This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

Probabilistic Tensor Canonical Polyadic Decomposition With Orthogonal Factors

Lei Cheng, Yik-Chung Wu and H. Vincent Poor

***Abstract*—Tensor canonical polyadic decomposition (CPD), which recovers the latent factor matrices from multidimensional data, is an important tool in signal processing. In many applications, some of the factor matrices are known to have orthogonality structure, and this information can be exploited to improve the accuracy of latent factors recovery. However, existing methods for CPD with orthogonal factors all require the knowledge of tensor rank, which is difficult to acquire, and have no mechanism to handle outliers in measurements. To overcome these disadvantages, in this paper, a novel tensor CPD algorithm based on the probabilistic inference framework is devised. In particular, the problem of tensor CPD with orthogonal factors is interpreted using a probabilistic model, based on which an inference algorithm is proposed that alternatively estimates the factor matrices, recovers the tensor rank and mitigates the outliers. Simulation results using synthetic data and real-world applications are presented to illustrate the excellent performance of the proposed algorithm in terms of accuracy and robustness.**

***Index Terms*—Tensor Canonical Polyadic Decomposition, Orthogonal Constraints, Robust Estimation, Multidimensional Signal Processing**

I. Introduction

Many problems in signal processing, such as independent component analysis (ICA) with matrix-based models [1]–[4], blind signal estimation in wireless communications [5]–[9], localization in array signal processing [10], [11], and linear image coding [12], [13], eventually reduce to the issue of finding a set of factor matrices $\{\mathbf{A}^{(n)} \in \mathbb{C}^{I_n \times R}\}_{n=1}^{N}$ from a complex-valued tensor $\mathcal{X} \in \mathbb{C}^{I_1 \times I_2 \times \ldots \times I_N}$ that satisfy

$$\begin{aligned}\mathcal{X} &= \sum_{r=1}^{R} \mathbf{A}_{:,r}^{(1)} \circ \mathbf{A}_{:,r}^{(2)} \circ \cdots \circ \mathbf{A}_{:,r}^{(N)} \\ &\triangleq [\![\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(N)}]\!] \end{aligned} \quad (1)$$

where $\mathbf{A}_{:,r}^{(n)} \in \mathbb{C}^{I_n \times 1}$ is the r^{th} column of the factor matrix $\mathbf{A}^{(n)}$, and $\circ$ denotes the vector outer product. This decomposition is called canonical polyadic decomposition (CPD), and the number of rank-1 component R is defined as the tensor rank [14]. Under rather mild conditions, the CPD is unique up to a trivial scalar and permutation ambiguity [14], and this fact has underlain its importance in signal processing [1]-[13].

To find the factor matrices in CPD, a common approach is to solve $\min_{\{\mathbf{A}^{(n)}\}_{n=1}^{N}} \| \mathcal{X} - [\![\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(N)}]\!] \|_F^2$. Unfortunately, it can be seen from (1) that all the factor matrices are nonlinearly coupled, and thus a closed-form solution does not exist. Consequently, the most popular solution is the alternating least squares (ALS) method, which iteratively optimizes one factor matrix at a time while holding the other factor matrices fixed [14], [15]. However, the ALS method does not take into account the potential orthogonality structure in the factor matrices, which can be found in a variety of applications. For example, the zero-mean uncorrelated signals in wireless communications [5]–[9], the prewhitening procedure in ICA [1], [4], and the basis matrices in linear image coding [12], [13], all give rise to orthogonal factors in the tensor model. Interestingly, the uniqueness of tensor CPD incorporating orthogonal factors is guaranteed under an even milder condition than the case without orthogonal factors. Pioneering work [17] formally established this fact, and extended the conventional methods to account for the orthogonality structure, among which the orthogonality constrained ALS (OALS) algorithm[1] shows remarkable efficiency in terms of accuracy and complexity.

However, there are at least two major challenges the algorithms in [17] (including the OALS) face in practical applications. Firstly, these algorithms are least-squares based, and thus lack robustness to outliers in measurements, such as ubiquitous impulsive noise in sensor arrays or networks [18], [19], and salt-and-pepper noise in images [20]. Secondly, knowledge of tensor rank is a prerequisite to implement these algorithms. Unfortunately, tensor rank acquisition from tensor data is known to be NP-hard [14]. Even though for applications in wireless communications, where the tensor rank can be assumed to be known as it is related to the number of users or sensors, existing decomposition algorithms are still susceptible to degradation caused by network dynamics, e.g., users joining and leaving the network, sudden sensor failures, etc.

In order to overcome the disadvantages presented in existing methods, we devise a novel algorithm for complex-valued tensor CPD with orthogonal factors based on the probabilistic inference framework. Probabilistic inference is well-known for providing an alternative formulation to principal component analysis (PCA) [22]. With the inception of probabilistic PCA, not only is the conventional singular

This research was supported in part by the U S. Army Research Office under MURI Grant W911NF-11-1-0036, U.S. National Science Foundation under Grant CCF-1420575 and Grant CCF-1435778.

Lei Cheng and Yik-Chung Wu are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: leicheng@eee.hku.hk, ycwu@eee.hku.hk).

H. Vincent Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (email: poor@princeton.edu).

[1]It was called the "first kind of ALS algorithm for tensor CPD with orthogonal factors (ALS1-CPO)" in [17]. For brevity of discussion, we just call it the OALS algorithm in this paper.

value decomposition (SVD) linked to statistical inference over a probabilistic model, advances in Bayesian statistics and machine learning can also be incorporated to achieve automatic relevance determination [23] and outlier removal [24]. Although the probabilistic approach is well established in matrix decomposition, extension to the tensor counterpart faces its unique challenges, since all the factor matrices are nonlinearly coupled via multiple Khatri-Rao products [14].

In this paper, we propose a probabilistic CPD algorithm for complex-valued tensor with some of the factors being orthogonal, under unknown tensor rank and in the presence of outliers in the observations. In particular, the tensor CPD problem is reformulated as an inference problem over a probabilistic model, wherein the uniform distribution over the Stiefel manifold is leveraged to encode the orthogonality structure. Since the complicatedly coupled factor matrices in the probabilistic model lead to analytically intractable integrations in exact Bayesian inference, variational inference is exploited to give an alternative solution. This results in an efficient algorithm that alternatively estimates the factor matrices, recovers the tensor rank and mitigates the outliers. Interestingly, the OALS in [17] can be interpreted as a special case of the proposed algorithm.

The remainder of this paper is organized as follows. Section II presents the motivating examples and the problem formulation. In Section III, the CPD problem is interpreted using probability density functions, and the corresponding probabilistic model is established. In Section IV, based on variational inference framework, a robust algorithm for tensor CPD with orthogonal factors is derived, and its relationship to the OALS algorithm is revealed. Simulation results using synthetic data and real-world applications are reported in Section V. Finally, conclusions are drawn in Section VI.

Notation: Boldface lowercase and uppercase letters will be used for vectors and matrices, respectively. Tensors are written as calligraphic letters. $\mathbb{E}[\,\cdot\,]$ denotes the expectation of its argument and $j \triangleq \sqrt{-1}$. Superscripts T, $*$ and H denote transpose, conjugate and Hermitian respectively. $\delta(\cdot)$ denotes the Dirac delta function. The operator $\mathrm{Tr}(\mathbf{A})$ denotes the trace of a matrix $\mathbf{A}$ and $\|\cdot\|_F$ represents the Frobenius norm of the argument. The symbol $\propto$ represents a linear scalar relationship between two real-valued functions. $\mathcal{CN}(\mathbf{x}|\mathbf{u},\mathbf{R})$ stands for the probability density function of a circularly-symmetric complex Gaussian vector $\mathbf{x}$ with mean $\mathbf{u}$ and covariance matrix $\mathbf{R}$. $\mathcal{CMN}(\mathbf{X}|\mathbf{M},\mathbf{\Sigma}_r,\mathbf{\Sigma}_c)$ denotes the complex-valued matrix normal probability density function $p(\mathbf{X}) \propto \exp\{-\mathrm{Tr}(\mathbf{\Sigma}_c^{-1}(\mathbf{X}-\mathbf{M})^H\mathbf{\Sigma}_r^{-1}(\mathbf{X}-\mathbf{M}))\}$, and $\mathcal{VMF}(\mathbf{X}|\mathbf{F})$ stands for the complex-valued von Mises-Fisher matrix probability density function $p(\mathbf{X}) \propto \exp\{-\mathrm{Tr}(\mathbf{F}\mathbf{X}^H + \mathbf{X}\mathbf{F}^H)\}$. The $N \times N$ diagonal matrix with diagonal components a_1 through a_N is represented as $\mathbf{diag}\{a_1, a_2, ..., a_N\}$, while $\mathbf{I}_M$ represents the $M \times M$ identity matrix. The $(i,j)^{th}$ element and the j^{th} column of a matrix $\mathbf{A}$ is represented by $\mathbf{A}_{i,j}$ and $\mathbf{A}_{:,j}$, respectively.

II. Motivating Examples and Problem Formulation

Tensor CPD with orthogonal factors has been widely exploited in various signal processing applications [1]-[13]. In this section, we briefly mention two motivating examples, and then we give the general problem formulation.

A. Motivating Example 1: Blind Receiver Design for DS-CDMA Systems

In a direct-sequence code division multiple access (DS-CDMA) system, the transmitted signal $s_r(k)$ from the r^{th} user at the k^{th} symbol period is multiplied by a spreading sequence $[c_{1r}, c_{2r}, \cdots, c_{Zr}]$ where c_{zr} is the z^{th} chip of the applied spreading code. Assuming R users transmit their signals simultaneously to a base station (BS) equipped with M receive antennas, the received data is given by

$$y_{mz}(k) = \sum_{r=1}^{R} h_{mr}c_{zr}s_r(k) + w_{mz}(k), \quad 1 \le m \le M, \quad 1 \le z \le Z, \tag{2}$$

where h_{mr} denotes the flat fading channel between the r^{th} user and the m^{th} receive antenna at the base station, and $w_{mz}(k)$ denotes white Gaussian noise. By introducing $\mathbf{H} \in \mathbb{C}^{M\times R}$ with its $(m,r)^{th}$ element being h_{mr}, and $\mathbf{C} \in \mathbb{C}^{Z\times R}$ with its $(z,r)^{th}$ element being c_{zr}, the model (2) can be written in matrix form as $\mathbf{Y}(k) = \sum_{r=1}^{R} \mathbf{H}_{:,r} \circ \mathbf{C}_{:,r}s_r(k) + \mathbf{W}(k)$, where $\mathbf{Y}(k), \mathbf{W}(k) \in \mathbb{C}^{M\times Z}$ are matrices with their $(m,z)^{th}$ elements being $y_{mz}(k)$ and $w_{mz}(k)$, respectively. After collecting T samples along the time dimension and defining $\mathbf{S} \in \mathbb{C}^{T\times R}$ with its $(k,r)^{th}$ element being $s_r(k)$, the system model can be further written in tensor form as [5]

$$\begin{aligned}\mathcal{Y} &= \sum_{r=1}^{R} \mathbf{H}_{:,r} \circ \mathbf{C}_{:,r} \circ \mathbf{S}_{:,r} + \mathcal{W} \\ &= [\![\mathbf{H}, \mathbf{C}, \mathbf{S}]\!] + \mathcal{W}\end{aligned} \tag{3}$$

where $\mathcal{Y} \in \mathbb{C}^{M\times Z\times T}$ and $\mathcal{W} \in \mathbb{C}^{M\times Z\times T}$ are third-order tensors, which take $y_{mz}(k)$ and $w_{mz}(k)$ as their $(m,z,k)^{th}$ elements, respectively.

It is shown in [5] that under certain mild conditions, the CPD of tensor $\mathcal{Y}$, which solves $\min_{\mathbf{H},\mathbf{C},\mathbf{S}} \| \mathcal{Y} - [\![\mathbf{H}, \mathbf{C}, \mathbf{S}]\!] \|_F^2$, can blindly recover the transmitted signals $\mathbf{S}$. Furthermore, since the transmitted signals are usually uncorrelated and with zero mean, the orthogonality structure[2] of $\mathbf{S}$ can further be taken into account to give better performance for blind signal recovery [9]. Similar models can also be found in blind data detection for cooperative communication systems [6]-[7], and in topology learning for wireless sensor networks (WSNs) [8].

B. Motivating Example 2: Linear Image Coding for a Collection of Images

Given a collection of images representing a class of objects, linear image coding extracts the commonalities of these images, which is important in image compression and recognition

[2]Strictly speaking, $\mathbf{S}$ is only approximately orthogonal. But the approximation gets better and better when observation length T increases.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

[12], [13]. The k^{th} image of size $M \times Z$ naturally corresponds to a matrix $\mathbf{B}(k)$ with its $(m,z)^{th}$ element being the image's intensity at that position. Linear image coding seeks the orthogonal basis matrices $\mathbf{U} \in \mathbb{C}^{M\times R}$ and $\mathbf{V} \in \mathbb{C}^{Z\times R}$ that capture the directions of the largest R variances in the image data, and this problem can be written as [12], [13]

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V},\{d_r(k)\}_{r=1}^R} & \sum_{k=1}^{K} \| \mathbf{B}(k) - \mathbf{U}\mathbf{diag}\{d_1(k),...,d_R(k)\}\mathbf{V}^T \|_F^2 \\ \text{s.t.} \quad & \mathbf{U}^H\mathbf{U} = \mathbf{I}_R, \mathbf{V}^H\mathbf{V} = \mathbf{I}_R. \end{aligned} \tag{4}$$

Obviously, if there is only one image (i.e., $K = 1$), problem (4) is equivalent to the well-studied SVD problem. Notice that the expression inside the Frobenius norm in (4) can be written as $\mathbf{B}(k) - \sum_{r=1}^{R} \mathbf{U}_{:,r} \circ \mathbf{V}_{:,r} d_r(k)$. Further introducing the matrix $\mathbf{D}$ with its $(k,r)^{th}$ element being $d_r(k)$, it is easy to see that problem (4) can be rewritten in tensor form as

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V},\mathbf{D}} \| \mathcal{B} - \underbrace{\sum_{r=1}^{R} \mathbf{U}_{:,r} \circ \mathbf{V}_{:,r} \circ \mathbf{D}_{:,r}}_{=[\![\mathbf{U},\mathbf{V},\mathbf{D}]\!]} \|_F^2 \\ \text{s.t.} \quad \mathbf{U}^H\mathbf{U} = \mathbf{I}_R, \mathbf{V}^H\mathbf{V} = \mathbf{I}_R, \end{aligned} \tag{5}$$

where $\mathcal{B} \in \mathbb{C}^{M\times Z\times K}$ is a third-order tensor with $\mathbf{B}(k)$ as its k^{th} slice. Therefore, linear image coding for a collection of images is equivalent to solving a tensor CPD with two orthonormal factor matrices.

