb{E}_q[\cdot]$ denotes the posterior expectation w.r.t. all variables involved. $\mathbb{I}(\mathcal{O}_{i_n} = 1)$ is a sample function denoting a subset of the observed entries, whose mode-n index is i_n. The most complex term in (13) is $\mathbf{A}_{i_n}^{(\backslash n)T} = \left(\bigodot_{k \neq n} \mathbf{A}^{(k)}\right)^T_{\mathbb{I}(\mathcal{O}_{i_n}=1)}$, where $(\cdot)_{\mathbb{I}(\mathcal{O}_{i_n}=1)}$ denotes a subset of columns sampled according to the subtensor $\mathcal{O}_{i_n} = 1$. Note that the update of $\mathbf{V}_{i_n}^{(n)}$ involves expectation of the Khatri-Rao product, which can not be evaluated straightforwardly. Hence, we introduce the following results:

Lemma IV.1. *Given a set of independent random matrices $\{\mathbf{A}^{(n)}|n = 1, \ldots, N\}$, we assume that $\forall n, \forall i_n$, the row vectors $\{\mathbf{a}_{i_n}^{(n)}\}$ are independent, then*

$$\mathbb{E}\left[\left(\bigodot_n \mathbf{A}^{(n)}\right)^T \left(\bigodot_n \mathbf{A}^{(n)}\right)\right] = \sum_{i_1,\ldots,i_N} \underset{n}{\circledast}\left(\mathbb{E}\left[\mathbf{a}_{i_n}^{(n)}\mathbf{a}_{i_n}^{(n)T}\right]\right).$$

Proof. See Sec. 3 of Appendix for details. $\square$

According to Lemma IV.1, we can obtain that

$$\mathbb{E}_q\big[\mathbf{A}_{i_n}^{(\backslash n)T}\mathbf{A}_{i_n}^{(\backslash n)}\big] = \sum_{(i_1,\ldots,i_N)\in\Omega} \underset{k\neq n}{\circledast}\left(\mathbb{E}\left[\mathbf{a}_{i_k}^{(k)}\mathbf{a}_{i_k}^{(k)T}\right]\right). \quad (14)$$

For simplicity, let $\mathbf{B}^{(n)}$ of size $I_n \times R^2$ denote an expectation of a quadratic form related to $\mathbf{A}^{(n)}$ by defining i_nth-row vector $\mathbf{b}_{i_n}^{(n)} = \mathrm{vec}\left(\mathbb{E}_q\left[\mathbf{a}_{i_n}^{(n)}\mathbf{a}_{i_n}^{(n)T}\right]\right) = \mathrm{vec}\left(\tilde{\mathbf{a}}_{i_n}^{(n)}\tilde{\mathbf{a}}_{i_n}^{(n)T} + \mathbf{V}_{i_n}^{(n)}\right)$, then (14) can be written as

$$\mathrm{vec}\left(\mathbb{E}_q\big[\mathbf{A}_{i_n}^{(\backslash n)T}\mathbf{A}_{i_n}^{(\backslash n)}\big]\right) = \left(\bigodot_{k\neq n} \mathbf{B}^{(k)}\right)^T \mathrm{vec}(\boldsymbol{\mathcal{O}}_{\cdots i_n \cdots}). \quad (15)$$

Note that the Khatri-Rao product in (15) is computed by all mode factors except nth mode, while the sum is performed according to the indices of observations, implying that only factors that interact with $\mathbf{a}_{i_n}^{(n)}$ are taken into account.

An intuitive interpretation of (13) is given as follows. $\mathbf{V}_{i_n}^{(n)}$ is updated by combining $\mathbb{E}_q[\boldsymbol{\Lambda}]$, denoting the factor prior, and covariance of other factor matrices computed by (15), while the tradeoff between these two terms is controlled by $\mathbb{E}_q[\tau]$ that is related to model fitness. In other words, the better fitness leads to more information from the current model than from the factor prior. $\tilde{\mathbf{a}}_{i_n}^{(n)}$ is updated firstly by a linear combination of all other factors, while the combination coefficients are observed values, which implies that the larger observation leads to more similarity of its corresponding latent factors. Subsequently, $\tilde{\mathbf{a}}_{i_n}^{(n)}$ is rotated by $\mathbf{V}_{i_n}^{(n)}$ and is scaled according to the model fitness $\mathbb{E}_q[\tau]$.

2) Posterior distribution of hyperparameters $\boldsymbol{\lambda}$: From Fig. 2, the inference of $\boldsymbol{\lambda}$ can be performed by receiving messages from N factor matrices and incorporating the messages from its hyperprior. We can show the posteriors of $\lambda_r, \forall r \in [1, R]$ are independent Gamma distribution, $q(\boldsymbol{\lambda}) = \prod_{r=1}^R \mathrm{Ga}(\lambda_r|c_M^r, d_M^r)$, where c_M^r, d_M^r denote the posterior parameters learned from M observations and can be updated by (see Sec. 4 of Appendix for details)

$$c_M^r = c_0 + \frac{1}{2}\sum_{n=1}^N I_n, \quad d_M^r = d_0 + \frac{1}{2}\sum_{n=1}^N \mathbb{E}_q\left[\mathbf{a}_{\cdot r}^{(n)T}\mathbf{a}_{\cdot r}^{(n)}\right]. \quad (16)$$

The posterior expectation term in (16) can be evaluated using the posterior parameters in (13), thus we have $\mathbb{E}_q\left[\mathbf{a}_{\cdot r}^{(n)T}\mathbf{a}_{\cdot r}^{(n)}\right] = \tilde{\mathbf{a}}_{\cdot r}^{(n)T}\tilde{\mathbf{a}}_{\cdot r}^{(n)} + \sum_{i_n}\left(\mathbf{V}_{i_n}^{(n)}\right)_{rr}$. Therefore, we can further simplify the computation of $\mathbf{d}_M = [d_M^1, \ldots d_M^R]^T$ by

$$\mathbf{d}_M = \sum_{n=1}^N \left\{\mathrm{diag}\left(\tilde{\mathbf{A}}^{(n)T}\tilde{\mathbf{A}}^{(n)} + \sum_{i_n}\mathbf{V}_{i_n}^{(n)}\right)\right\}. \quad (17)$$

Based on the updated posterior of $\boldsymbol{\lambda}$, we can obtain $\mathbb{E}_q[\boldsymbol{\Lambda}] = \mathrm{diag}([c_M^1/d_M^1, \ldots, c_M^R/d_M^R])$.

An intuitive interpretation is that the smaller $\sum_n \|\mathbf{a}_{\cdot r}^{(n)}\|_2^2$ leads to larger $\mathbb{E}_q[\lambda_r]$, which thus updates the prior over $\{\mathbf{a}_{\cdot r}^{(n)}\}_{n=1}^N$, resulting in that the rth component is enforced more strongly to be zero. Therefore, the smaller components can be diminished eventually to exact zero and effectively pruned out after several iterations, while the larger components are enhanced to explain the data. This sparsity technique plays an key role to obtain the minimum number of components and automatic rank determination.