C. Problem Formulation

From the two motivating examples above, we take a step further and consider a generalized problem in which the observed data tensor $\mathcal{Y} \in \mathbb{C}^{I_1\times I_2\times ...\times I_N}$ obeys the following model:

$$\mathcal{Y} = [\![\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(N)}]\!] + \mathcal{W} + \mathcal{E} \tag{6}$$

where $\mathcal{W}$ represents an additive noise tensor with each element $w_{i_1,i_2,...,i_N} \sim \mathcal{CN}\left(w_{i_1,i_2,...,i_N}|0,\beta^{-1}\right)$ and with correlation $\mathbb{E}(w^*_{i_1,i_2,...,i_N} w_{\tau_1,\tau_2,...,\tau_N}) = \beta^{-1}\prod_{n=1}^{N}\delta(\tau_n - i_n)$; $\mathcal{E}$ denotes potential outliers in measurements with each element $e_{i_1,i_2,...,i_N}$ taking an unknown value if an outlier emerges, and otherwise taking the value zero. Since the number of orthogonal factor matrices could be known a priori in specific application, it is assumed that $\{\mathbf{A}^{(n)}\}_{n=1}^{P}$ are known to be orthogonal where $P < N$, while the remaining factor matrices are unconstrained.

Due to the orthogonality structure of the first P factor matrices $\{\mathbf{A}^{(n)}\}_{n=1}^{P}$, they can be written as $\mathbf{A}^{(n)} = \mathbf{U}^{(n)}\mathbf{\Lambda}^{(n)}$ where $\mathbf{U}^{(n)}$ is an orthonormal matrix and $\mathbf{\Lambda}^{(n)}$ is a diagonal matrix. Putting $\mathbf{A}^{(n)} = \mathbf{U}^{(n)}\mathbf{\Lambda}^{(n)}$ for $1 \le n \le P$ into the definition of the tensor CPD in (1), it is easy to show that

$$[\![\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(N)}]\!] = [\![\mathbf{\Xi}^{(1)}, \mathbf{\Xi}^{(2)}, \cdots, \mathbf{\Xi}^{(N)}]\!] \tag{7}$$

with $\mathbf{\Xi}^{(n)} = \mathbf{U}^{(n)}\mathbf{\Pi}$ for $1 \le n \le P$, $\mathbf{\Xi}^{(n)} = \mathbf{A}^{(n)}\mathbf{\Pi}$ for $P+1 \le n \le N-1$, and $\mathbf{\Xi}^{(N)} = \mathbf{A}^{(N)}\mathbf{\Lambda}^{(1)}\mathbf{\Lambda}^{(2)}\cdots\mathbf{\Lambda}^{(P)}\mathbf{\Pi}$, where $\mathbf{\Pi} \in \mathbb{C}^{R\times R}$ is a permutation matrix. From (7), it can be seen that up to the scaling and permutation indeterminacy, the tensor CPD under orthogonal constraints is equivalent to that under orthonormal constraints. In general, the scaling and permutation ambiguity can be easily resolved using side information [5]. On the other hand, for those applications that seek the subspaces spanned by the factor matrices, such as linear image coding described in Section II.B, the scaling and permutation ambiguity can be ignored. Thus, without loss of generality, our goal is to estimate an N-tuplet of factor matrices $(\mathbf{\Xi}^{(1)}, \mathbf{\Xi}^{(2)}, ..., \mathbf{\Xi}^{(N)})$ with the first P (where $P < N$) of them being orthonormal, based on the observation $\mathcal{Y}$ and in the absence of the knowledge of noise power β^{-1}, outlier statistics and the tensor rank R. In particular, since we do not know the exact value of R, it is assumed that there are L columns in each factor matrix $\mathbf{\Xi}^{(n)}$, where L is the maximum possible value of the tensor rank R. Thus, the problem to be solved can be stated as

$$\begin{aligned}\min_{\{\mathbf{\Xi}^{(n)}\}_{n=1}^{N},\mathcal{E}} & \beta \| \mathcal{Y} - [\![\mathbf{\Xi}^{(1)}, \mathbf{\Xi}^{(2)}, \cdots, \mathbf{\Xi}^{(N)}]\!] - \mathcal{E} \|_F^2 \\ & + \sum_{l=1}^{L} \gamma_l \Big(\sum_{n=1}^{N} \mathbf{\Xi}_{:,l}^{(n)H} \mathbf{\Xi}_{:,l}^{(n)} \Big) \\ \text{s.t.} \quad & \mathbf{\Xi}^{(n)H}\mathbf{\Xi}^{(n)} = \mathbf{I}_L, \quad n = 1, 2, \cdots, P, \end{aligned} \tag{8}$$

where the regularization term $\sum_{l=1}^{L} \gamma_l \left(\sum_{n=1}^{N} \mathbf{\Xi}_{:,l}^{(n)H} \mathbf{\Xi}_{:,l}^{(n)} \right)$ is added to control the complexity of the model and avoid overfitting of noise [21], since more columns (thus more degrees of freedom) in $\mathbf{\Xi}^{(n)}$ than the true model are introduced, and $\{\gamma_l\}_{l=1}^{L}$ are regularization parameters trading off the relative importance of the square error term and the regularization term.

Existing algorithms [17] for tensor CPD with orthonormal factors cannot be used to solve problem (8), since they have no mechanism to handle outliers $\mathcal{E}$. Furthermore, the choice of regularization parameters plays an important role, since setting γ_l too large results in excessive residual squared error, while setting γ_l too small risks overfitting of noise. In general, determining the optimal regularization parameters (e.g., using cross-validation [27], or the L-curve [28]) requires exhaustive search, and thus is computationally demanding. To overcome these problems, we propose a novel algorithm based on the framework of probabilistic inference, which effectively mitigates the outliers $\mathcal{E}$ and automatically learns the regularization parameters.

III. Probabilistic Model for Tensor CPD with Orthogonal Factors

Before solving problem (8), we interpret different terms in (8) as probability density functions, based on which a probabilistic model that encodes our knowledge of the observation and the unknowns can be established.

Firstly, since the elements of the additive noise $\mathcal{W}$ is white, zero-mean and circularly-symmetric complex Gaussian, the squared error term in problem (8) can be interpreted as the negative log of the likelihood given by [21]:

$$\begin{aligned} & p\left(\mathcal{Y} \mid \mathbf{\Xi}^{(1)}, \mathbf{\Xi}^{(2)}, ..., \mathbf{\Xi}^{(N)}, \mathcal{E}, \beta\right) \\ & \propto \exp\left(-\beta \| \mathcal{Y} - [\![\mathbf{\Xi}^{(1)}, \mathbf{\Xi}^{(2)}, ..., \mathbf{\Xi}^{(N)}]\!] - \mathcal{E} \|_F^2\right). \end{aligned} \tag{9}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

Secondly, the regularization term in problem (8) can be interpreted as arising from a circularly-symmetric complex Gaussian prior distribution over the columns of the factor matrices, i.e., $\prod_{n=1}^{N}\prod_{l=1}^{L}\mathcal{CN}\Big(\mathbf{\Xi}_{:,l}^{(n)}|\ \mathbf{0}_{I_n\times 1},\gamma_l^{-1}\mathbf{I}_L\Big)$ [21]. Note that the columns of the factor matrices are independent of each other, and the l^{th} columns in all factor matrices $\{\mathbf{\Xi}^{(n)}\}_{n=1}^{N}$ share the same variance γ_l^{-1}. This has the physical interpretation that if γ_l is large, the l^{th} columns in all $\mathbf{\Xi}^{(n)}$'s will be effectively "switched off". On the other hand, for the first P factor matrices $\{\mathbf{\Xi}^{(n)}\}_{n=1}^{P}$, there are additional hard constraints in problem (8), which correspond to the Stiefel manifold [33] $V_L(\mathbb{C}^{I_n})=\{\mathbf{A}\in\mathbb{C}^{I_n\times L}:\mathbf{A}^H\mathbf{A}=\mathbf{I}_L\}$ for $1\le n\le P$. Since the orthonormal constraints result in $\mathbf{\Xi}_{:,l}^{(n)H}\mathbf{\Xi}_{:,l}=1$, the hard constraints would dominate the Gaussian distribution of the columns in $\{\mathbf{\Xi}^{(n)}\}_{n=1}^{P}$. Therefore, $\mathbf{\Xi}^{(n)}$ can be interpreted as being uniformly distributed over the Stiefel manifold $V_L(\mathbb{C}^{I_n})$ for $1\le n\le P$, and Gaussian distributed for $P+1\le n\le N$:

$$\begin{aligned}&p(\mathbf{\Xi}^{(1)},\mathbf{\Xi}^{(2)},\cdots,\mathbf{\Xi}^{(P)})\propto\prod_{n=1}^{P}\mathbb{I}_{V_L(\mathbb{C}^{I_n})}(\mathbf{\Xi}^{(n)}),\\&p(\mathbf{\Xi}^{(P+1)},\mathbf{\Xi}^{(P+2)},\cdots,\mathbf{\Xi}^{(N)})\\&\qquad=\prod_{n=P+1}^{N}\prod_{l=1}^{L}\mathcal{CN}\Big(\mathbf{\Xi}_{:,l}^{(n)}|\mathbf{0}_{I_n\times 1},\gamma_l^{-1}\mathbf{I}_L\Big),\end{aligned}\tag{10}$$

where $\mathbb{I}_{V_L(\mathbb{C}^{I_n})}(\mathbf{\Xi}^{(n)})$ is an indicator function with $\mathbb{I}_{V_L(\mathbb{C}^{I_n})}(\mathbf{\Xi}^{(n)})=1$ when $\mathbf{\Xi}^{(n)}\in V_L(\mathbb{C}^{I_n})$, and otherwise $\mathbb{I}_{V_L(\mathbb{C}^{I_n})}(\mathbf{\Xi}^{(n)})=0$. For the parameters β and $\{\gamma_l\}_{l=1}^{L}$, which correspond to the inverse noise power and the variances of columns in the factor matrices, since we have no information about their distributions, non-informative a Jeffrey's prior [27] is imposed on them, i.e., $p(\beta)\propto\beta^{-1}$ and $p(\gamma_l)\propto\gamma_l^{-1}$ for $l=1,\cdots,L$.

Finally, although the generative model for outliers $\mathcal{E}_{i_1,\cdots,i_N}$ is unknown, the rare occurrence of outliers motivates us to employ student's t distribution as its prior [27], i.e., $p(\mathcal{E}_{i_1,\cdots,i_N})=\mathcal{T}(\mathcal{E}_{i_1,\cdots,i_N}|0,c_{i_1,\cdots,i_N},d_{i_1,\cdots,i_N})$. To facilitate the Bayesian inference procedure, student's t distribution can be equivalently represented as a Gaussian scale mixture as follows [34] :

$$\begin{aligned}&\mathcal{T}\left(\mathcal{E}_{i_1,\cdots,i_N}\mid 0,c_{i_1,\cdots,i_N},d_{i_1,\cdots,i_N}\right)\\&=\int\mathcal{CN}\left(\mathcal{E}_{i_1,\cdots,i_N}\mid 0,\zeta_{i_1,\cdots,i_N}^{-1}\right)\\&\times\mathrm{Gamma}\left(\zeta_{i_1,\cdots,i_N}\mid c_{i_1,\cdots,i_N},d_{i_1,\cdots,i_N}\right)\ \mathrm{d}\zeta_{i_1,\cdots,i_N}.\end{aligned}\tag{11}$$

This means that student's t distribution can be obtained by mixing an infinite number of zero-mean circularly-symmetric complex Gaussian distributions where the mixing distribution on the precision $\zeta_{i_1,\cdots,i_N}$ is the gamma distribution with parameters $c_{i_1,\cdots,i_N}$ and $d_{i_1,\cdots,i_N}$. In addition, since the statistics of outliers such as means and correlations are generally unavailable in practice, we set the hyper-parameters $c_{i_1,\cdots,i_N}$ and $d_{i_1,\cdots,i_N}$ as 10^{-6} to produce a non-informative prior on $\mathcal{E}_{i_1,\cdots,i_N}$, and assume outliers are independent of each other:

$$p\left(\mathcal{E}\right)=\prod_{i_1=1}^{I_1}\cdots\prod_{i_N=1}^{I_N}\mathcal{T}\big(\mathcal{E}_{i_1,\cdots,i_N}|0,c_{i_1,\cdots,i_N}=10^{-6},d_{i_1,\cdots,i_N}=10^{-6}\big)\,.\tag{12}$$

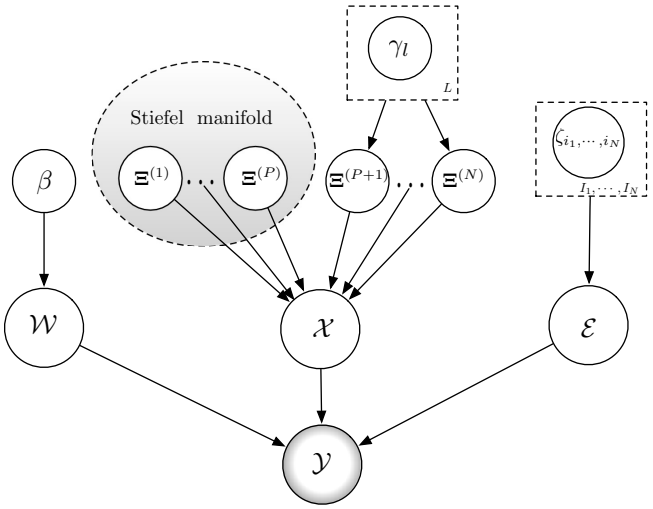


Figure 1: Probabilistic model for tensor CPD with orthogonal factors

The complete probabilistic model is shown in Figure 1. Notice that the proposed probabilistic model in this paper is different from that of existing works on tensor decompositions [30]–[32]. In particular, existing tensor probabilistic models do not take orthogonality structure into account. Furthermore, existing tenor decompositions [30]–[32], [43]–[45] are designed for real-valued tensors only, and thus cannot process the complex-valued data arising in applications such as wireless communications [5]-[9] and functional magnetic resonance imaging [3].

IV. Variational Inference for Tensor Factorization

Let $\mathbf{\Theta}$ be a set containing the factor matrices $\{\mathbf{\Xi}^{(n)}\}_{n=1}^{N}$, and other variables $\mathcal{E}$, $\{\gamma_l\}_{l=1}^{L}$, $\{\zeta_{i_1,\cdots,i_N}\}_{i_1=1,\cdots,i_N=1}^{I_1,\cdots,I_N}$, β . From the probabilistic model established above, the marginal probability density functions of the unknown factor matrices $\{\mathbf{\Xi}\}_{n=1}^{N}$ are given by

$$p(\mathbf{\Xi}^{(n)}|\mathcal{Y})=\int\frac{p(\mathcal{Y},\mathbf{\Theta})}{p(\mathcal{Y})}d\mathbf{\Theta}\backslash\mathbf{\Xi}^{(n)},\ n=1,2,\cdots,N,\tag{13}$$

where

$$\begin{aligned}&p(\mathcal{Y},\mathbf{\Theta})\\&\propto\prod_{n=1}^{P}\mathbb{I}_{V_L(\mathbb{C}^{I_n})}\Big(\mathbf{\Xi}^{(n)}\Big)\exp\left\{\left(\prod_{n=1}^{N}I_n-1\right)\ln\beta\right.\\&+\left(\sum_{n=P+1}^{N}I_n+1\right)\sum_{l=1}^{L}\ln\gamma_l-\mathrm{Tr}\left(\mathbf{\Gamma}\sum_{n=P+1}^{N}\mathbf{\Xi}^{(n)H}\mathbf{\Xi}^{(n)}\right)\\&+\left.\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\left[(c_{i_1,\cdots,i_N}-1)\ln\zeta_{i_1,\cdots,i_N}-d_{i_1,\cdots,i_N}\zeta_{i_1,\cdots,i_N}\right]\right.\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

$$+\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\left(\ln\zeta_{i_1,\cdots,i_N}-\zeta_{i_1,\cdots,i_N}\mathcal{E}^*_{i_1,\cdots,i_N}\mathcal{E}_{i_1,\cdots,i_N}\right)$$
$$-\beta\parallel\mathcal{Y}-[\![\boldsymbol{\Xi}^{(1)},\boldsymbol{\Xi}^{(2)},...,\boldsymbol{\Xi}^{(N)}]\!]-\mathcal{E}\parallel_F^2\Bigg\}\quad(14)$$

with $\boldsymbol{\Gamma}=\mathbf{diag}\{\gamma_1,\cdots,\gamma_R\}$.

Since the factor matrices and other variables are nonlinearly coupled in (14), the multiple integrations in (13) are analytically intractable, which prohibits exact Bayesian inference. To handle this problem, Monte Carlo statistical methods [25], [26], in which a large number of random samples are generated from the joint distributions and marginalization is approximated by operations on samples, can be explored. These Monte Carlo based approximations can approach the exact multiple integrations when the number of samples approaches infinity, which however is computationally demanding [27]. More recently, variational inference, in which another distribution that is close to the true posterior distribution in the Kullback-Leibler (KL) divergence sense is sought, has been exploited to give deterministic approximations to the intractable multiple integrations [29].

More specifically, in variational inference, a variational distribution with probability density function $Q(\boldsymbol{\Theta})$ that is the closest among a given set of distributions to the true posterior distribution $p(\boldsymbol{\Theta}\mid\mathcal{Y})=p(\boldsymbol{\Theta},\mathcal{Y})/p(\mathcal{Y})$ in the KL divergence sense is sought [29]:

$$\mathbf{KL}\big(Q(\boldsymbol{\Theta})\parallel p(\boldsymbol{\Theta}\mid\mathcal{Y})\big)\triangleq-\mathbb{E}_{Q(\boldsymbol{\Theta})}\left\{\ln\frac{p(\boldsymbol{\Theta}\mid\mathcal{Y})}{Q(\boldsymbol{\Theta})}\right\}.\quad(15)$$

The KL divergence vanishes when $Q(\boldsymbol{\Theta})=p(\boldsymbol{\Theta}\mid\mathcal{Y})$ if no constraint is imposed on $Q(\boldsymbol{\Theta})$, which however leads us back to the original intractable posterior distribution. A common approach is to apply the mean field approximation, which assumes that the variational probability density takes a fully factorized form $Q(\boldsymbol{\Theta})=\prod_k Q(\Theta_k)$, $\Theta_k\in\boldsymbol{\Theta}$. Furthermore, to facilitate the manipulation of hard constraints on the first P factor matrices, their variational densities are assumed to take a Dirac delta functional form $Q(\boldsymbol{\Xi}^{(k)})=\delta(\boldsymbol{\Xi}^{(k)}-\hat{\boldsymbol{\Xi}}^{(k)})$ for $k=1,2,\cdots,P$, where $\hat{\boldsymbol{\Xi}}^{(k)}$ is a parameter to be derived.

Under these approximations, the probability density functions $Q(\Theta_k)$ of the variational distribution can be analytically obtained via [29]

$$Q(\boldsymbol{\Xi}^{(k)})=\delta\Big(\boldsymbol{\Xi}^{(k)}-\underbrace{\arg\max_{\boldsymbol{\Xi}^{(k)}}\mathbb{E}_{\prod_{\Theta_j\neq\boldsymbol{\Xi}^{(k)}}Q(\Theta_j)}\big[\ln p(\mathcal{Y},\boldsymbol{\Theta})\big]}_{\triangleq\hat{\boldsymbol{\Xi}}^{(k)}}\Big),$$
$$k=1,2,\cdots,P,\quad(16)$$

and

$$Q(\Theta_k)\propto\exp\Big\{\mathbb{E}_{\prod_{j\neq k}Q(\Theta_j)}\big[\ln p(\mathcal{Y},\boldsymbol{\Theta})\big]\Big\},\ \Theta_k\in\boldsymbol{\Theta}\backslash\{\boldsymbol{\Xi}^{(k)}\}_{k=1}^P.\quad(17)$$

Obviously, these variational distributions are coupled in the sense that the computation of variational distribution of one parameter requires the knowledge of variational distributions of other parameters. Therefore, these variational distributions should be updated iteratively. In the following, explicit expression for each $Q(\cdot)$ is derived.

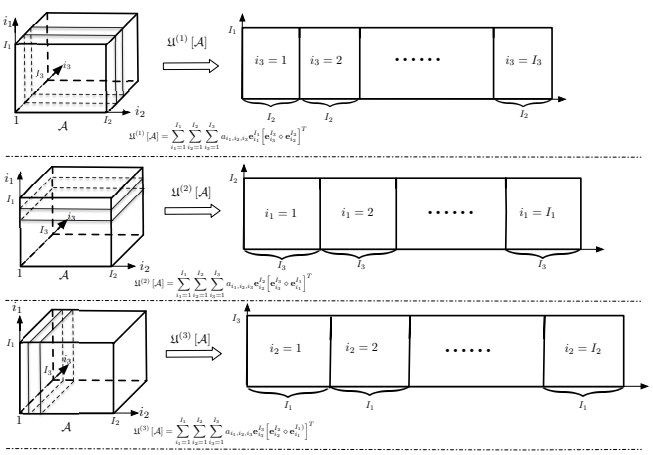


Figure 2: Unfolding operation for a third-order tensor

A. Derivation for $Q(\boldsymbol{\Xi}^{(k)}), 1\leq k\leq P$

By substituting (14) into (16) and only keeping the terms relevant to $\boldsymbol{\Xi}^{(k)}$ ($1\leq k\leq P$), we directly have

$$\hat{\boldsymbol{\Xi}}^{(k)}=\underset{\boldsymbol{\Xi}^{(k)}\in V_L(\mathbb{C}^{I_k})}{\arg\max}\ \mathbb{E}_{\prod_{\Theta_j\neq\boldsymbol{\Xi}^{(k)}}Q(\Theta_j)}\big[$$
$$-\beta\parallel\mathcal{Y}-[\![\boldsymbol{\Xi}^{(1)},\cdots,\boldsymbol{\Xi}^{(N)}]\!]-\mathcal{E}\parallel_F^2\big].\quad(18)$$

To expand the square of the Frobenius norm inside the expectation in (18), we use the result that $\parallel\mathcal{A}\parallel_F^2=\parallel\mathfrak{U}^{(k)}[\mathcal{A}]\parallel_F^2=\mathrm{Tr}\left(\mathfrak{U}^{(k)}[\mathcal{A}]\left(\mathfrak{U}^{(k)}[\mathcal{A}]\right)^H\right)$ [14], where the unfolding operation $\left\{\mathfrak{U}^{(k)}[\mathcal{A}]\right\}_{k=1,2,\cdots,N}$ on an N^{th}-order tensor $\mathcal{A}\in\mathbb{C}^{I_1\times\cdots\times I_N}$ along its k^{th} mode is specified as $\mathfrak{U}^{(k)}[\mathcal{A}]=\sum_{i_1=1}^{I_1}\cdots\sum_{i_n=1}^{I_N}a_{i_1,\cdots,i_N}\,\mathbf{e}_{i_k}^{I_k}\Big[\overset{N}{\underset{n=1,n\neq k}{\diamond}}\mathbf{e}_{i_n}^{I_n}\Big]^T$. In this expression, the elementary vector $\mathbf{e}_{i_n}^{I_n}\in\mathbb{R}^{I_n\times1}$ is all zeroes except for a 1 at the i_n^{th} location, and the multiple Khatri-Rao products $\overset{N}{\underset{n=1,n\neq k}{\diamond}}\mathbf{A}^{(n)}=\mathbf{A}^{(N)}\diamond\cdots\diamond\mathbf{A}^{(k+1)}\diamond\mathbf{A}^{(k-1)}\diamond\cdots\diamond\mathbf{A}^{(1)}$. For example, the unfolding operation for a third-order tensor is illustrated in Figure 2. After expanding the square of the Frobenius norm and taking expectations, the parameter $\hat{\boldsymbol{\Xi}}^{(k)}$ for each variational density in $\{Q(\boldsymbol{\Xi}^{(k)})\}_{k=1}^P$ can be obtained from the problem (19) at the top of the next page.

Using the fact that the feasible set for parameter $\boldsymbol{\Xi}^{(k)}$ is the Stiefel manifold $V_L(\mathbb{C}^{I_k})$, i.e., $\boldsymbol{\Xi}^{(k)H}\boldsymbol{\Xi}^{(k)}=\mathbf{I}_L$, the term $\mathbf{G}^{(k)}$ is irrelevant to the factor matrix of interest $\boldsymbol{\Xi}^{(k)}$. Consequently, problem (19) is equivalent to

$$\hat{\boldsymbol{\Xi}}^{(k)}=\underset{\boldsymbol{\Xi}^{(k)}\in V_L(\mathbb{C}^{I_k})}{\arg\max}\ \mathrm{Tr}\left(\mathbf{F}^{(k)}\boldsymbol{\Xi}^{(k)H}+\boldsymbol{\Xi}^{(k)}\mathbf{F}^{(k)H}\right),\quad(20)$$

where $\mathbf{F}^{(k)}$ was defined in the first line of (19). Problem (20) is a non-convex optimization problem, as its feasible set $V_L(\mathbb{C}^{I_k})$ is non-convex [37]. While in general (20) can be solved by numerical iterative algorithms based on a geometric approach or the alternating direction method of multipliers [37], a closed-form optimal solution can be obtained by noticing that the objective function in (20) has the same functional form as the log of the von Mises-Fisher matrix distribution

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

$$\hat{\mathbf{\Xi}}^{(k)} = \underset{\mathbf{\Xi}^{(k)} \in V_L(\mathbb{C}^{I_k})}{\arg\max}\ \mathrm{Tr}\Bigg(\underbrace{\mathbb{E}_{Q(\beta)}[\beta]\mathfrak{U}^{(k)}\Big[\mathcal{Y} - \mathbb{E}_{Q(\mathcal{E})}\left[\mathcal{E}\right]\Big]\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbb{E}_{Q(\mathbf{\Xi}^{(n)})}\left[\mathbf{\Xi}^{(n)}\right]\Big)^*}_{\triangleq \mathbf{F}^{(k)}} \mathbf{\Xi}^{(k)H} + \mathbf{\Xi}^{(k)}\mathbf{F}^{(k)H}\Bigg)$$
$$- \mathrm{Tr}\Bigg(\underbrace{\mathbf{\Xi}^{(k)H}\mathbf{\Xi}^{(k)}\Big[\mathbb{E}_{Q(\beta)}\left[\beta\right]\mathbb{E}_{\prod_{n=1,n\neq k}^{N} Q(\mathbf{\Xi}^{(n)})}\Big[\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbf{\Xi}^{(n)}\Big)^T\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbf{\Xi}^{(n)}\Big)^*\Big] + \mathbb{E}_{\prod_{l=1}^{L} Q(\gamma_l)}\left[\mathbf{\Gamma}\right]\Big]}_{\triangleq \mathbf{G}^{(k)}}\Bigg) \tag{19}$$

with parameter matrix $\mathbf{F}^{(k)}$, and the feasible set in (20) also coincides with the support of this von Mises-Fisher matrix distribution [35]. As a result, we have

$$\hat{\mathbf{\Xi}}^{(k)} = \underset{\mathbf{\Xi}^{(k)}}{\arg\max} \ln \mathcal{VMF}\left(\mathbf{\Xi}^{(k)} \mid \mathbf{F}^{(k)}\right). \tag{21}$$

Then, the closed-form solution for problem (21) can be acquired using **Property 1** below, which has been proved in [33].

Property 1. *Suppose the matrix $\mathbf{A} \in \mathbb{C}^{\kappa_1 \times \kappa_2}$ follows a von Mises-Fisher matrix distribution with parameter matrix $\mathbf{F} \in \mathbb{C}^{\kappa_1 \times \kappa_2}$. If $\mathbf{F} = \mathbf{U}\mathbf{\Phi}\mathbf{V}^H$ is the SVD of the matrix $\mathbf{F}$, then the unique mode of $\mathcal{VMF}\left(\mathbf{A} \mid \mathbf{F}\right)$ is $\mathbf{U}\mathbf{V}^H$.*

From **Property 1**, it is easy to conclude that $\hat{\mathbf{\Xi}}^{(k)} = \mathbf{\Upsilon}^{(k)}\mathbf{\Pi}^{(k)H}$, where $\mathbf{\Upsilon}^{(k)}$ and $\mathbf{\Pi}^{(k)}$ are the left-orthonormal matrix and right-orthonormal matrix from the SVD of $\mathbf{F}^{(k)}$, respectively.