3) Posterior distribution of sparse tensor $\boldsymbol{\mathcal{S}}$: By combining the priors in (8) and the likelihood in (6), we can derive the posterior approximation of $\boldsymbol{\mathcal{S}}$ as (see Sec. 5 of Appendix for details)

$$q(\boldsymbol{\mathcal{S}}) = \prod_{(i_1,\ldots,i_N)\in\Omega} \mathcal{N}\left(\mathcal{S}_{i_1\ldots i_N}\Big|\tilde{\mathcal{S}}_{i_1\ldots i_N}, \sigma_{i_1\ldots i_N}^2\right), \quad (18)$$

where the posterior parameters can be updated by

$$\tilde{\mathcal{S}}_{i_1\ldots i_N} = \sigma_{i_1\ldots i_N}^2 \mathbb{E}_q[\tau]\left(\mathcal{Y}_{i_1\ldots i_N} - \mathbb{E}_q\left[\left\langle\mathbf{a}_{i_1}^{(1)}, \ldots, \mathbf{a}_{i_N}^{(N)}\right\rangle\right]\right),$$
$$\sigma_{i_1\ldots i_N}^2 = (\mathbb{E}_q[\gamma_{i_1\ldots i_N}] + \mathbb{E}_q[\tau])^{-1}. \quad (19)$$

Observe that $\boldsymbol{\mathcal{S}}$ captures the information which is not explained by the low-rank CP approximation, while the magnitude is controlled by $\sigma_{i_1\ldots i_N}$ that is affected by the prior parameter $\mathbb{E}_q[\gamma_{i_1\ldots i_N}]$ and the precision of Gaussian noise $\mathbb{E}_q[\tau]$. The intuitive interpretation is that $\boldsymbol{\mathcal{S}}$ can model individual noises from total residuals, which are non-Gaussian. An alternative interpretation is that $[\![\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!]$ explains the global information by using minimum number of rank-one tensors, while $\boldsymbol{\mathcal{S}}$ explains the local information that is too expensive to be represented by increasing the model complexity.

4) Posterior distribution of hyperparameters γ: By incorporating the prior and hyperprior of $\mathcal{S}_\Omega$ in (8), we show that the posterior of γ is also factorized as independent distributions of each entries (see Sec. 6 of Appendix), given by

$$q(\boldsymbol{\gamma}) = \prod_{(i_1,\ldots,i_N)\in\Omega} \mathrm{Ga}(\gamma_{i_1\ldots i_N}|a_M^{\gamma_{i_1\ldots i_N}}, b_M^{\gamma_{i_1\ldots i_N}}), \quad (20)$$

whose posterior parameters can be updated by

$$a_M^{\gamma_{i_1\ldots i_N}} = a_0^\gamma + \frac{1}{2}, \; b_M^{\gamma_{i_1\ldots i_N}} = b_0^\gamma + \frac{1}{2}(\tilde{\mathcal{S}}_{i_1\ldots i_N}^2 + \sigma_{i_1\ldots i_N}^2). \quad (21)$$

This indicates that the smaller $\mathbb{E}_q[\mathcal{S}_{i_1\ldots i_N}^2]$ leads to larger $\mathbb{E}_q[\gamma_{i_1\ldots i_N}]$ which enforces $\tilde{\mathcal{S}}_{i_1\ldots i_N}$ to be zero more strongly by (19), and vice versa. In other words, the elements with small magnitude are forced to be zero, while the elements with large magnitude are further enhanced. It should be noted that the sparsity on $\mathcal{S}$ is essentially important due to the fact that Gaussian with individual hyperparameters can easily capture the whole information of data.

5) Posterior distribution of hyperparameter τ: The inference of noise precision τ can be performed by receiving the messages from observed data, and incorporating the messages from its hyperprior. We can show that the variational posterior is a Gamma distribution (see Sec. 7 of Appendix), i.e., $q(\tau) = \mathrm{Ga}(\tau|a_M^\tau, b_M^\tau)$ where the posterior parameters can be updated by

$$a_M^\tau = a_0^\tau + \frac{1}{2} \sum_{i_1,\ldots,i_N} \mathcal{O}_{i_1\ldots i_N},$$
$$b_M^\tau = b_0^\tau + \frac{1}{2}\mathbb{E}_q\left[\left\|\boldsymbol{\mathcal{O}} \circledast \left(\boldsymbol{\mathcal{Y}} - [\![\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(N)}]\!] - \boldsymbol{\mathcal{S}}\right)\right\|_F^2\right]. \quad (22)$$

However, the posterior expectation of model residuals in the above expression can not be computed straightforward, we need to introduce the following results.

Lemma IV.2. *Given a set of independent random matrices $\{\mathbf{A}^{(n)}|n = 1,\ldots,N\}$, we assume that $\forall n, \forall i_n$, the row vectors $\{\mathbf{a}_{i_n}^{(n)}\}$ are independent, then*

$$\mathbb{E}\left[\left\|[\![\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(N)}]\!]\right\|_F^2\right]$$
$$= \sum_{i_1,\ldots,i_N} \left\langle \mathbb{E}\left[\mathbf{a}_{i_1}^{(1)}\mathbf{a}_{i_1}^{(1)T}\right],\ldots,\mathbb{E}\left[\mathbf{a}_{i_N}^{(N)}\mathbf{a}_{i_N}^{(N)T}\right]\right\rangle. \quad (23)$$

Proof. See Sec. 8 of Appendix for details. □

By using Lemma IV.2, the posterior expectation in (22) can be evaluated explicitly (see Sec. 9 of Appendix for details), that is

$$\mathbb{E}_q\left[\left\|\boldsymbol{\mathcal{O}} \circledast \left(\boldsymbol{\mathcal{Y}} - [\![\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(N)}]\!] - \boldsymbol{\mathcal{S}}\right)\right\|_F^2\right]$$
$$= \|\boldsymbol{\mathcal{Y}}_\Omega\|_F^2 - 2\mathrm{vec}^T(\boldsymbol{\mathcal{Y}}_\Omega)\mathrm{vec}\left([\![\tilde{\mathbf{A}}^{(1)},\ldots,\tilde{\mathbf{A}}^{(N)}]\!]_\Omega\right)$$
$$+ \mathrm{vec}^T(\boldsymbol{\mathcal{O}})\left(\bigodot_n \mathbf{B}^{(n)}\right)\mathbf{1}_{R^2} - 2\mathrm{vec}^T(\boldsymbol{\mathcal{Y}}_\Omega)\mathrm{vec}(\tilde{\boldsymbol{\mathcal{S}}}_\Omega)$$
$$+ 2\mathrm{vec}^T([\![\tilde{\mathbf{A}}^{(1)},\ldots,\tilde{\mathbf{A}}^{(N)}]\!]_\Omega)\mathrm{vec}(\tilde{\boldsymbol{\mathcal{S}}}_\Omega) + \mathbb{E}_q[\|\boldsymbol{\mathcal{S}}_\Omega\|_F^2]. \quad (24)$$

Hence, the posterior expectation of τ can be updated by $\mathbb{E}_q[\tau] = a_M^\tau/b_M^\tau$, where a_M^τ is related to the number of observations and b_M^τ is related to the posterior expectation of model residuals measured by squared Frobenius norm.

6) Lower bound of model evidence: We can also evaluate the variational lower bound in (10) for our model. Since at each step of the iterative re-estimation procedure the value of this bound should not decrease, we can monitor the bound in order to test for convergence. The lower bound on the log marginal likelihood can be also written as

$$\mathcal{L}(q) = \mathbb{E}_{q(\Theta)}[\ln p(\boldsymbol{\mathcal{Y}}_\Omega, \Theta)] + H(q(\Theta)), \quad (25)$$

where the first term denotes the posterior expectation of joint probability density, and the second term denotes the entropy of q distribution. Taking the parametric form of q distributions derived in the previous section, it can then be evaluated by an explicit form (See Sec. 10 of Appendix for details).