B. Derivation for $Q\left(\mathbf{\Xi}^{(k)}\right), P+1 \leq k \leq N$

Using (14) and (17), the variational density $Q\left(\mathbf{\Xi}^{(k)}\right)$ $(P+1 \leq k \leq N)$ is derived in Appendix A to be a circularly-symmetric complex matrix normal distribution [35] as

$$Q\left(\mathbf{\Xi}^{(k)}\right) = \mathcal{CMN}(\mathbf{\Xi}^{(k)} \mid \mathbf{M}^{(k)}, \mathbf{I}_{I_k}, \mathbf{\Sigma}^{(k)}) \tag{22}$$

where

$$\mathbf{\Sigma}^{(k)} = \Bigg(\mathbb{E}_{Q(\beta)}\left[\beta\right]\mathbb{E}_{\prod_{n=1,n\neq k}^{N} Q(\mathbf{\Xi}^{(n)})}\Big[\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbf{\Xi}^{(n)}\Big)^T\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbf{\Xi}^{(n)}\Big)^*\Big] + \mathbb{E}_{\prod_{l=1}^{L} Q(\gamma_l)}\left[\mathbf{\Gamma}\right]\Bigg)^{-1} \tag{23}$$

$$\mathbf{M}^{(k)} = \mathbb{E}_{Q(\beta)}\left[\beta\right]\mathfrak{U}^{(k)}\Big[\mathcal{Y} - \mathbb{E}_{Q(\mathcal{E})}\left[\mathcal{E}\right]\Big] \times \Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbb{E}_{Q(\mathbf{\Xi}^{(n)})}\left[\mathbf{\Xi}^{(n)}\right]\Big)^*\mathbf{\Sigma}^{(k)}. \tag{24}$$

Due to the fact that $Q(\mathbf{\Xi}^{(k)})$ is Gaussian, the parameter $\mathbf{M}^{(k)}$ is both the expectation and the mode of the variational density $Q\left(\mathbf{\Xi}^{(k)}\right)$.

To calculate $\mathbf{M}^{(k)}$, some expectation computations are required as shown in (23) and (24). For those with the form $\mathbb{E}_{Q(\Theta_k)}\left[\Theta_k\right]$ where $\Theta_k \in \mathbf{\Theta}$, the value can be easily obtained if the corresponding $Q\left(\Theta_k\right)$ is available. The remaining challenge stems from the expectation $\mathbb{E}_{\prod_{n=1,n\neq k}^{N} Q(\mathbf{\Xi}^{(n)})}\Big[\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbf{\Xi}^{(n)}\Big)^T\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbf{\Xi}^{(n)}\Big)^*\Big]$ in (23). But its calculation becomes straightforward after exploiting the orthonormal structure of $\{\hat{\mathbf{\Xi}}^{(k)}\}_{k=1}^{P}$ and the property of multiple Khatri-Rao products, as presented in the following property.

Property 2. *Suppose the matrix $\mathbf{A}^{(n)} \in \mathbb{C}^{\kappa_n \times \rho} \sim \delta(\mathbf{A}^{(n)} - \hat{\mathbf{A}}^{(n)})$ for $1 \leq n \leq P$, where $\hat{\mathbf{A}}^{(n)} \in V_\rho(\mathbb{C}^{\kappa_n})$ and $P < N$, and the matrix $\mathbf{A}^{(n)} \in \mathbb{C}^{\kappa_n \times \rho} \sim \mathcal{CMN}(\mathbf{A}^{(n)} \mid \mathbf{M}^{(n)}, \mathbf{I}_{\kappa_n}, \mathbf{\Sigma}^{(n)})$ for $P+1 \leq n \leq N$. Then,*

$$\mathbb{E}_{\prod_{n=1,n\neq k}^{N} p(\mathbf{A}^{(n)})}\Big[\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbf{A}^{(n)}\Big)^T\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}} \mathbf{A}^{(n)}\Big)^*\Big] = \mathfrak{D}\Big[\overset{N}{\underset{n=P+1,n\neq k}{\odot}} \Big(\mathbf{M}^{(n)H}\mathbf{M}^{(n)} + \kappa_n\mathbf{\Sigma}^{(n)}\Big)^*\Big] \tag{25}$$

where $\mathfrak{D}[\mathbf{A}]$ is a diagonal matrix taking the diagonal element from $\mathbf{A}$, and the multiple Hadamard products $\overset{N}{\underset{n=1,n\neq k}{\odot}} \mathbf{A}^{(n)} = \mathbf{A}^{(N)} \odot \cdots \odot \mathbf{A}^{(k+1)} \odot \mathbf{A}^{(k-1)} \odot \cdots \odot \mathbf{A}^{(1)}$.

Proof: See Appendix B. ■

C. Derivation for $Q\left(\mathcal{E}\right)$

The variational density $Q\left(\mathcal{E}\right)$ can be obtained by taking only the terms relevant to $\mathcal{E}$ after substituting (14) into (17), and can be expressed as

$$Q\left(\mathcal{E}\right) \propto \prod_{i_1=1}^{I_1}\cdots\prod_{i_N=1}^{I_N} \exp\Bigg\{\mathbb{E}_{\prod_{\Theta_j\neq\mathcal{E}} Q(\Theta_j)}\Bigg[-\zeta_{i_1,\cdots,i_N}\Big|\mathcal{E}_{i_1,\cdots,i_N}\Big|^2 - \beta\Big|\mathcal{Y}_{i_1,\cdots,i_n} - \sum_{l=1}^{L}\prod_{n=1}^{N}\mathbf{\Xi}_{i_n,l}^{(n)} - \mathcal{E}_{i_1,\cdots,i_N}\Big|^2\Bigg]\Bigg\}. \tag{26}$$

After taking expectations, the term inside the exponent of (26) is

$$- \mathcal{E}_{i_1,\cdots,i_N}^*\underbrace{\Big(\mathbb{E}_{Q(\beta)}[\beta] + \mathbb{E}_{Q(\zeta_{i_1,\cdots,i_N})}\left[\zeta_{i_1,\cdots,i_N}\right]\Big)}_{\triangleq p_{i_1,\cdots,i_N}}\mathcal{E}_{i_1,\cdots,i_N} + 2\mathfrak{Re}\Bigg[\mathcal{E}_{i_1,\cdots,i_N}^* p_{i_1,\cdots,i_N} \times \underbrace{\mathbb{E}_{Q(\beta)}[\beta]p_{i_1,\cdots,i_N}^{-1}\Big(\mathcal{Y}_{i_1,\cdots,i_N} - \sum_{l=1}^{L}\Big(\prod_{n=1}^{N}\mathbb{E}_{Q(\mathbf{\Xi}^{(n)})}[\mathbf{\Xi}_{i_n,l}^{(n)}]\Big)\Big)}_{\triangleq m_{i_1,\cdots,i_N}}\Bigg]. \tag{27}$$

Since (27) is a quadratic function with respect to $\mathcal{E}_{i_1,\cdots,i_N}$, it is easy to show that

$$Q\left(\mathcal{E}\right) = \prod_{i_1=1}^{I_1}\cdots\prod_{i_N=1}^{I_N}\mathcal{CN}\left(\mathcal{E}_{i_1,\cdots,i_N} \mid m_{i_1,\cdots,i_N}, p_{i_1,\cdots,i_N}^{-1}\right). \tag{28}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

Notice that from (27), the computation of outlier mean $m_{i_1,\cdots,i_N}$ can be rewritten as $m_{i_1,\cdots,i_N} = \mathfrak{n}_1\mathfrak{n}_2$, where $\mathfrak{n}_1 = \frac{\left(\mathbb{E}_{Q(\zeta_{i_1,\cdots,i_N})}[\zeta_{i_1,\cdots,i_N}]\right)^{-1}}{\left(\mathbb{E}_{Q(\zeta_{i_1,\cdots,i_N})}[\zeta_{i_1,\cdots,i_N}]\right)^{-1}+\left(\mathbb{E}_{Q(\beta)}[\beta]\right)^{-1}}$ and $\mathfrak{n}_2 = \mathcal{Y}_{i_1,\cdots,i_N} - \sum_{l=1}^{L}\left(\prod_{n=1}^{N}\mathbb{E}_{Q(\boldsymbol{\Xi}^{(n)})}[\boldsymbol{\Xi}^{(n)}_{i_n,l}]\right)$. From the general data model in (6), it can be seen that $\mathfrak{n}_2$ consists of the estimated outliers plus noise. On the other hand, since $\left(\mathbb{E}_{Q(\zeta_{i_1,\cdots,i_N})}[\zeta_{i_1,\cdots,i_N}]\right)^{-1}$ and $\left(\mathbb{E}_{Q(\beta)}[\beta]\right)^{-1}$ can be interpreted as the estimated power of the outliers and the noise respectively, $\mathfrak{n}_1$ represents the strength of the outliers in the estimated outliers plus noise. Therefore, if the estimated power of the outliers $\left(\mathbb{E}_{Q(\zeta_{i_1,\cdots,i_N})}[\zeta_{i_1,\cdots,i_N}]\right)^{-1}$ goes to zero, the outlier mean $m_{i_1,\cdots,i_N}$ becomes zero accordingly.

D. Derivations for $Q(\gamma_l)$, $Q(\zeta_{i_1,\cdots,i_N})$, and $Q(\beta)$

Using (14) and (17) again, the variational density $Q(\gamma_l)$ can be expressed as

$$Q(\gamma_l) \propto \exp\Bigg\{\underbrace{(\sum_{n=P+1}^{N} I_n - 1)}_{\triangleq\tilde{a}_l}\ln\gamma_l - \gamma_l\underbrace{\Big[\sum_{n=P+1}^{N}\mathbb{E}_{Q(\boldsymbol{\Xi}^{(n)})}\big[\boldsymbol{\Xi}^{(n)H}_{:,l}\boldsymbol{\Xi}^{(n)}_{:,l}\big]\Big]}_{\triangleq\tilde{b}_l}\Bigg\} \quad (29)$$

which has the same functional form as the probability density function of gamma distribution, i.e., $Q(\gamma_l) = \text{gamma}(\gamma_l \mid \tilde{a}_l, \tilde{b}_l)$. Since $\mathbb{E}_{Q(\gamma_l)}[\gamma_l] = \tilde{a}_l/\tilde{b}_l$ is required for updating the variational distributions of other variables in $\boldsymbol{\Theta}$, we need to compute $\tilde{a}_l$ and $\tilde{b}_l$. While computation of $\tilde{a}_l$ is straightforward, the computation of $\tilde{b}_l$ can be facilitated by using the correlation property of the matrix normal distribution $\mathbb{E}_{Q(\boldsymbol{\Xi}^{(n)})}[\boldsymbol{\Xi}^{(n)H}_{:,l}\boldsymbol{\Xi}^{(n)}_{:,l}] = \mathbf{M}^{(n)H}_{:,l}\mathbf{M}^{(n)}_{:,l} + I_n\boldsymbol{\Sigma}^{(n)}_{l,l}$ [35] for $P+1 \le n \le N$.

Similarly, using (14) and (17), the variational densities $Q(\zeta_{i_1,\cdots,i_N})$ and $Q(\beta)$ can be found to be gamma distributions as

$$Q(\zeta_{i_1,\cdots,i_N}) = \text{gamma}\left(\zeta_{i_1,\cdots,i_N} \mid \tilde{c}_{i_1,\cdots,i_N}, \tilde{d}_{i_1,\cdots,i_N}\right) \quad (30)$$

$$Q(\beta) = \text{gamma}\left(\beta \mid \tilde{e}, \tilde{f}\right) \quad (31)$$

with parameters $\tilde{c}_{i_1,\cdots,i_N} = c_{i_1,\cdots,i_N} + 1$, $\tilde{d}_{i_1,\cdots,i_N} = d_{i_1,\cdots,i_N} + (m_{i_1,\cdots,i_N})^* m_{i_1,\cdots,i_N} + p^{-1}_{i_1,\cdots,i_N}$, $\tilde{e} = \prod_{n=1}^{N} I_n$, and $\tilde{f} = \mathbb{E}_{\prod_{n=1}^{N}Q(\boldsymbol{\Xi}^{(n)})Q(\mathcal{E})}\Big[\parallel \mathcal{Y} - [\![\boldsymbol{\Xi}^{(1)},\cdots,\boldsymbol{\Xi}^{(N)}]\!] - \mathcal{E}\parallel_F^2\Big]$. For $\tilde{c}_{i_1,\cdots,i_N}$, $\tilde{d}_{i_1,\cdots,i_N}$ and $\tilde{e}$, the computations are straightforward. $\tilde{f}$ is derived in Appendix C to be

$$\begin{aligned}\tilde{f} =& \parallel \mathcal{Y} - \mathcal{M} \parallel_F^2 + \sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N} p^{-1}_{i_1,\cdots,i_N} \\ &+ \text{Tr}\Big(\mathfrak{D}\Big[\bigodot_{n=P+1}^{N}\Big(\mathbf{M}^{(n)H}\mathbf{M}^{(n)} + I_n\boldsymbol{\Sigma}^{(n)}\Big)^*\Big]\Big) \\ &- 2\mathfrak{Re}\Big(\text{Tr}\Big(\mathfrak{U}^{(1)}\Big[\mathcal{Y} - \mathcal{M}\Big]\Big[\Big(\diamond_{n=P+1}^{N}\mathbf{M}^{(n)}\Big) \\ &\diamond\Big(\diamond_{n=2}^{P}\hat{\boldsymbol{\Xi}}^{(n)}\Big)\Big]^*\hat{\boldsymbol{\Xi}}^{(1)H}\Big)\Big) \quad (32)\end{aligned}$$

where $\mathcal{M}$ is a tensor with its $(i_1,\cdots,i_N)^{th}$ element being $m_{i_1,\cdots,i_N}$, and $\mathfrak{Re}(\cdot)$ denotes the real part of its argument. Although equation (32) for computing $\tilde{f}$ is complicated, its meaning is clear when we refer to its definition below (31), from which it can be seen that $\tilde{f}$ represents the estimate of the overall noise power.