The top level hyperparameters $a_0^\gamma, b_0^\gamma, a_0^\tau, b_0^\tau, c_0, d_0$ are usually fixed to be very small values leading to a noninformative prior or set to zero leading to a Jeffrey's prior. Note that a_0^γ, b_0^γ are related to sparsity degree, we seek a strategy to automatically adopt these hyperparameters to various types of outliers. This can be easily achieved by maximizing the lower bound w.r.t. a_0^γ, b_0^γ, expressed by

$$\mathcal{L}(a_0^\gamma, b_0^\gamma) = -M\ln\Gamma(a_0^\gamma) + Ma_0^\gamma\ln b_0^\gamma$$
$$+ (a_0^\gamma - 1)\sum_{(i_1\ldots i_N)\in\Omega}\left\{(\psi(a_M^{\gamma_{i_1\ldots i_N}}) - \ln b_M^{\gamma_{i_1\ldots i_N}})\right\}$$
$$- b_0^\gamma \sum_{(i_1\ldots i_N)\in\Omega} \frac{a_M^{\gamma_{i_1\ldots i_N}}}{b_M^{\gamma_{i_1\ldots i_N}}}. \quad (26)$$

7) Initialization of model inference: The variational Bayesian inference is only guaranteed to converge to a local minimum. To alleviate getting stuck in poor local solutions, it is important to choose an initialization point. In our model, the top level hyperparameters including $c_0, d_0, a_0^\tau, b_0^\tau, a_0^\gamma, b_0^\gamma$ are set to 10^{-6}, resulting in a noninformative prior. Thus the expectation of hyperparameters can be initialized by $\mathbb{E}[\boldsymbol{\Lambda}] = \mathbf{I}$, $\mathbb{E}[\tau] = 1$ and $\forall n, \forall i_n, \mathbb{E}[\gamma_{i_1\ldots i_N}] = 1$. For the factor matrices, $\mathbb{E}[\mathbf{A}^{(n)}], \forall n \in [1, N]$ can be initialized by two different schemes. One is randomly drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ for each row vector $\{\mathbf{a}_{i_n}^{(n)}\}$. The other is set to $\mathbf{A}^{(n)} = \mathbf{U}^{(n)}\boldsymbol{\Sigma}^{(n)\frac{1}{2}}$, where $\mathbf{U}^{(n)}$ denotes the left singular vectors and $\boldsymbol{\Sigma}^{(n)}$ denotes the diagonal singular values matrix, obtained by SVD of mode-n matricization of $\boldsymbol{\mathcal{Y}}$. $\mathbf{V}^{(n)}$ is simply set to $\mathbb{E}[\boldsymbol{\Lambda}^{-1}]$. For sparse tensor $\boldsymbol{\mathcal{S}}$, $\mathbb{E}[\mathcal{S}_{i_1\ldots i_N}]$ is drawn from $\mathcal{N}(0, 1)$, while $\sigma_{i_1\ldots i_N}^2$ is set to $\mathbb{E}[\gamma_{i_1\ldots i_N}^{-1}]$. The tensor rank R is usually initialized by the maximum rank, i.e. $R \leq \min_n P_n$, where $P_n = \prod_{i\neq n} I_i$. For efficiency, we can also manually set the initialization value of R.

The whole procedure of model inference is summarized in Algorithm 1, where the posterior factors in (11) are updated in an order that from bottom to top (see Fig. 2), which indicates that the message passing is started from observed data.

Algorithm 1 Bayesian Robust Tensor Factorization

Input: An Nth-order incomplete tensor $\mathcal{Y}$ and an indicator tensor $\mathcal{O}$.
Initialization: $\tilde{\mathbf{A}}^{(n)}, \mathbf{V}^{(n)}_{i_n}, \forall i_n \in [1, I_n], \forall n \in [1, N]$, $\tilde{\mathcal{S}}, \sigma^2$, hyperparameters $\boldsymbol{\lambda}, \boldsymbol{\gamma}, \tau$, top level hyperparameters $c_0, d_0, a_0^\gamma, b_0^\gamma, a_0^\tau, b_0^\tau$.
repeat
    for $n = 1$ **to** N **do**
        Update the posterior $q(\mathbf{A}^{(n)})$ by (13);
    end for
    Update the posterior $q(\boldsymbol{\lambda})$ by (16);
    Update the posterior $q(\tau)$ by (22);
    Update the posterior $q(\mathcal{S})$ by (19);
    Update the posterior $q(\boldsymbol{\gamma})$ by (21);
    Evaluate the lower bound by (25);
    Update a_0^γ, b_0^γ by maximizing (26);
    Model reduction by eliminating zero-components in $\{\mathbf{A}^{(n)}\}$;
until convergence.

C. Predictive Distribution

The predictive distribution over missing entries, given observed entries, is also analytically intractable. Hence, we can approximate it by using the variational posteriors of all parameters in Θ, yielding a Student-t distribution (see Sec. 11 of Appendix for details)

$$p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{Y}}_\Omega) = \prod_{i_1,\dots,i_N} \mathcal{T}(\mathcal{Y}_{i_1\dots i_N}|\tilde{\mathcal{Y}}_{i_1\dots i_N}, \Psi_{i_1\dots i_N}, \nu_y) \quad (27)$$

with its parameters given by

$$\tilde{\mathcal{Y}}_{i_1\dots i_N} = \left\langle \tilde{\mathbf{a}}^{(1)}_{i_1}, \cdots, \tilde{\mathbf{a}}^{(n)}_{i_N} \right\rangle,$$

$$\Psi_{i_1\dots i_N} = \left\{ \frac{b_M^\tau}{a_M^\tau} + \sum_n \left\{ \left(\circledast_{k\neq n} \tilde{\mathbf{a}}^{(k)}_{i_k} \right)^T \mathbf{V}^{(n)}_{i_n} \left(\circledast_{k\neq n} \tilde{\mathbf{a}}^{(k)}_{i_k} \right) \right\} \right\}^{-1},$$

and $\nu_y = 2a_M^\tau$. Thus, the uncertainty of predictions can be obtained by $\text{var}(\mathcal{Y}_{i_1\dots i_N}) = \frac{\nu_y}{\nu_y - 2} \Psi^{-1}_{i_1\dots i_N}$.

D. Computational Complexity

The computation cost of N factor matrices in (13) is $O(R^2 M \sum_n I_n + R^3 \sum_n I_n)$, where M denotes the number of observations, R denotes model complexity and is generally much smaller than the data size, i.e., $R \ll M$. The computational costs are $O(R^2 \sum_n I_n)$ for $\boldsymbol{\lambda}$, $O(R^2 M)$ for τ, and $\mathcal{O}(MNR)$ for $\mathcal{S}$. Therefore, the overall complexity of our algorithm is $O((R^2 M + R^3) \sum_n I_n)$, which scales linearly with the data size but polynomially with the tensor rank. Note that due to the automatic model reduction, the excessive latent components are pruned out in the first few iterations such that R reduces rapidly in practice. The solution presented in this paper mainly focuses on a general tensor factorization problem. However, when data is extremely sparse or large scale, an alternative strategy for approximate inference updated by each observed entry can be developed correspondingly.

E. Case of Complete Tensor

For fully observed tensor data, we can simply define $\mathcal{O}$ with all elements being 1 and apply the model inference as described previously. However, several essentially different properties arise during inference, which leads to the possibility of more efficient computation for approximate posteriors. For the inference of factor matrices shown in (13), since $\mathbf{A}^{(\backslash n)}_{i_n}$ are same for any $i_n \in [1, I_n]$, such that $\{\mathbf{V}^{(n)}_{i_n}\}^{I_n}_{i_n=1}$ are all equivalent. Hence, only one $\mathbf{V}^{(n)}$ needs to be computed for each mode-n, and $\{\tilde{\mathbf{a}}^{(n)}_{i_n}\}^{I_n}_{i_n=1}$ can be updated simultaneously by

$$\tilde{\mathbf{A}}^{(n)} = \mathbb{E}_q[\tau] \left(\mathbf{Y}_{(n)} - \mathbb{E}_q[\mathbf{S}_{(n)}] \right) \mathbb{E}_q\left[\mathbf{A}^{(\backslash n)}\right] \mathbf{V}^{(n)},$$
$$\mathbf{V}^{(n)} = \left(\mathbb{E}_q[\tau]\mathbb{E}_q\left[\mathbf{A}^{(\backslash n)T}\mathbf{A}^{(\backslash n)}\right] + \mathbb{E}_q[\boldsymbol{\Lambda}] \right)^{-1}, \quad (28)$$

where $\mathbf{Y}_{(n)}$ denotes mode-n matricization of tensor $\mathcal{Y}$. For computational efficiency, we introduce another solution related to Lemma IV.1.

Lemma IV.3. *Given a set of matrices $\{\mathbf{A}^{(n)}|n = 1,\dots,N\}$, if the row vectors $\{\mathbf{a}^{(n)}_{i_n}\}^{I_n}_{i_n=1}$ are independent and $cov[\mathbf{a}^{(n)}_{i_n}] = \mathbf{V}^{(n)}, \forall i_n \in [1, I_n]$, then*

$$\mathbb{E}\left[\left(\bigodot_n \mathbf{A}^{(n)} \right)^T \left(\bigodot_n \mathbf{A}^{(n)} \right) \right] = \circledast_n \left(\mathbb{E}\left[\mathbf{A}^{(n)T}\mathbf{A}^{(n)} \right] \right),$$

where $\mathbb{E}[\mathbf{A}^{(n)T}\mathbf{A}^{(n)}] = \left\{ \mathbb{E}[\mathbf{A}^{(n)T}]\mathbb{E}[\mathbf{A}^{(n)}] + I_n \mathbf{V}^{(n)} \right\}$.

Proof. See Sec. 12 of Appendix for details. $\square$

According to Lemma IV.3, the term in (28) can be computed efficiently by

$$\mathbb{E}_q\left[\mathbf{A}^{(\backslash n)T}\mathbf{A}^{(\backslash n)}\right] = \circledast_{k\neq n} \left\{ \tilde{\mathbf{A}}^{(k)T}\tilde{\mathbf{A}}^{(k)} + I_k \mathbf{V}^{(k)} \right\}. \quad (29)$$

Hence the computational cost for factor matrices are reduced to $\mathcal{O}(R^2 \sum_n I_n + NR^3)$. For hyperparameters $\boldsymbol{\lambda}$, the update rules in (17) can be simplified by

$$\mathbf{d}_M = \sum_{n=1}^N \left\{ \text{diag}\left(\tilde{\mathbf{A}}^{(n)T}\tilde{\mathbf{A}}^{(n)} + I_n \mathbf{V}^{(n)} \right) \right\}. \quad (30)$$

The inference for $\mathcal{S}$ and $\boldsymbol{\gamma}$ are similar to the case of incomplete data. For hyperparameter τ, the posterior expectation of squared Frobenius norm of CP approximation can be computed more efficiently by introducing the following result.

Lemma IV.4. *Given a set of independent random matrices $\{\mathbf{A}^{(n)}|n = 1,\dots,N\}$, we assume that $\forall i_n \in [1, I_n]$, the row vectors $\{\mathbf{a}^{(n)}_{i_n}\}$ are independent, then*

$$\mathbb{E}\left[\left\| [\![\mathbf{A}^{(1)},\dots,\mathbf{A}^{(N)}]\!] \right\|_F^2 \right] =$$
$$\left\langle \mathbb{E}\left[\mathbf{A}^{(1)T}\mathbf{A}^{(1)}\right], \dots, \mathbb{E}\left[\mathbf{A}^{(N)T}\mathbf{A}^{(N)}\right] \right\rangle.$$

Proof. See Sec. 13 of Appendix for details. $\square$

Hence, the computational cost for τ is reduced to $\mathcal{O}(R^2 \sum_n I_n)$. In addition, the computation of lower bound and predictive distributions can be also simplified easily, which would not been presented in details.

V. ADVANTAGES

Since our model is based on a hierarchical probabilistic framework and fully Bayesian treatment, several advantages are gained and discussed as follows:

- Our method is characterized as a *tuning parameter free* approach and all model parameters can be learned automatically from observed data. By contrast, the existing tensor factorization methods require either predefined rank or penalty parameter and tensor completion methods using nuclear norm need to tune regularization parameters.

- The *automatic rank determination* enables us to discover the ground-truth of CP rank, while the *automatic sparsity model* can adapt the model to various types of outliers or non-Gaussian noises. Furthermore, the most elegant characteristic is that the tradeoff between the low-rank approximation and the sparse representation can be learned automatically in the sense of maximizing the model evidence.

- In contrast to point estimations by most existing tensor methods, the *uncertainty information* over all model parameters are taken into account, which can effectively prevent overfitting problem. The full posteriors of factor matrices and predicted missing entries can provide confidence information regarding the solutions.

- An efficient and deterministic algorithm for Bayesian inference is developed, which empirically shows a *fast convergence* and its computational complexity scales linearly with the data size.

VI. EXPERIMENTAL RESULTS

A. Validation on Synthetic Data

We firstly assess the performance quantitatively on synthetic data. The true low-rank tensor $\mathcal{X}$ of size $30 \times 30 \times 30$ was generated by rank-3 factor matrices, i.e., $\mathbf{A}^{(n)} \in \mathbb{R}^{30 \times 3}, n = 1, 2, 3$. Three components of the nth factor matrix are $[\sin(2\pi \frac{n}{30} i_n), \cos(2\pi \frac{n}{30} i_n), \text{sgn}(\sin(0.5\pi i_n))]$, indicating that the first two components possess different frequencies related to n, and the third components are common in all modes. A random fraction of tensor entries were corrupted by outliers drawn from an uniform distribution $\mathcal{U}(-|H|, |H|)$. To mimic more realistic settings, a small noise drawn from $\mathcal{N}(0, 0.01)$ was also considered. Subsequently, a fraction of entries were randomly selected to be observed tensor $\mathcal{Y}_\Omega$, when missing data was considered. We utilized the root relative square error (RRSE), defined by $\frac{\|\hat{\mathcal{X}} - \mathcal{X}\|_2}{\|\mathcal{X}\|_2}$, to evaluate the performance of tensor recovery. As for recovering the underlying factors, factor match error (FME) [10] between the estimated factors and ground-truth was also evaluated. We compared our Bayesian robust tensor factorization (BRTF) with state-of-the-art methods including tensor factorization (CP-ALS [1], HOSVD [7], CP-ARD [19]), tensor factorization with missing data (CPWOPT [9] and CPNLS [10]), nonparametric Bayesian tensor factorization (MGPCP [20]) and robust tensor factorization with missing data (HORPCA [38]). It should be emphasized that the tuning parameters are necessary for most methods and have been carefully tuned based on

TABLE I
RESULTS ON A COMPLETE TENSOR $\mathcal{Y}$ OF SIZE $30 \times 30 \times 30$ WITH $R = 50$

	BRTF	HORPCA	CP-ALS	HOSVD	CP-ARD
Rank	50 (Auto)	N/A	48	(23,23,23)	68 (Auto)
RRSE	0.04	0.20	0.31	0.56	0.35
FME	0.02	0.96	0.29	0.96	0.20
Sensitivity	N/A	0.32	0.20	0.14	N/A
Runtime	3s	10s	31s	5s	4s

ground-truth data, which is generally impractical for real applications. Specifically, CP-ALS, HOSVD, CPWOPT, CPNLS need to tune the parameter of *tensor rank*, and HORPCA needs to tune the penalty parameter. By contrast, BRTF, MGPCP and CP-ARD can automatically estimate tensor rank and do not require any tuning parameters. Two scenarios were considered: 1) tensor recovery from fully observed data, and 2) tensor completion from partially observed data.

(a) $H = 10 \cdot \text{std}(\text{vec}(\mathcal{X}))$
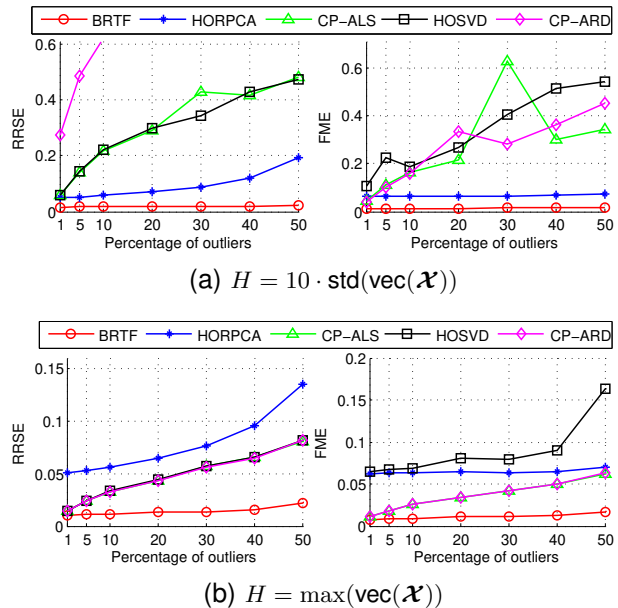
(b) $H = \max(\text{vec}(\mathcal{X}))$

Fig. 3. Performance of recovering true tensor $\mathcal{X}$ and latent factors $\{\mathbf{A}^{(n)}\}_{n=1}^N$ under varying percentages of outliers. There are two types of outliers: (a) the magnitude is much larger than that of true signals, and (b) is within the range of true signals.