E. Summary of the Iterative Algorithm

From the expressions for $Q(\Theta_k)$ evaluated above, it is seen that the calculation of a particular $Q(\Theta_k)$ relies on the statistics of other variables in $\boldsymbol{\Theta}$. As a result, the variational distribution for each variable in $\boldsymbol{\Theta}$ should be iteratively updated. The iterative algorithm is summarized as follows.

Initializations:

Choose $L > R$ and initial values $\{\hat{\boldsymbol{\Xi}}^{(n,0)}\}_{n=1}^{P}$, $\{\mathbf{M}^{(n,0)}, \boldsymbol{\Sigma}^{(n,0)}\}_{n=P+1}^{N}$, $\tilde{b}_l^0$, $\{\tilde{c}^0_{i_1,\cdots,i_N}, \tilde{d}^0_{i_1,\cdots,i_N}\}$ and $\tilde{f}^0$ for all l and $i_1,\cdots,i_N$.

Let $\tilde{a}_l = \sum_{n=P+1}^{N} I_n$ and $\tilde{e} = \prod_{n=1}^{N} I_n$.

Iterations: For the t^{th} iteration ($t \ge 1$),

Update the statistics of outliers: $\{p_{i_1,\cdots,i_N}, m_{i_1,\cdots,i_N}\}_{i_1=1,\cdots,i_N=1}^{I_1,\cdots,I_N}$

$$p^t_{i_1,\cdots,i_N} = \frac{\tilde{e}}{\tilde{f}^{t-1}} + \frac{\tilde{c}^{t-1}_{i_1,\cdots,i_N}}{\tilde{d}^{t-1}_{i_1,\cdots,i_N}} \quad (33)$$

$$\begin{aligned}m^t_{i_1,\cdots,i_N} =& \frac{\tilde{e}}{\tilde{f}^{t-1}p^t_{i_1,\cdots,i_N}}\Bigg(\mathcal{Y}_{i_1,\cdots,i_N} \\ &- \sum_{l=1}^{L}\Big[\Big(\prod_{n=1}^{P}\hat{\boldsymbol{\Xi}}^{(n,t-1)}_{i_n,l}\Big)\Big(\prod_{n=P+1}^{N}\mathbf{M}^{(n,t-1)}_{i_n,l}\Big)\Big]\Bigg)\end{aligned} \quad (34)$$

Update the statistics of factor matrices: $\{\mathbf{M}^{(k)}, \boldsymbol{\Sigma}^{(k)}\}_{k=P+1}^{N}$

$$\begin{aligned}&\boldsymbol{\Sigma}^{(k,t)} \\ &= \Bigg(\frac{\tilde{e}}{\tilde{f}^{t-1}}\mathfrak{D}\Big[\bigodot_{n=P+1,n\neq k}^{N}\Big(\mathbf{M}^{(n,t-1)H}\mathbf{M}^{(n,t-1)} + I_n\boldsymbol{\Sigma}^{(n,t-1)}\Big)^*\Big] \\ &\qquad + \mathbf{diag}\Big\{\frac{\tilde{a}_1}{\tilde{b}_1^{t-1}}, ..., \frac{\tilde{a}_L}{\tilde{b}_L^{t-1}}\Big\}\Bigg)^{-1}\end{aligned} \quad (35)$$

$$\begin{aligned}&\mathbf{M}^{(k,t)} = \frac{\tilde{e}}{\tilde{f}^{t-1}}\Big(\mathfrak{U}^{(k)}\left[\mathcal{Y} - \mathcal{M}^t\right]\Big) \\ &\times\Big[\Big(\diamond_{n=P+1,n\neq k}^{N}\mathbf{M}^{(n,t-1)}\Big)\diamond\Big(\diamond_{n=1}^{P}\hat{\boldsymbol{\Xi}}^{(n,t-1)}\Big)\Big]^*\boldsymbol{\Sigma}^{(k,t)}\end{aligned} \quad (36)$$

Update the orthonormal factor matrices $\{\hat{\boldsymbol{\Xi}}^{(k)}\}_{k=1}^{P}$

$$\begin{aligned}\left[\boldsymbol{\Upsilon}^{(k,t)}, \boldsymbol{\Pi}^{(k,t)}\right] =& \text{SVD}\Bigg[\frac{\tilde{e}}{\tilde{f}^{t-1}}\mathfrak{U}^{(k)}\Big[\mathcal{Y} - \mathcal{M}^t\Big] \\ &\times\Big[\Big(\diamond_{n=P+1}^{N}\mathbf{M}^{(n,t)}\Big)\diamond\Big(\diamond_{n=1,n\neq k}^{P}\hat{\boldsymbol{\Xi}}^{(n,t-1)}\Big)\Big]^*\Bigg] \\ \hat{\boldsymbol{\Xi}}^{(k,t)} =& \boldsymbol{\Upsilon}^{(k,t)}\boldsymbol{\Pi}^{(k,t)H}\end{aligned} \quad (37)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

Update $\{\tilde{b}_l\}_{l=1}^{L}$, $\{\tilde{d}_{i_1,\cdots,i_N}\}_{i_1=1,\cdots,i_N=1}^{I_1,\cdots,I_N}$ and $\tilde{f}$

$$\tilde{b}_l^t = \sum_{n=P+1}^{N} \mathbf{M}_{:,l}^{(n,t)H}\mathbf{M}_{:,l}^{(n,t)} + I_n\mathbf{\Sigma}_{l,l}^{(n,t)} \tag{38}$$

$$\tilde{c}_{i_1,\cdots,i_N}^t = \tilde{c}_{i_1,\cdots,i_N}^0 + 1 \tag{39}$$

$$\tilde{d}_{i_1,\cdots,i_N}^t = \tilde{d}_{i_1,\cdots,i_N}^0 + \left(m_{i_1,\cdots,i_N}^t\right)^* m_{i_1,\cdots,i_N}^t + 1/p_{i_1,\cdots,i_N}^t \tag{40}$$

$$\begin{aligned}\tilde{f}^t =& \| \mathcal{Y} - \mathcal{M}^t \|_F^2 + \sum_{i_1=1}^{I_1}\cdots\sum_{i_n=1}^{I_N}(p_{i_1,\cdots,i_N}^t)^{-1} \\ &+\mathrm{Tr}\left(\mathfrak{D}\Big[\bigodot_{n=P+1}^{N}\left(\mathbf{M}^{(n,t)H}\mathbf{M}^{(n,t)}+I_n\mathbf{\Sigma}^{(n,t)}\right)^*\Big]\right) \\ &- 2\mathfrak{Re}\Bigg(\mathrm{Tr}\Bigg(\mathfrak{U}^{(1)}\Big[\mathcal{Y}-\mathcal{M}^t\Big]\Big[\Big(\diamondsuit_{n=P+1}^{N}\mathbf{M}^{(n,t)}\Big) \\ &\qquad \diamond\Big(\diamondsuit_{n=2}^{P}\hat{\mathbf{\Xi}}^{(n,t)}\Big)\Big]^*\hat{\mathbf{\Xi}}^{(1,t)H}\Bigg)\Bigg)\end{aligned} \tag{41}$$

Until Convergence

F. Further Discussions

To gain more insights from the above proposed CPD algorithm, discussions on its convergence property, automatic rank determination, relationship to the OALS algorithm and computational complexity are presented in the following.

1) Convergence Property: Although the functional minimization of the KL divergence in (15) is non-convex over the mean-field family $Q(\mathbf{\Theta}) = \prod_k Q(\Theta_k)$, it is convex with respect to a single variational density $Q(\Theta_k)$ when the others $\{Q(\Theta_j)|j \neq k\}$ are fixed [29]. Therefore, the proposed algorithm, which iteratively updates the optimal solution for each Θ_k, is essentially a coordinate-descent algorithm in the functional space of variational distributions with each update solving a convex problem. This guarantees monotonic decrease of the KL divergence in (15), and the proposed algorithm is guaranteed to converge to at least a stationary point [Theorem 2.1, 39].

2) Automatic Rank Determination: The automatic rank determination for the tensor CPD uses an idea from the Bayesian model selection (or Bayesian Occam's razor) [pp. 157, 27]. More specifically, the parameters $\{\gamma_l\}_{l=1}^{L}$ control the model complexity, and their optimal variational densities are obtained together with those of other parameters by minimizing the KL divergence. After convergence, if some $\mathbb{E}[\gamma_l]$ are very large, e.g., 10^6, this indicates that their corresponding columns in $\{\mathbf{M}^{(n)}\}_{n=P+1}^{N}$ can be "switched off", as they play no role in explaining the data. Furthermore, according to the definition of the tensor CPD in (1), the corresponding columns in $\{\hat{\mathbf{\Xi}}^{(n)}\}_{n=1}^{P}$ should also be pruned accordingly. Finally, the learned tensor rank R is the number of remaining columns in each estimated factor matrix $\hat{\mathbf{\Xi}}^{(n)}$.

3) Relationship to the OALS: If the tensor rank R is known, the regularization term in (8) is not needed, and consequently there are no parameters $\{\tilde{a}_l, \tilde{b}_l\}_{l=1}^{L}$. Further restricting $Q(\mathbf{\Xi}^{(k)})$ to be $\delta(\mathbf{\Xi}^{(k)} - \mathbf{M}^{(k)})$ for $P+1 \leq k \leq N$, it can be shown that all the equations in the above algorithm still hold except the term $I_n\mathbf{\Sigma}^{(n,t-1)}$ in (35) and the term $I_n\mathbf{\Sigma}^{(n,t)}$ in (41) are removed. Then, the proposed algorithm is a robust version of OALS, even covering the case of $P = N$. If we further have the knowledge that outliers do not exist, only (35)-(37) remain. Interestingly, this resulting algorithm is exactly the OALS algorithm in [17]. In this regard, the proposed algorithm not only provides a probabilistic interpretation of the OALS algorithm, but also has the additional properties in automatic rank determination, outlier removal and learning of the noise power.

4) Computational Complexity: For each iteration, the complexity is dominated by updating each factor matrix, costing $O(\sum_{n=1}^{N} I_n L^2 + N\prod_{n=1}^{N} I_n L)$. Thus, the overall complexity is about $O(q(\sum_{n=1}^{N} I_n L^2 + N\prod_{n=1}^{N} I_n L))$ where q is the number of iterations needed for convergence. On the other hand, for the OALS algorithm with exact tensor rank R, its complexity is $O(m(\sum_{n=1}^{N} I_n R^2 + N\prod_{n=1}^{N} I_n R))$ where m is the number of iterations needed for convergence. Therefore, for each iteration, the complexity of the proposed algorithm is comparable to that of the OALS algorithm.

V. Simulation Results and Discussions

In this section, numerical simulations are presented to assess the performance of the proposed algorithm (labeled as VB) using synthetic data and two applications, in comparison with various state-of-the-art tensor CPD algorithms. The algorithms being compared include the ALS [15], the simultaneous diagonalization method for coupled tensor CPD (labeled as SD) [40], the direct algorithm for CPD followed by enhanced ALS (labeled as DIAG-A) [41], the Bayesian tensor CPD (labeled as BCPD) [32], the robust iteratively reweighed ALS (labeled as IRALS) [42], and the OALS algorithm (labeled as OALS) [17]. In all experiments, three outlier models are considered, and they are listed in Table I. For all the simulated algorithms, the initial factor matrix $\hat{\mathbf{\Xi}}^{(n,0)}$ is set as the matrix consisting of L leading left singular vectors of $\mathfrak{U}^{(n)}[\mathcal{Y}]$ where $L = \max\{I_1, I_2, \cdots, I_N\}$ for the proposed algorithm and the BCPD, and $L = R$ for other algorithms. The initial parameters of the proposed algorithm $\{\tilde{c}_{i_1,\cdots,i_N}^0, \tilde{d}_{i_1,\cdots,i_N}^0\}$ are set as 10^{-6} for all $i_1,\cdots,i_N$, $\tilde{b}_l^0 = \sum_{n=P+1}^{N} I_n$ for all l, $\tilde{f}^0 = \prod_{n=1}^{N} I_n$, and $\{\mathbf{\Sigma}^{(n,0)}\}_{n=P+1}^{N}$ are all set to be $\mathbf{I}_L$. All the algorithms terminate at the t^{th} iteration when $\| [\![\mathbf{A}^{(1,t)}, \mathbf{A}^{(2,t)}, \cdots, \mathbf{A}^{(N,t)}]\!] - [\![\mathbf{A}^{(1,t-1)}, \mathbf{A}^{(2,t-1)}, \cdots, \mathbf{A}^{(N,t-1)}]\!] \|_F^2 < 10^{-6}$ or the iteration number exceeds 2000.