The results for fully observed tensor are shown in Fig. 3. Rank R was initialized as $\max_n I_n$ for BRTF and CP-ARD. Observe that CPALS, CP-ARD and HOSVD are non-robust due to the sensitivity of L_2-norm loss function to outliers. HORPCA is shown to be robust when outliers are much larger than true data, while it performs poorly when the magnitude of outliers is within the range of true data. It should be noted that BRTF significantly outperforms competing methods under all conditions in terms of recovering the low-rank tensor and latent factors, indicating that it is more robust to outliers and

TABLE II
THE RUNTIME (SECONDS) OF SIMULATIONS IN FIG. 4(B)

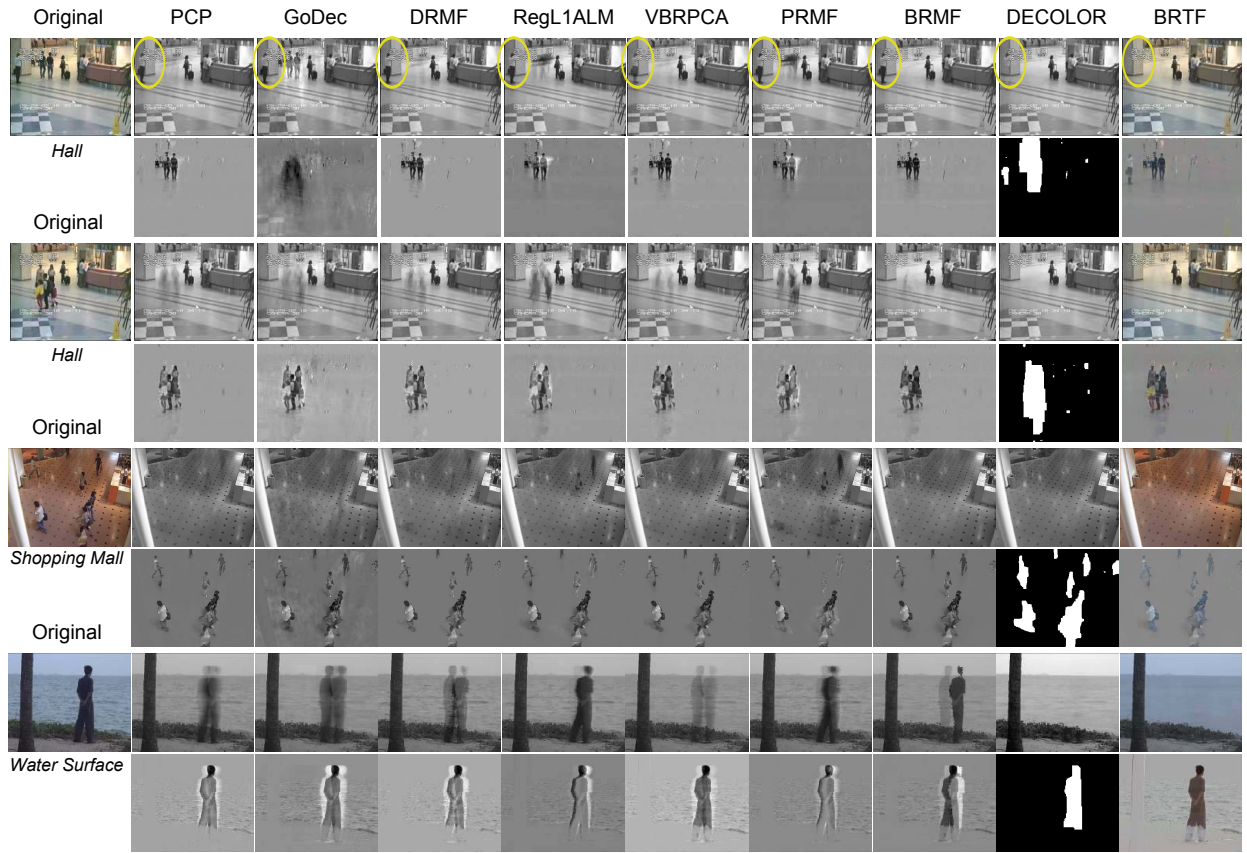| BRTF	HORPCA	CPWOPT	CPNLS	MGPCP
2 ± 0.4	26 ± 13	26 ± 8	91 ± 10	68 ± 34

Fig. 5. Results of background modeling. Four frames from three video sequences are shown from top to bottom. For each frame, there are two rows corresponding to background and foreground. From left to right, the results obtained by nine methods are shown in an order of the published years. The inferred CP ranks from three video sequences by BRTF are 2, 3, 3 respectively.
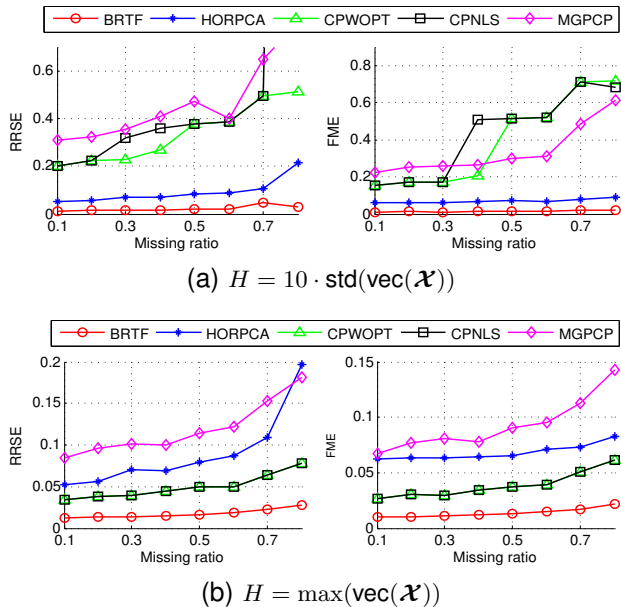

(a) $H = 10 \cdot \mathrm{std}(\mathrm{vec}(\boldsymbol{\mathcal{X}}))$

(b) $H = \max(\mathrm{vec}(\boldsymbol{\mathcal{X}}))$

Fig. 4. Predictive performance on true tensor $\boldsymbol{\mathcal{X}}$ and latent factors $\{\mathbf{A}^{(n)}\}_{n=1}^{N}$ against fraction of missing entries. The percentage of outliers is fixed to 10%, and two different magnitudes are considered in (a), (b).

non-Gaussian noises. In addition, the performance of BRTF is unaffected by the percentage and magnitude of outliers, which

confirms the capability of automatically adapting the model to various types of outliers.

A special case when the true *CP Rank* is larger than data dimensions, i.e., $R > \max_n I_n$, was investigated under the condition of 1% outliers and the detailed results are shown in Table I. The initial rank was set to 100 for both BRTF and CP-ARD. We observe that BRTF can correctly estimate rank R while CP-ARD overestimates it due to the corruptions by outliers. For CP-ALS and HOSVD, the optimal rank was selected from all possible values within $[1, R]$ by multiple runs. The sensitivity of tuning parameters is also reported by standard deviation of RRSE under varying selections. Although ground-truth data was used to tune parameters by other methods, BRTF still significantly outperforms all other methods in terms of RRSE and FME, while the runtime also shows its high efficiency. It should be noted that this experiment shows an essentially different property of tensor in contrast to matrix where $R \le \min_n I_n$ is always satisfied. Therefore, the straightforward extension of many matrix based techniques is not applicable to this situation, which has been demonstrated by the low performance of HORPCA that employs the robust matrix technique to each mode-n matricization of the tensor alternately.

The results for partially observed tensor are shown in Fig. 4. Several tensor factorization based completion methods are compared under varying missing ratios. For computation

efficiency, the rank R was initialized as 10 in BRTF and MGPCP. Observe that HORPCA is more robust than MGPCP, CWOPT and CPNLS, which cannot handle outliers explicitly, only when outliers are much larger than true data. By contrast, BRTF achieves the best performance among competing methods and its performances are quite stable under varying missing ratios, which demonstrates its robustness to both outliers and missing data. In addition, the results confirm that BRTF can accurately estimate the ground-truth of tensor rank in all cases. It should be emphasized that all competing methods had multiple runs to tune the parameters and the best possible predictive performance on missing data was reported, while BRTF and MGPCP only needs to run once. These results demonstrate that the superiorities of BRTF is not only in automatic model selection, but also in the accuracy of tensor factorization and completion. In addition, Table II shows the mean and standard deviation of runtime under different missing ratios, indicating the high efficiency of BRTF.

B. Video Background Modeling

Anomaly detection is another important ability of BRTF which can model the local information explicitly. Hence, we now consider a real-world application in surveillance video sequences with aim to separate the foreground objects from the background. Since the background is highly correlated along the frames, thus can be modeled by a low-rank tensor, while the foreground objects are moving along frames, thus can be modeled by a sparse tensor. We conducted experiments on the popular video sequences[2] by extracting 100 frames from *Shopping mall* sequence with frame size of 256×320, 300 frames from *Hall* sequence with frame size of 144×176 and 300 frames from *WaterSurface* sequence with frame size of 128×160. The state-of-the-art methods for robust matrix/tensor factorization and background modeling were employed for comparisons. The matrix-based methods, including PCP [26], GoDec [34], DRMF [30], RegL1ALM [31], VBRPCA [43], PRMF [41], BRMF [42], and DECOLOR [46], were performed on grayscale videos represented by matrices (e.g., 81920×100 for *Shopping mall*) while HORPCA [38] and BRTF were performed on original color videos represented by tensors with an additional RGB mode (e.g., $81920 \times 3 \times 100$ for *Shopping mall*). For competing methods, if necessary, the tuning parameters were selected close to optimal by visual quality due to the lack of ground-truth. The initialized rank in BRTF is set to 10.

As shown in Fig. 5, BRTF successfully separates the background and foreground on all the sequences. From one frame of *Hall*, we observe that BRTF can completely separate the person, who stands for a while and then moves away, from the background, while all other methods capture this person by both background and foreground. In another frame of *Hall*, the ghosting effects are observed by all methods except DECOLOR and BRTF. For the sequence of *Shopping Mall*, VBRPCA, BRMF, DECOLOR and BRTF obtain clearer background than those by other methods. For the sequence of *Water Surface*, DECOLOR and BRTF are clearly superior

[2]http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html
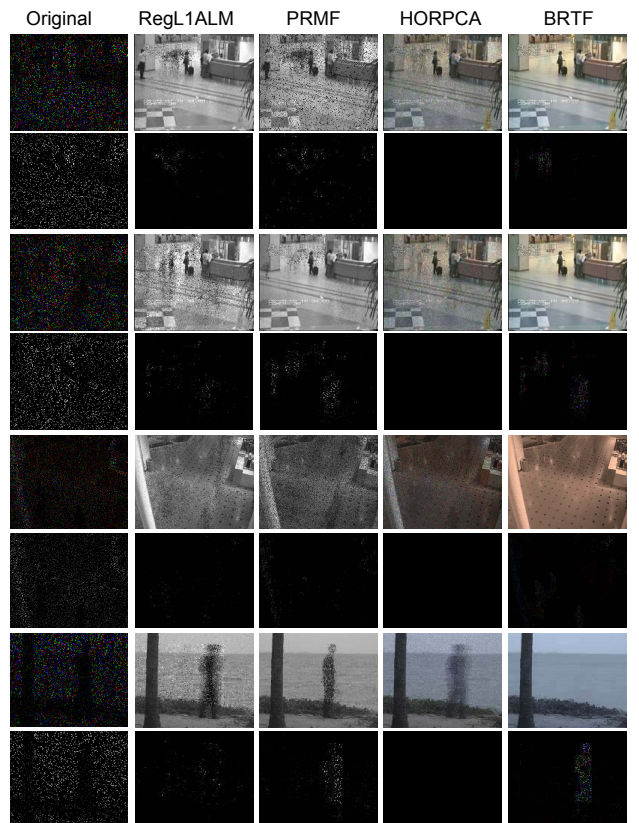

Fig. 6. Background modeling when 90% pixels are missing. Four frames from several sequences are shown and each frame has two rows corresponding to background and foreground. The original frame in color and grayscale forms are shown in the first column, followed by results obtained by different methods. The inferred CP ranks from three video sequences by BRTF are 2, 1, 1 respectively.

to other methods. Note that DECOLOR obtains comparable results with BRTF, since it incorporates the auxiliary information that is specially designed for this application. By contrast, BRTF, as a general tool for robust factorization, has shown the superiorities not only in handling tensor data, but also in more robustness than existing robust matrix factorizations. An intuitive interpretation is that BRTF can be considered as robust matrix factorizations on R, G, B matrices simultaneously with a constraint of common factors, which is more effective to capture the low-rank structure than applying matrix factorization independently.

To illustrate the property of simultaneous robust completion and anomaly detection, we conducted additional experiments on the same sequences by randomly dropping 90% pixels and compared BRTF with RegL1ALM, PRMF and HORPCA, which can handle both missing data and outliers. As illustrated in Fig. 6, BRTF is significantly superior to matrix based RegL1ALM and PRMF and tensor based HORPCA in terms of recovering the background. We observe that in the presence of missing pixels, the ghost effects are more severe by other methods, while BRTF is unaffected, indicating its better robustness to missing data and outliers. Although HORPCA is also a robust tensor factorization method, it shows comparable results with matrix based methods and cannot recover well the color of background. It should be noted that all competing

methods require tuning parameters whose selection is quite time consuming, while BRTF works in a fully automatic fashion.

C. Facial Image Denoising

In this section, we further investigate the model property in terms of robustness to non-Gaussion noises. We use the CMU-PIE face database [47] for multilinear model analysis and evaluation. All the facial images are first aligned by their eye position and then cropped to size 32×32. Next, we use 1500 facial images selected from the first 30 subjects with 5 poses and 10 illumination changes to construct a forth-order tensor $\mathcal{X} \in \mathbb{R}^{1024 \times 30 \times 5 \times 10}$. Two common types of noise that arises in the images are salt-and-pepper noise (impulse noise) and poisson noise (shot photon noise), which are both non-Gaussian. We consider noise removal for the facial images corrupted by both poisson and salt-and-pepper noise with ratio of 10% (see Fig. 7) and compare BRTF with tensor based methods including HORPCA, CP-ALS, Tucker-ARD [19], and HOSVD. Note that the tuning parameters of HORPCA were selected carefully by using the ground-truth of images while BRTF and Tucker-ARD can automatically learn the model parameters without requiring fine-tuning. The rank is initialized as 300 in BRTF, while *CP* rank was finally inferred as 297. CP-ALS was performed using the *CP* rank obtained from BRTF and HOSVD was performed using the multilinear rank obtained from Tucker-ARD.