A. Validation on Synthetic Data

Synthetic tensors are used in this subsection to assess the performance of the proposed algorithm on convergence, rank learning ability and factor matrix recovery under different outlier models. A complex-valued third-order tensor $[\![\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}]\!] \in \mathbb{C}^{12\times12\times12}$ with rank $R = 5$ is considered, where the orthogonal factor matrix $\mathbf{A}^{(1)}$ is constructed from the R leading left singular vectors of a matrix drawn from $\mathcal{CMN}(\mathbf{A}|\mathbf{0}_{12\times5,}, \mathbf{I}_{12\times12}, \mathbf{I}_{5\times5})$, and the factor matrices $\{\mathbf{A}^{(n)}\}_{n=2}^{3}$ are drawn from $\mathcal{CMN}(\mathbf{A}|\ \mathbf{0}_{12\times5,}, \mathbf{I}_{12\times12}, \mathbf{I}_{5\times5})$. Parameters for outlier models are set as: $\pi = 0.05$, $\sigma_e^2 = 100$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

Table I: Three Different Outlier Models

Scenario	Variable Description
Bernoulli-Gaussian	$\mathcal{E}_{i_1,\cdots,i_N} \sim \mathcal{CN}(0,\sigma_e^2)$ with a probability π
Bernoulli-Uniform	$\mathcal{E}_{i_1,\cdots,i_N} \sim \mathcal{U}(-H,H)$ with a probability π
Bernoulli-Student's t	$\mathcal{E}_{i_1,\cdots,i_N} \sim \mathcal{T}(\mu,\lambda,\nu)$ with a probability π

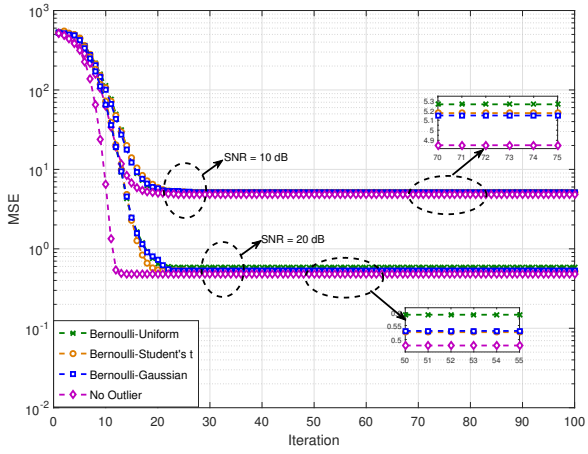
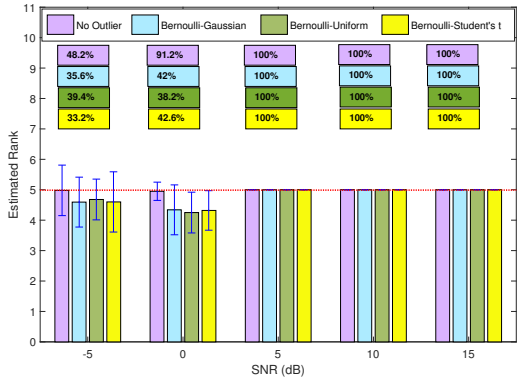


Figure 3: Convergence of the proposed algorithm under different outlier models

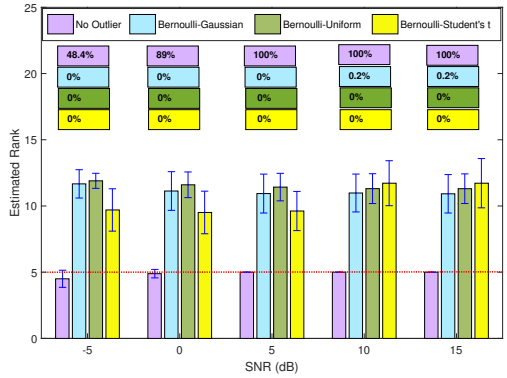
$H = 10\arg\max_{i_1,\cdots,i_N} |[\![\mathbf{A}^{(1)},\mathbf{A}^{(2)},\mathbf{A}^{(3)}]\!]_{i_1,\cdots,i_N}|$, $\mu = 3$, $\lambda = 1/50$ and $\nu = 10$. The signal-to-noise ratio (SNR) is defined as $10\log_{10}(\| [\![\mathbf{A}^{(1)},\mathbf{A}^{(2)},\mathbf{A}^{(3)}]\!] \|_F^2 \ / \ \| \mathcal{W} \|_F^2)$ [5], [17]. Each result in this subsection is obtained by averaging 500 Monte-Carlo runs.

Figure 3 presents the convergence performance of the proposed algorithm under different outlier models, where the mean-square-error (MSE) $\| [\![\hat{\mathbf{\Xi}}^{(1)},\hat{\mathbf{\Xi}}^{(2)},\hat{\mathbf{\Xi}}^{(3)}]\!] - [\![\mathbf{A}^{(1)},\mathbf{A}^{(2)},\mathbf{A}^{(3)}]\!] \|_F^2$ is chosen as the assessment criterion. From Figure 3, it can be seen that the MSEs decrease significantly in the first few iterations and converge to stable values quickly, demonstrating the rapid convergence property. Furthermore, by comparing the simulation results with outliers to that without outlier, it is clear that the proposed algorithm is effective in mitigating outliers.

For tensor rank learning, the simulation results of the proposed algorithm are shown in Figure 4 (a), while those of the Bayesian tensor CPD algorithm are shown in Figure 4 (b). Each vertical bar in the figures shows the mean and standard deviation of rank estimates, with the red horizontal dotted lines indicating the true tensor rank. The percentages of correct estimates are also shown on top of the figures. From Figure 4 (a), it is seen that the proposed method can recover the true tensor rank with 100% accuracy when $\mathrm{SNR} \geq 5$ dB, both with or without outliers. This shows the accuracy and robustness of the proposed algorithm when the noise power is moderate. Even though the performance at low SNRs is not as impressive as that at high SNRs, it can be observed that the proposed algorithm still gives estimates close to the true tensor rank with the true rank lying mostly within one standard deviation from the mean estimate. On the other hand, in Figure 4 (b), it is observed that while the Bayesian tensor CPD algorithm performs nearly the same as the proposed



(a)



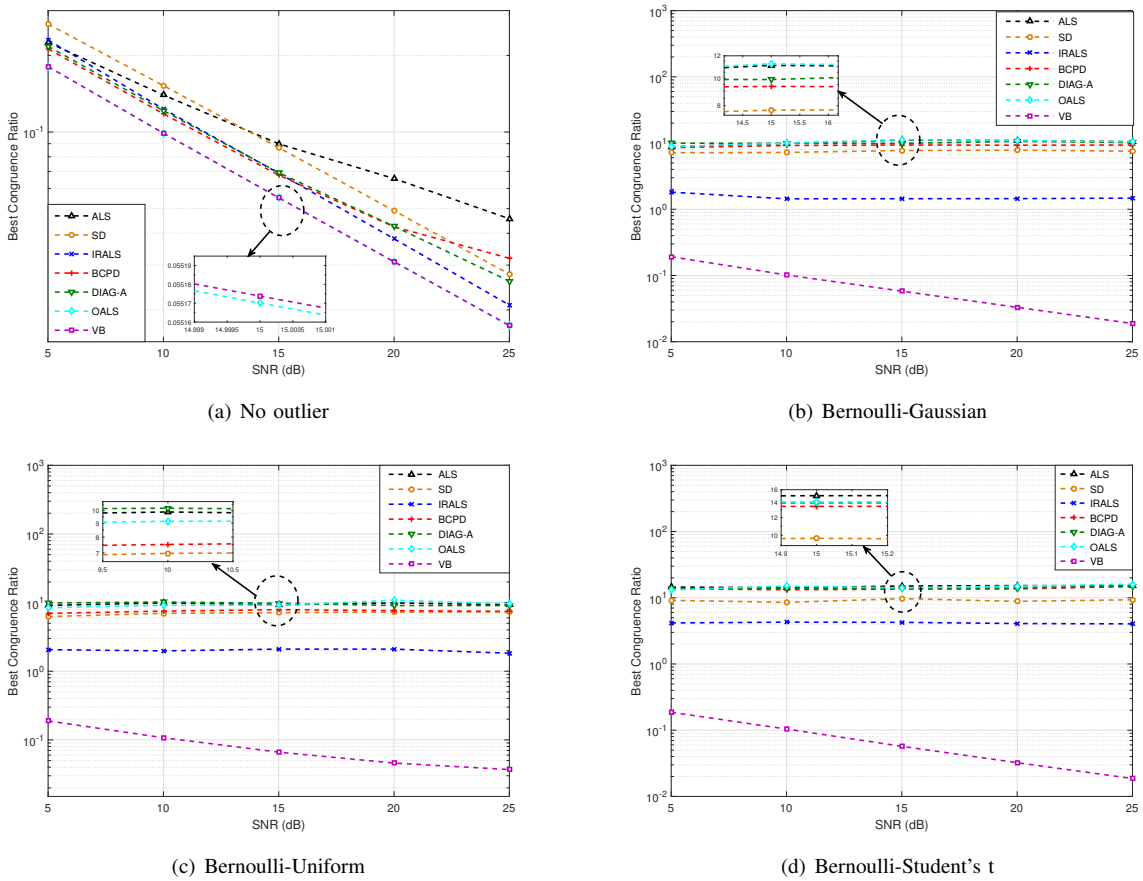
(b)

Figure 4: Rank determination using (a) the proposed method and (b) the Bayesian tensor CPD [32]

algorithm without outliers, it gives tensor rank estimates very far away from the true value when outliers are present.

Figure 5 compare the proposed algorithm to other state-of-the-art CPD algorithms in terms of recovery accuracy of the orthogonal factor matrix $\mathbf{A}^{(1)}$ under different outlier models. The criterion is set as the best congruence ratio defined as $\min_{\mathbf{\Delta P}} \| \mathbf{A}^{(1)} - \hat{\mathbf{\Xi}}^{(1)}\mathbf{P\Delta} \|_F \ / \ \| \mathbf{A}^{(1)} \|_F$, where the diagonal matrix $\mathbf{\Delta}$ and the permutation matrix $\mathbf{P}$ are found via the greedy least-squares column matching algorithm [5]. From Figure 5 (a), it is seen that both the proposed algorithm and OALS perform better than other algorithms when outliers are absent. This shows the importance of incorporating the orthogonality information of the factor matrix. On the other hand, while OALS offers the same performance as the proposed algorithm when there is no outlier, its performance is significantly different in the presence of outliers, as presented in Figure 5 (b)-(d). Furthermore, since all the algorithms except VB and IRALS have not taken the outliers into account, their performances degrade significantly as shown in Figure 5 (b)-(d). Even though the IRALS uses the robust l_p $(0 < p \leq 1)$ norm optimization to alleviate the effects of outliers, it cannot learn the statistical information of the outliers, leading to its worse performance in outliers mitigation than that of the proposed algorithm.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing



(a) No outlier

(b) Bernoulli-Gaussian

(c) Bernoulli-Uniform

(d) Bernoulli-Student's t

Figure 5: Performance of factor matrix recovery versus SNR under different outlier models

B. Blind Data Detection for DS-CDMA Systems

In this subsection, we consider an uplink DS-CDMA system, in which $R = 5$ users communicate with the BS equipped with $M = 8$ antennas over flat fading channels $h_{mr} \sim \mathcal{CN}(h_{mr}|0,1)$. The transmitted data $s_r(k)$ are random binary phase-shift keying (BPSK) symbols. The spreading code is of length $Z = 6$, and with each code element $c_{zr} \sim \mathcal{CN}(c_{zr}|0,1)$. After observing the received tensor $\mathcal{Y} \in \mathbb{C}^{8\times 6\times 100}$, the proposed algorithm and other state-of-the-art tensor CPD algorithms, combined with ambiguity removal and constellation mapping [5], [9], are executed to blindly detect the transmitted data. Their performance is measured in terms of bit error rate (BER).

The BERs versus SNR under different outlier models are presented in Figure 6, which are averaged over 10000 independent trials. The parameter settings for different outlier models are the same as those in the last subsection. It is seen from Figure 6 (a) that when there is no outlier, the proposed algorithm and OALS behave the same, and both outperform other CPDs. However, when outliers exist, it is seen from Figure 6 (b)-(d) that the proposed algorithm performs significantly better than other algorithms.

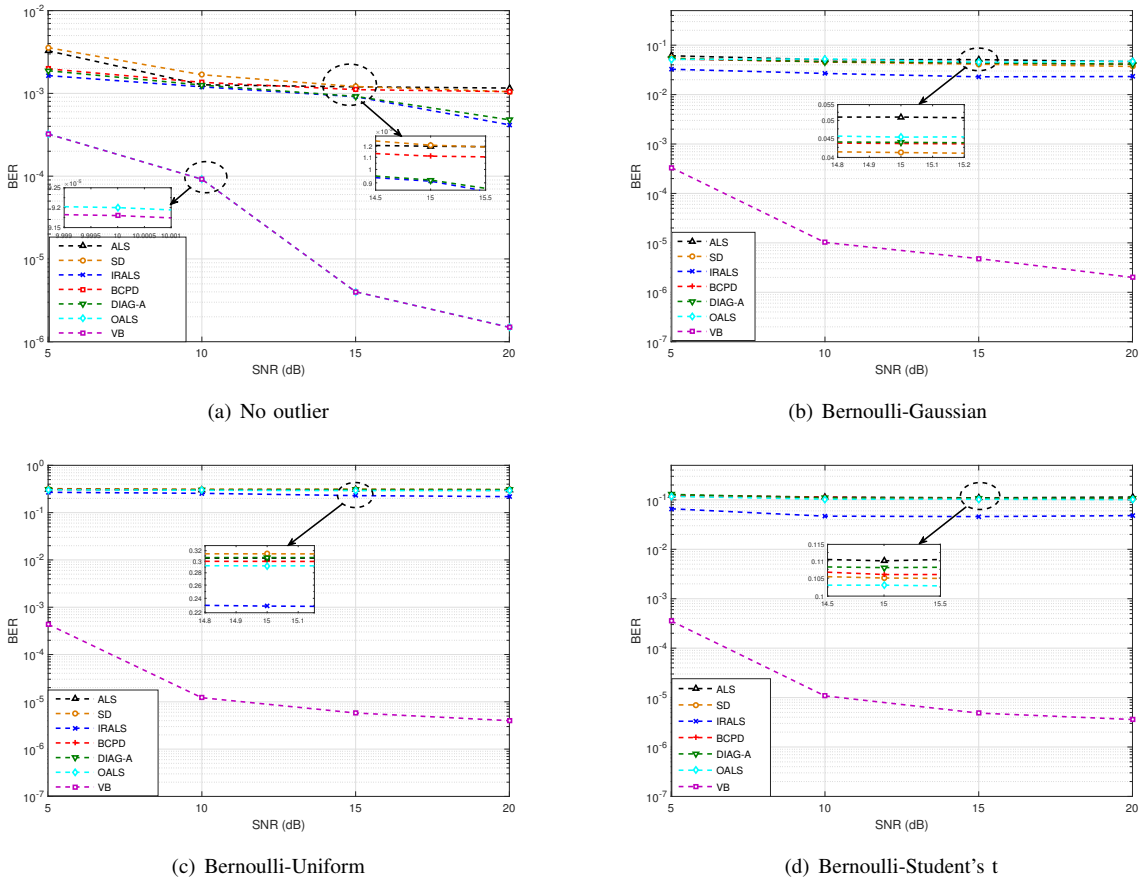
C. Linear Image Coding for Face Images

In this subsection, we conduct experiments on 165 face images from the Yale Face Database[3] [38], representing different facial expressions (also with or without sunglasses) of 15 people (11 images for each person). In each classification experiment, we randomly pick two people's images. Among these 22 images, 12 (6 from each person) are used for training. In particular, each image is of size 240×320, and the training data can be naturally represented by a third-order tensor $\mathcal{Y} \in \mathbb{R}^{240\times 320\times 12}$. Various state-of-the-art tensor CPD algorithms and the proposed algorithm are run to learn the two orthogonal basis matrices (see (4)). Then, the feature vectors of these 12 training images, which are obtained by projecting them onto the multilinear subspaces spanned by the two orthogonal basis matrices, are used to train a support vector machine (SVM) classifier. For the 10 testing images, their feature vectors are fed into the SVM classifier to determine which person is in each image. The parameters of various outlier models are: $\pi = 0.05$, $\sigma_e = 100$, $H = 100$, $\mu = 1$, $\lambda = 1/1000$ and $\nu = 20$.