Fig. 7 shows the qualitative results of some example images including 2 people under 5 poses and 2 lighting variations. We can observe that BRTF and HORPCA are robust to non-Gaussian noises and obtain the satisfactory visual quality, while CP-ALS, HOSVD and Tucker-ARD cannot perform well for non-Gaussian noise removal. The detailed quantitative evaluation results are shown in Table III which contains recovery performance evaluated by RRSE and PSNR as well as computational efficiency measured by runtime. BRTF and HORPCA outperform the other methods significantly because of robust property of their model assumptions. It should be noted that although HORPCA tuned parameters by the ground-truth, resulting in the best possible performance, BRTF still outperforms HORPCA in terms of recovery performance. Due to multiple runs for tuning parameter selections, the computation of HORPCA is much more expensive than BRTF. Tucker-ARD is another method that can achieve automatic model selection, however, it cannot handle non-Gaussian noise and its computational efficiency is quite low. These results demonstrate the superiority of BRTF in terms of denoising performance, robustness to non-Gaussian noise and computational efficiency. In summary, BRTF is a robust method which is similar to HORPCA; BRTF can achieve automatic rank determination which is similar to Tucker-ARD; BRTF is based on CP factorization which is similar to CP-ALS. The most significant characteristic of BRTF is the hierarchical probabilistic tensor model and full posterior inference of all unknown variables.

TABLE III
PERFORMANCE COMPARISONS ON FACIAL IMAGE DENOISING. R DENOTES THE INFERRED RANK.

	BRTF	HORPCA	CP-ALS	Tucker-ARD	HOSVD
RRSE	0.0892	0.1040	0.2551	0.2223	0.2109
PSNR	29.38	28.05	20.25	21.45	21.91
Runtime	234 s	1089 s	76 s	4528 s	0.7 s
R	297	N/A	N/A	[9, 29, 4, 9]	N/A

VII. CONCLUSION

In this paper, we have proposed a fully Bayesian generative model for robust tensor factorization, which can naturally handle missing data and outliers, together with the corresponding algorithm for efficient model inference. Our method has several significant characteristics: 1) a general framework for an arbitrary order tensor; 2) robustness to outliers, non-Gaussian noises, and overfitting; 3) tuning parameters free due to an automatic rank determination and automatic parameter selection; 4) the closed-form posterior update for efficient and deterministic inference. Comprehensive experiments and comparisons on synthetic and real-world datasets have confirmed the superiorities of BRTF over state-of-the-art robust methods and tensor factorization methods. Therefore, BRTF has proven to be promising for robust tensor factorization, robust tensor completion and outlier detection.

The Appendix, Matlab codes and demonstration videos are provided in supplementary materials.

ACKNOWLEDGMENTS

REFERENCES

[1] T. Kolda and B. Bader, "Tensor Decompositions and Applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.
[2] A. Cichocki, R. Zdunek, A. H. Phan, and S. I. Amari, *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons, 2009.
[3] L. Xiong, X. Chen, T.-K. Huang, J. G. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with bayesian probabilistic tensor factorization," in *SDM*, vol. 10. SIAM, 2010, pp. 211–222.
[4] S. Gao, L. Denoyer, P. Gallinari, and J. GUO, "Probabilistic latent tensor factorization model for link pattern prediction in multi-relational networks," *The Journal of China Universities of Posts and Telecommunications*, vol. 19, pp. 172–181, 2012.
[5] Q. Zhao, C. F. Caiafa, D. P. Mandic, Z. C. Chao, Y. Nagasaka, N. Fujii, L. Zhang, and A. Cichocki, "Higher order partial least squares (HOPLS): A generalized multilinear regression method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1660–1673, Jul. 2013.
[6] Y.-L. Chen, C.-T. Hsu, and H.-Y. Liao, "Simultaneous tensor decomposition and completion using factor priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 577–591, Mar. 2014.
[7] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.
[8] M. Sørensen, L. D. Lathauwer, P. Comon, S. Icart, and L. Deneire, "Canonical polyadic decomposition with a columnwise orthonormal factor matrix," *SIAM Journal on Matrix Analysis and Applications*, vol. 33, no. 4, pp. 1190–1213, 2012.

(a) Corrupted images   (b) BRTF   (c) HORPCA

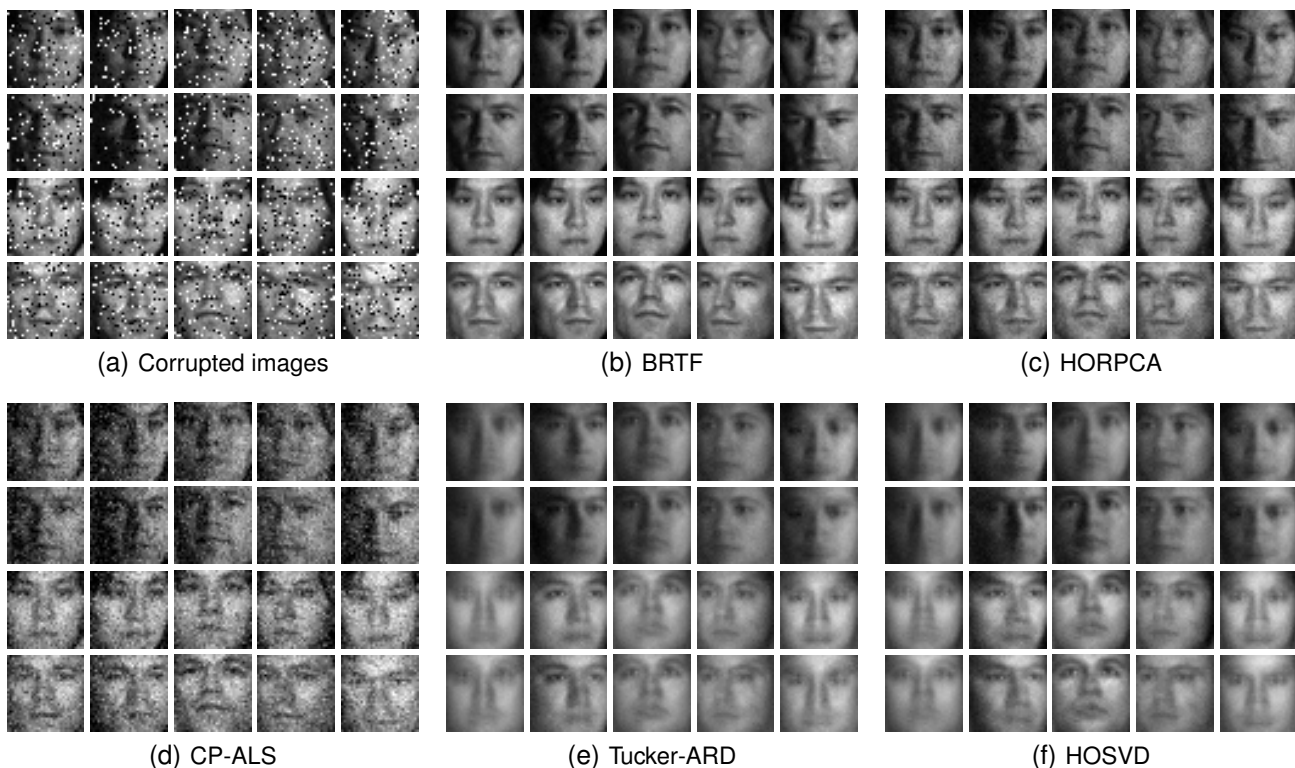(d) CP-ALS   (e) Tucker-ARD   (f) HOSVD

Fig. 7. (a) The examples of facial images corrupted by poisson and salt-and-pepper noises. (b)-(f) The reconstructed images obtained by five tensor factorization and completion methods.

[9] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 41–56, 2011.

[10] L. Sorber, M. Van Barel, and L. De Lathauwer, "Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-$(L_r, L_r, 1)$ terms, and a new generalization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 695–720, 2013.

[11] W. Chu and Z. Ghahramani, "Probabilistic models for incomplete multi-dimensional arrays," in *AISTATS Twelfth International Conference on Artificial Intelligence and Statistics*, Florida, USA, Apr. 2009, pp. 89–96.

[12] K. Hayashi, T. Takenouchi, T. Shibata, Y. Kamiya, D. Kato, K. Kunieda, K. Yamada, and K. Ikeda, "Exponential family tensor factorization for missing-values prediction and anomaly detection," in *IEEE 10th International Conference on Data Mining (ICDM)*, Sydney, NSW, Dec. 2010, pp. 216–225.

[13] Z. Xu, F. Yan, and A. Qi, "Infinite Tucker decomposition: Nonparametric Bayesian models for multiway data analysis," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, Scotland, UK, Jun. 2012, pp. 1023–1030.

[14] E. S. Allman, P. D. Jarvis, J. A. Rhodes, and J. G. Sumner, "Tensor rank, invariants, inequalities, and applications," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 1014–1045, 2013.

[15] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *Journal of the ACM (JACM)*, vol. 60, no. 6, pp. 45:1–45:39, Nov. 2013.

[16] V. De Silva and L.-H. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1084–1127, 2008.

[17] B. Alexeev, M. A. Forbes, and J. Tsimerman, "Tensor rank: Some lower and upper bounds," in *2011 IEEE 26th Annual Conference on Computational Complexity*, San Jose, CA, Jun. 2011, pp. 283–291.

[18] P. Bürgisser and C. Ikenmeyer, "Geometric complexity theory and tensor rank," in *Proceedings of the 43rd annual ACM symposium on Theory of computing*, San Jose, California, USA, Jun. 2011, pp. 509–518.

[19] M. Mørup and L. K. Hansen, "Automatic relevance determination for multi-way models," *Journal of Chemometrics*, vol. 23, no. 7-8, pp. 352–363, 2009.

[20] P. Rai, Y. Wang, S. Guo, G. Chen, D. Dunson, and L. Carin, "Scalable bayesian low-rank decomposition of incomplete multiway tensors," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, Beijing, CHINA, Jun. 2014, pp. 1800–1808.

[21] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jun. 2013.

[22] H. Tan, B. Cheng, W. Wang, Y.-J. Zhang, and B. Ran, "Tensor completion via a multi-linear low-n-rank factorization model," *Neurocomputing*, vol. 133, pp. 161–169, 2014.

[23] M. Signoretto, Q. T. Dinh, L. De Lathauwer, and J. A. Suykens, "Learning with tensors: a framework based on convex optimization and spectral regularization," *Machine Learning*, vol. 94, no. 3, pp. 303–351, 2014.

[24] J. Huang, S. Zhang, H. Li, and D. Metaxas, "Composite splitting algorithms for convex optimization," *Computer Vision and Image Understanding*, vol. 115, no. 12, pp. 1610–1622, 2011.

[25] A. Narita, K. Hayashi, R. Tomioka, and H. Kashima, "Tensor factorization using auxiliary information," *Data Mining and Knowledge Discovery*, vol. 25, no. 2, pp. 298–324, 2012.

[26] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[27] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit." *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3047–3064, 2012.

[28] J. Feng, H. Xu, and S. Yan, "Online robust PCA via stochastic optimization," in *Advances in Neural Information Processing Systems*, Harrahs and Harveys, Lake Tahoe, USA, Dec. 2013, pp. 404–412.

[29] F. Nie, H. Huang, C. Ding, D. Luo, and H. Wang, "Robust principal component analysis with non-greedy L1-norm maximization," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, Vancouver, CANADA, Jul. 2011, pp. 1433–1438.

[30] L. Xiong, X. Chen, and J. Schneider, "Direct robust matrix factorizatoin for anomaly detection," in *IEEE 11th International Conference on Data Mining (ICDM)*, Vancouver, BC, CANADA, Dec. 2011, pp. 844–853.

[31] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi, "Practical low-rank matrix approximation under robust L_1-norm," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, Jun. 2012, pp. 1410–1417.

[32] F. Nie, H. Wang, H. Huang, and C. Ding, "Joint schatten p-norm and l_p-

norm robust matrix completion for missing value recovery," *Knowledge and Information Systems*, vol. 42, no. 3, pp. 525–544, 2013.

[33] A. Eriksson and A. Van den Hengel, "Efficient computation of robust weighted low-rank matrix approximations using the l_1 norm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1681–1690, Sep. 2012.

[34] T. Zhou and D. Tao, "GoDec: randomized low-rank & sparse matrix decomposition in noisy case," in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, Bellevue, Washington, USA, Jun. 2011, pp. 33–40.

[35] H. Huang and C. Ding, "Robust tensor factorization using R_1 norm," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, Jun. 2008, pp. 1–8.

[36] M. Zhang and C. Ding, "Robust Tucker tensor decomposition for effective image representation," in *The IEEE International Conference on Computer Vision (ICCV)*, Sydney, NSW, Dec. 2013, pp. 1550–5499.

[37] J. Li, G. Han, J. Wen, and X. Gao, "Robust tensor subspace learning for anomaly detection," *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 2, pp. 89–98, 2011.

[38] D. Goldfarb and Z. Qin, "Robust low-rank tensor recovery: Models and algorithms," *SIAM Journal on Matrix Analysis and Applications*, vol. 35, no. 1, pp. 225–253, 2014.

[39] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, May 2011.

[40] J. Luttinen, A. Ilin, and J. Karhunen, "Bayesian robust PCA of incomplete data," *Neural processing letters*, vol. 36, no. 2, pp. 189–202, 2012.

[41] N. Wang, T. Yao, J. Wang, and D.-Y. Yeung, "A probabilistic approach to robust matrix factorization," in *Computer Vision–ECCV*. Florence, Italy: Springer, Oct. 2012, vol. 7578, pp. 126–139.

[42] N. Wang and D.-Y. Yeung, "Bayesian robust matrix factorization for image and video processing," in *Proceedings of International Conference on Computer Vision*, Sydney, VIC, Dec., pp. 1785–1792.

[43] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, May 2012.

[44] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, Sep. 2001.

[45] J. M. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, pp. 661–694, Dec. 2005.

[46] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[47] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.

Guoxu Zhou received the Ph.D degree in intelligent signal and information processing from South China University of Technology, Guangzhou, China, in 2010. He is currently a research scientist of the laboratory for Advanced Brain Signal Processing, at RIKEN Brain Science Institute (JAPAN). His research interests include statistical signal processing, tensor analysis, intelligent information processing, and machine learning.
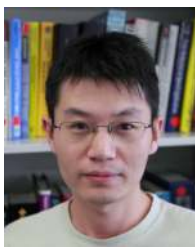

Liqing Zhang received the Ph.D. degree from Zhongshan University, Guangzhou, China, in 1988. He is now a Professor with Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests cover computational theory for cortical networks, visual perception and computational cognition, statistical learning and inference. He has published more than 210 papers in international journals and conferences.

Andrzej Cichocki received the Ph.D. and Dr.Sc. (Habilitation) degrees, all in electrical engineering, from Warsaw University of Technology (Poland). He is the senior team leader of the Laboratory for Advanced Brain Signal Processing, at RIKEN BSI (Japan). He is coauthor of more than 400 scientific papers and 4 monographs (two of them translated to Chinese). He served as AE of IEEE Trans. on Signal Processing, TNNLS, Cybernetics and J. of Neuroscience Methods.

Shun-ichi Amari received Dr. Eng. degree from the University of Tokyo in 1963. He had worked at the University of Tokyo and is now Professor-Emeritus. He served as Director of RIKEN Brain Science Institute and is now its senior advisor. Dr. Amari served as President of Institute of Electronics, Information and Communication Engineers, Japan and President of International Neural Networks Society. He received Emanuel A. Piore Award and Neural Networks Pioneer Award from IEEE, the Japan Academy Award, Order of Cultural Merit of Japan, Gabor Award from INNS, among many others.

Qibin Zhao received the Ph.D. degree from Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently a research scientist at Laboratory for Advanced Brain Signal Processing in RIKEN Brain Science Institute, Japan and is also a visiting professor in Saitama Institute of Technology, Japan. His research interests include machine learning, tensor factorization, computer vision and brain computer interface. He has published more than 50 papers in international journals and conferences.