Since the tensor rank is not known in the image data, it should be carefully chosen. For the algorithms (ALS, SD, IRALS, DIAG-A and OALS) that cannot automatically determine the rank, it can be obtained by first running the these algorithms with tensor rank ranges from 1 to 12, and then

[3] http://vision.ucsd.edu/content/yale-face-database

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing



(a) No outlier

(b) Bernoulli-Gaussian

(c) Bernoulli-Uniform

(d) Bernoulli-Student's t

Figure 6: BER versus SNR under different outlier models

Table II: Classification Error and CPD Computation Time in Face Recognition

Algorithm	No Outlier		Bernoulli-Gaussian		Bernoulli-Uniform		Bernoulli-Student's t	
	Classification Error	CPD Time (s)	Classification Error	CPD Time (s)	Classification Error	CPD Time (s)	Classification Error	CPD Time (s)
ALS [15]	9%	1.9635	51%	46.3041	47%	63.6765	48%	58.1047
SD [40]	12%	0.5736	49%	0.6287	44%	0.6181	49%	0.6594
IRALS [42]	9%	4.3527	43%	15.8361	27%	16.6151	36%	17.5181
BCPD [32]	2%	4.1546	53%	20.7338	35%	17.9896	48%	17.1151
DIAG-A [41]	11%	3.7384	50%	28.9961	41%	30.6317	51%	22.5127
OALS [17]	11%	1.0174	58%	40.2731	34%	38.1754	47%	24.0162
VB	2%	2.4827	10%	2.7806	6%	2.4912	7%	2.8895

finding the knee point of the reconstruction error decrement [27]. When there is no outlier, it is able to find the appropriate tensor rank. However, when outliers exist, the knee point cannot be found and we set the rank as the upper bound 12. For the BCPD, although it learns the appropriate rank when there is no outliers, it learns the rank as 12 when outliers exist. On the other hand, no matter whether there are outliers or not, the proposed algorithm automatically learns the appropriate tensor rank without exhaustive search, and thus saves considerable computational complexity.

The average classification errors of 10 independent experiments and the corresponding average CPD computation times (benchmarked in Matlab on a personal computer with an i7 CPU) are shown in Table II, and it can be seen that the proposed algorithm provides the smallest classification error under all considered scenarios.

VI. Conclusions

In this paper, a probabilistic CPD algorithm with orthogonal factors has been proposed for complex-valued tensors, under unknown tensor rank and in the presence of outliers. It has been shown that, without knowledge of noise power and outlier statistics, the proposed algorithm alternatively estimates the factor matrices, recovers the tensor rank and mitigates the outliers. Interestingly, the popular OALS in [17] has been shown to be a special case of the proposed algorithm. Simulation results using synthetic data and real-world applications have demonstrated the excellent performance of the proposed algorithm in terms of accuracy and robustness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

APPENDIX A
DERIVATIONS OF $Q(\boldsymbol{\Xi}^{(k)}), P+1 \leq k \leq N$

By substituting (14) into (17) and only taking the terms relevant to $\boldsymbol{\Xi}^{(k)}$, we directly have

$$Q\left(\boldsymbol{\Xi}^{(k)}\right) \propto \exp\left\{\mathbb{E}_{\prod_{\Theta_j \neq \boldsymbol{\Xi}^{(k)}} Q(\Theta_j)}\Big[- \beta \parallel \mathcal{Y} - [\![\boldsymbol{\Xi}^{(1)}, \cdots, \boldsymbol{\Xi}^{(N)}]\!] - \mathcal{E} \parallel_F^2 - \mathrm{Tr}\left(\boldsymbol{\Gamma}\boldsymbol{\Xi}^{(k)H}\boldsymbol{\Xi}^{(k)}\right)\Big]\right\}. \tag{42}$$

Using that $\parallel \mathcal{A} \parallel_F^2 = \parallel \mathfrak{U}^{(n)}[\mathcal{A}] \parallel_F^2 = \mathrm{Tr}\Big(\mathfrak{U}^{(n)}[\mathcal{A}]\left(\mathfrak{U}^{(n)}[\mathcal{A}]\right)^H\Big)$, the square of the Frobenius norm inside expectation in (42) can be expressed as

$$\begin{aligned}\mathrm{Tr}\Bigg(&\boldsymbol{\Xi}^{(k)}\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\boldsymbol{\Xi}^{(n)}\big)^T\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\boldsymbol{\Xi}^{(n)}\big)^*\boldsymbol{\Xi}^{(k)H}\\ &- \boldsymbol{\Xi}^{(k)}\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\boldsymbol{\Xi}^{(n)}\big)^T\Big(\mathfrak{U}^{(k)}\big[\mathcal{Y}-\mathcal{E}\big]\Big)^H\\ &- \mathfrak{U}^{(k)}\big[\mathcal{Y}-\mathcal{E}\big]\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\boldsymbol{\Xi}^{(n)}\big)^*\boldsymbol{\Xi}^{(k)H}\\ &+ \mathfrak{U}^{(k)}\big[\mathcal{Y}-\mathcal{E}\big]\Big(\mathfrak{U}^{(k)}\big[\mathcal{Y}-\mathcal{E}\big]\Big)^H\Bigg).\end{aligned} \tag{43}$$

By putting (43) into (42), and distributing the expectations into various terms, (42) becomes (44) at the top of the next page.

After completing the square over $\boldsymbol{\Xi}^{(k)}$, it can be seen that (44) corresponds to the functional form of a circularly-symmetric complex matrix normal distribution [35]. In particular, with $\mathcal{CMN}(\mathbf{X}|\mathbf{M},\boldsymbol{\Sigma}_r,\boldsymbol{\Sigma}_c)$ denoting the distribution $p(\mathbf{X}) \propto \exp\{-\mathrm{Tr}(\boldsymbol{\Sigma}_c^{-1}(\mathbf{X}-\mathbf{M})^H\boldsymbol{\Sigma}_r^{-1}(\mathbf{X}-\mathbf{M}))\}$, we have $Q\left(\boldsymbol{\Xi}^{(k)}\right) = \mathcal{CMN}(\boldsymbol{\Xi}^{(k)}|\mathbf{M}^{(k)},\mathbf{I}_{I_k},\boldsymbol{\Sigma}^{(k)})$.

APPENDIX B
PROOF OF PROPERTY 2

To prove **Property 2**, we first introduce the following lemma.

Lemma 1. *Suppose the matrix $\mathbf{A}^{(n)} \in \mathbb{C}^{\kappa_n \times \rho}$ $(1 \leq n \leq N)$. Then, if $N \geq 2$, $\big(\overset{N}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)^T\big(\overset{N}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)^*$ can be computed as*

$$\big(\overset{N}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)^T\big(\overset{N}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)^* = \overset{N}{\underset{n=1}{\odot}}\mathbf{A}^{(n)T}\mathbf{A}^{(n)*}. \tag{45}$$

Proof: Mathematical induction is used to prove this lemma. For $N=2$, it is proved in [36] that

$$\begin{aligned}&\left(\mathbf{A}^{(2)} \diamond \mathbf{A}^{(1)}\right)^T\left(\mathbf{A}^{(2)} \diamond \mathbf{A}^{(1)}\right)^*\\ &= \left(\mathbf{A}^{(2)T}\mathbf{A}^{(2)*}\right) \odot \left(\mathbf{A}^{(1)T}\mathbf{A}^{(1)*}\right).\end{aligned} \tag{46}$$

Without loss of generality, assume that (45) holds for some $K \geq 2$, i.e.,

$$\big(\overset{K}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)^T\big(\overset{K}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)^* = \overset{K}{\underset{n=1}{\odot}}\mathbf{A}^{(n)T}\mathbf{A}^{(n)*}. \tag{47}$$

Now consider $\big[\mathbf{A}^{(K+1)} \diamond \big(\overset{K}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)\big]^T\big[\mathbf{A}^{(K+1)} \diamond \big(\overset{K}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)\big]^*$. Treating $\overset{K}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}$ as a matrix and using (46), we have $\big[\mathbf{A}^{(K+1)} \diamond \big(\overset{K}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)\big]^T\big[\mathbf{A}^{(K+1)} \diamond \big(\overset{K}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)\big]^* = \big(\mathbf{A}^{(K+1)T}\mathbf{A}^{(K+1)*}\big) \odot \big[\big(\overset{K}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)^T\big(\overset{K}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)^*\big]$. Further using (47), we have

$$\begin{aligned}&\big(\overset{K+1}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)^T\big[\big(\overset{K+1}{\underset{n=1}{\diamond}}\mathbf{A}^{(n)}\big)\big]^*\\ &= \Big(\mathbf{A}^{(K+1)T}\mathbf{A}^{(K+1)*}\Big) \odot \Big[\overset{K}{\underset{n=1}{\odot}}\mathbf{A}^{(n)T}\mathbf{A}^{(n)*}\Big]\\ &= \overset{K+1}{\underset{n=1}{\odot}}\mathbf{A}^{(n)T}\mathbf{A}^{(n)*}.\end{aligned} \tag{48}$$

Then, (45) is shown to hold for $N = K+1$. Thus, by mathematical induction, (45) holds for any $N \geq 2$. ■

Using **Lemma 1** and taking expectations, it is easy to prove that

$$\begin{aligned}&\mathbb{E}_{\prod_{n=1,n\neq k}^N p(\mathbf{A}^{(n)})}\Big[\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\mathbf{A}^{(n)}\big)^T\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\mathbf{A}^{(n)}\big)^*\Big]\\ &= \Big(\overset{P}{\underset{n=1,n\neq k}{\odot}}\mathbb{E}_{p(\mathbf{A}^{(n)})}\big[\mathbf{A}^{(n)T}\mathbf{A}^{(n)*}\big]\Big)\\ &\odot \Big(\overset{N}{\underset{n=P+1,n\neq k}{\odot}}\mathbb{E}_{p(\mathbf{A}^{(n)})}\big[\mathbf{A}^{(n)T}\mathbf{A}^{(n)*}\big]\Big).\end{aligned} \tag{49}$$

Since the matrix $\mathbf{A}^{(n)} \sim \delta(\mathbf{A}^{(n)} - \hat{\mathbf{A}}^{(n)})$ for $1 \leq n \leq P$ where $\hat{\mathbf{A}}^{(n)} \in V_\rho(\mathbb{C}^{\kappa_n})$, and the matrix $\mathbf{A}^{(n)} \sim \mathcal{CMN}(\mathbf{A}^{(n)} \mid \mathbf{M}^{(n)}, \mathbf{I}_{\kappa_n}, \boldsymbol{\Sigma}^{(n)})$ for $P+1 \leq n \leq N$, we have $\mathbb{E}_{p(\mathbf{A}^{(n)})}\big[\mathbf{A}^{(n)T}\mathbf{A}^{(n)*}\big] = \hat{\mathbf{A}}^{(n)T}\hat{\mathbf{A}}^{(n)*} = \mathbf{I}_\rho$ for $1 \leq n \leq P$, and $\mathbb{E}_{p(\mathbf{A}^{(n)})}\big[\mathbf{A}^{(n)T}\mathbf{A}^{(n)*}\big] = \mathbf{M}^{(n)T}\mathbf{M}^{(n)*} + \kappa_n\boldsymbol{\Sigma}^{(n)*}$ for $P+1 \leq n \leq N$. Then, (49) becomes

$$\begin{aligned}&\mathbb{E}_{\prod_{n=1,n\neq k}^N p(\mathbf{A}^{(n)})}\Big[\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\mathbf{A}^{(n)}\big)^T\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\mathbf{A}^{(n)}\big)^*\Big]\\ &= \mathbf{I}_\rho \odot \Big[\overset{N}{\underset{n=P+1,n\neq k}{\odot}}\Big(\mathbf{M}^{(n)H}\mathbf{M}^{(n)} + \kappa_n\boldsymbol{\Sigma}^{(n)}\Big)^*\Big]\\ &= \mathfrak{D}\Big[\overset{N}{\underset{n=P+1,n\neq k}{\odot}}\Big(\mathbf{M}^{(n)H}\mathbf{M}^{(n)} + \kappa_n\boldsymbol{\Sigma}^{(n)}\Big)^*\Big].\end{aligned} \tag{50}$$

APPENDIX C
DERIVATION OF $\tilde{f}$

Recall that $\parallel \mathcal{Y} - [\![\boldsymbol{\Xi}^{(1)}, \cdots, \boldsymbol{\Xi}^{(N)}]\!] - \mathcal{E} \parallel_F^2$ is expressed in (43). Using this result and taking expectations with respect to $Q(\boldsymbol{\Xi}^{(n)})$ for $1 \leq n \leq N$ and $Q(\mathcal{E})$, we obtain (51) at the top of the next page. Since $Q(\mathcal{E}_{i_1,\cdots,i_N}) = \mathcal{CN}(\mathcal{E}_{i_1,\cdots,i_N}|m_{i_1,\cdots,i_N}, p^{-1}_{i_1,\cdots,i_N})$, it is easy to show that $\mathrm{Tr}(\mathfrak{w}_1) = \mathrm{Tr}\Big(\big(\mathfrak{U}^{(1)}\big[\mathcal{Y}-\mathcal{M}\big]\big)^H\mathfrak{U}^{(1)}\big[\mathcal{Y}-\mathcal{M}\big]\big)\Big) + \sum_{i_1,\cdots,i_N}^{I_1,\cdots,I_N} p^{-1}_{i_1,\cdots,i_N} = \parallel \mathcal{Y} - \mathcal{M} \parallel_F^2 + \sum_{i_1,\cdots,i_N}^{I_1,\cdots,I_N} p^{-1}_{i_1,\cdots,i_N}$, where $\mathcal{M}$ is a tensor with its $(i_1,\cdots,i_N)^{th}$ element being $m_{i_1,\cdots,i_N}$. On the other hand, using **Property 2**, we have $\mathfrak{w}_2 = \mathfrak{D}\Big[\overset{N}{\underset{n=P+1}{\odot}}\Big(\mathbf{M}^{(n)H}\mathbf{M}^{(n)} + I_n\boldsymbol{\Sigma}^{(n)}\Big)^*\Big]$. Substituting these two results into (51) at the top of the next page, we have equation (32).

REFERENCES

[1] L. D. Lathauwer, "A short introduction to tensor-based methods for factor analysis and blind source separation," in *Proc. of the IEEE Int. Symp. on Image and Signal Processing and Analysis (ISPA 2011)*, Dubrovnik, Croatia, Sep. 4-6, 2011, pp. 558-563.

[2] J. F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no.1, pp. 157-192, 1999.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2603969, IEEE Transactions on Signal Processing

$$Q\left(\boldsymbol{\Xi}^{(k)}\right) \propto \exp\Bigg\{-\mathrm{Tr}\Bigg(\boldsymbol{\Xi}^{(k)}\Big[\mathbb{E}_{Q(\beta)}\left[\beta\right]\underbrace{\mathbb{E}_{\prod_{n=1,n\neq k}^{N}Q(\boldsymbol{\Xi}^{(n)})}\Big[\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\boldsymbol{\Xi}^{(n)}\big)^{T}\big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\boldsymbol{\Xi}^{(n)}\big)^{*}\Big]+\mathbb{E}_{\prod_{l=1}^{L}Q(\gamma_l)}\left[\boldsymbol{\Gamma}\right]}_{\triangleq\left[\boldsymbol{\Sigma}^{(k)}\right]^{-1}}\Big]\boldsymbol{\Xi}^{(k)H}$$
$$-\boldsymbol{\Xi}^{(k)}\left[\boldsymbol{\Sigma}^{(k)}\right]^{-1}\underbrace{\Big[\mathbb{E}_{Q(\beta)}\left[\beta\right]\Big(\mathfrak{U}^{(k)}\Big[\mathcal{Y}-\mathbb{E}_{Q(\mathcal{E})}\left[\mathcal{E}\right]\Big]\Big)\Big(\overset{N}{\underset{n=1,n\neq k}{\diamond}}\mathbb{E}_{Q(\boldsymbol{\Xi}^{(n)})}\Big[\boldsymbol{\Xi}^{(n)}\Big]\Big)^{*}\boldsymbol{\Sigma}^{(k)}}_{\triangleq\mathbf{M}^{(k)}}\Big]^{H}$$
$$-\mathbf{M}^{(k)}\left[\boldsymbol{\Sigma}^{(k)}\right]^{-1}\boldsymbol{\Xi}^{(k)H}\Bigg)\Bigg\}. \tag{44}$$

$$\begin{aligned}\tilde{f}=&\mathrm{Tr}\Bigg(\underbrace{\mathbb{E}_{Q(\mathcal{E})}\Big[\Big(\mathfrak{U}^{(1)}\Big[\mathcal{Y}-\mathcal{E}\Big]\Big)^{H}\mathfrak{U}^{(1)}\Big[\mathcal{Y}-\mathcal{E}\Big]\Big]}_{\triangleq\mathfrak{w}_1}\\&+\mathbb{E}_{Q(\boldsymbol{\Xi}^{(1)})}\Big[\boldsymbol{\Xi}^{(1)}\Big]\underbrace{\mathbb{E}_{\prod_{n=2}^{N}Q(\boldsymbol{\Xi}^{(n)})}\Big[\big(\overset{N}{\underset{n=2}{\diamond}}\boldsymbol{\Xi}^{(n)}\big)^{T}\big(\overset{N}{\underset{n=2}{\diamond}}\boldsymbol{\Xi}^{(n)}\big)^{*}\Big]}_{\triangleq\mathfrak{w}_2}\mathbb{E}_{Q(\boldsymbol{\Xi}^{(1)})}\Big[\boldsymbol{\Xi}^{(1)}\Big]^{H}\\&-\mathbb{E}_{Q(\boldsymbol{\Xi}^{(1)})}\Big[\boldsymbol{\Xi}^{(1)}\Big]\Big(\overset{N}{\underset{n=2}{\diamond}}\mathbb{E}_{Q(\boldsymbol{\Xi}^{(n)})}\Big[\boldsymbol{\Xi}^{(n)}\Big]\Big)^{T}\Big(\mathfrak{U}^{(1)}\Big[\mathcal{Y}-\mathbb{E}_{Q(\mathcal{E})}\left[\mathcal{E}\right]\Big]\Big)^{H}\\&-\mathfrak{U}^{(1)}\Big[\mathcal{Y}-\mathbb{E}_{Q(\mathcal{E})}\left[\mathcal{E}\right]\Big]\Big(\overset{N}{\underset{n=2}{\diamond}}\mathbb{E}_{Q(\boldsymbol{\Xi}^{(n)})}\Big[\boldsymbol{\Xi}^{(n)}\Big]\Big)^{*}\mathbb{E}_{Q(\boldsymbol{\Xi}^{(1)})}\Big[\boldsymbol{\Xi}^{(1)}\Big]^{H}\Bigg).\end{aligned} \tag{51}$$

[3] C. F. Beckmann and S. M. Smith, "Tensorial extensions of independent component analysis for multisubject FMRI analysis," *Neuroimage*, vol. 25, no.1, pp. 294-31, 2005.

[4] L. De Lathauwer, "Algebraic methods after prewhitening," in *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, Academic Press, New York, 2010, pp. 155-177.

[5] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Process.*, vol. 48, no.3, pp. 810-823, 2000.

[6] C. A. R. Fernandes, A. L. F. de Almeida, and D. B. da Costa, "Unified tensor modeling for blind receivers in multiuser uplink cooperative systems," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 247-250, May 2012.

[7] A. Y. Kibangou and A. De Almeida, "Distributed PARAFAC based DS-CDMA blind receiver for wireless sensor networks", in *Proc. of the IEEE Int. Conference on Signal Processing Advances in Wireless Communications (SPAWC 2010)*, Marrakech, Jun. 20-23, 2010, pp. 1-5.

[8] A. L. F. de Almeida, A. Y. Kibangou, S. Miron and D. C. Araujo, "Joint data and connection topology recovery in collaborative wireless sensor networks," in *Proc. of the IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, May 26-31, 2013, pp. 5303-5307.

[9] M. Sorensen, L. D. Lathauwer, L. Deneire, "PARAFAC with orthogonality in one mode and applications in DS-CDMA systems", in *Proc. of the IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP 2010)*, Dallas, Texas, Mar. 2010, pp. 4142-4145.

[10] D. Nion and N. D. Sidiropoulos, "Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5693-5705, 2010.

[11] W. Sun, H. C. So and F. K. W. Chan and L. Huang, "Tensor approach for eigenvector-based multi-dimensional harmonic retrieval," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3378-3388, 2013.

[12] B. Pesquet-Popescu, J.-C. Pesquet and A. P. Petropulu, "Joint singular value decomposition - A new tool for separable representation of images," in *Proc. of the IEEE Int. Conference on Image Processing (ICIP 2001)*, Thessaloniki, Greece, Oct. 2001, pp. 569-572.

[13] A. Shashua and A. Levin, "Linear image coding for regression and classification using the tensor-rank principle," in *Proc. of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, Kauai, Hawaii, Dec. 2001, pp. 42-49.

[14] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, pp. 455-500, 2009.

[15] J. D. Carroll and J. J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, pp. 283-319, 1970.

[16] De Silva and L. H. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM J. Matrix Anal. Appl.*, vol. 30, pp. 1084-1127, 2008.

[17] M. Sorensen, L. D. Lathauwer, P. Comon, S. Icart, and L. Deneire, "Canonical polyadic decomposition with a columnwise orthonormal factor matrix," *SIAM J. Matrix Anal. Appl.*, vol.33, no.4, pp. 1190-1213, 2012.

[18] M. Muma, Y. Cheng, F. Roemer, M. Haardt, and A. M. Zoubir, "Robust source number enumeration for R-dimensional arrays in case of brief sensor failures," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2012)*, Kyoto, Mar. 2012, pp. 3709-3712.

[19] Z. Lu, Y. Chakhchoukh, and A. M. Zoubir, "Source number estimation in impulsive noise environments using bootstrap techniques and robust statistics," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, Signal Processing (ICASSP)*, Prague, Czech Republic, May 22-27, 2011, pp. 2712-2715.

[20] R. H. Chan, C. W. Ho, and M. Nikolova, "Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization," *IEEE Trans. Image Process.* vol.14, no. 10, pp. 1479-1485, 2005.

[21] D. J. C. MacKay, "Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 469-505, 1995.

[22] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B*, vol. 61, no. 3, pp. 611-622, 1999.

[23] C. M. Bishop, "Bayesian principal component analysis," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 382-388, 1999.

[24] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol.29, no.12, pp. 3419-3430, 2011.

[25] L. Xiong, X. Chen, T. K. Huang, J. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with Bayesian probabilistic tensor factorization," in *Proc. of the SIAM Int. Conference on Data Mining*, Columbus, May, 2010, pp. 211-222.

[26] I. Porteous, E. Bart, and M. Welling, "Multi-HDP: A non-parametric Bayesian model for tensor factorization," in *the AAAI Conference on Artificial Intelligence (AAAI 2008)*, Chicago, July, 2008, pp. 1487-1490.

[27] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[28] P. C. Hansen, *Rank Deficient and Discrete Ill-Posed Problems-Numerical Aspects of Linear Inversion*. Philadelphia, PA: SIAM, 1998.

[29] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol.1, no. 102, pp. 1-305, Jan. 2008.

[30] B. Ermis and A. T. Cemgil, "A Bayesian tensor factorization model via variational inference for link prediction", submitted on 29 Sep 2014. Online Available: http://arxiv.org/pdf/1409.8276v1.pdf.

[31] Z. Xu, F. Yan and Y.Qi, "Bayesian nonparametric models for multiway data analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.37, no.2, pp. 475-487, Nov. 2013.

[32] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian CP factorization of incomplete tensors with automatic rank determination," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no.9, pp. 1751-1753, Sep. 2015.

[33] C. G. Khat and K. V. Mardia "The von Mises-Fisher distribution in orientation statistics," *Journal of Royal Statistical Society*, vol. 39, pp. 95-106, 1977.

[34] M. West, "On scale mixtures of normal distributions," *Biometrika*, vol.74, no.3, pp.646-648, 1987.

[35] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions* , CRC Press,1999.

[36] S. Liu and G. Trenkler, "Hadamard, Khatri-Rao, Kronecker and other matrix products" *Int. J. Inform. Syst. Sci*, vol. 4, no.1, pp.160-177, 2008.

[37] T. Kanamori and A. Takeda, "Non-convex optimization on Stiefel manifold and applications to machine learning," *Neural Information Processing*, pp.109-16, 2012.

[38] A. Georghiades, D. Kriegman, and P. Belhumeur, "From few to many: Generative models for recognition under variable pose and illumination," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 643-660, 2001.

[39] M. J. Beal, "Variational algorithms for approximate Bayesian inference," PhD thesis, Gatsby Computational Neuroscience Unit, University College, London, 2003.

[40] M. Srensen, D. Ignat, and L. D. Lieven. "Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(L_r, n, L_r, n, 1)$ terms—Part II: Algorithms," *SIAM Journal on Matrix Analysis and Applications*, vol. 36, no. 3, pp. 1015-1045, Mar. 2015.

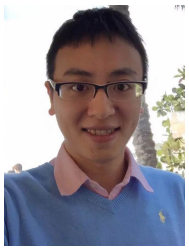
[41] X. Luciani, and L. Albera. "Canonical polyadic decomposition based on joint eigenvalue decomposition," *Chemometrics and Intelligent Lab. Systems*, vol. 132, pp. 152-167, Mar. 2014.

[42] X. Fu, K. Huang, W.-K. Ma, N.D. Sidiropoulos, and R. Bro, "Joint tensor factorization and outlying slab suppression with applications," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6315-6328, Dec. 1, 2015.

[43] H. Huang and C. H. Q. Ding, "Robust tensor factorization using r1 norm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Anchorage, AK, 2008, pp. 1-8.

[44] D. Goldfarb and Z. Qin, "Robust low-rank tensor recovery: Models and algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 35, no. 1, pp. 225-253, 2014.

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



Lei Cheng received the B.Eng. degree from Zhejiang University (ZJU) in 2013. He is currently working towards the Ph.D. degree at the University of Hong Kong (HKU). His research interests are in general areas of signal processing and machine learning, and in particular statistical inference for multidimensional data.

Yik-Chung Wu received the B.Eng. (EEE) degree in 1998 and the M.Phil. degree in 2001 from the University of Hong Kong (HKU). He received the Croucher Foundation scholarship in 2002 to study Ph.D. degree at Texas A&M University, College Station, and graduated in 2005. From August 2005 to August 2006, he was with the Thomson Corporate Research, Princeton, NJ, as a Member of Technical Staff. Since September 2006, he has been with HKU, currently as an Associate Professor. He was a visiting scholar at Princeton University, in summers of 2011 and 2015. His research interests are in general areas of signal processing, machine learning and communication systems, and in particular distributed signal processing and robust optimization theories with applications to communication systems and smart grid. Dr. Wu served as an Editor for IEEE Communications Letters, is currently an Editor for IEEE Transactions on Communications and Journal of Communications and Networks.

H. Vincent Poor (S'72, M'77, SM'82, F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor of Electrical Engineering. From 2006 till 2016, he served as Dean of Princeton's School of Engineering and Applied Science. Dr. Poor's research interests are in the areas of statistical signal processing, stochastic analysis and information theory, and their applications in wireless networks and related fields. Among his publications in these areas is the recent book *Mechanisms and Games for Dynamic Spectrum Allocation* (Cambridge University Press, 2014).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and a foreign member of the Royal Society. He is also a Fellow of the American Academy of Arts and Sciences and the National Academy of Inventors, and of other national and international academies. He received the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011, respectively. Recent recognition of his work includes the 2014 URSI Booker Gold Medal, the 2015 EURASIP Athanasios Papoulis Award, the 2016 John Fritz Medal, and honorary doctorates from Aalborg University, Aalto University, HKUST and the University of Edinburgh.