Non-negative Multiple Tensor Factorization

Koh Takeuchi*, Ryota Tomioka†, Katsuhiko Ishiguro*, Akisato Kimura*, and Hiroshi Sawada*
* NTT Communication Science Laboratories, Kyoto, Japan
{takeuchi.koh, ishiguro.katsuhiko, hiroshi.sawada}@lab.ntt.co.jp, akisato@ieee.org
†Toyota Technological Institute at Chicago, Chicago, Illinois, USA
tomioka@ttic.edu

***Abstract*—Non-negative Tensor Factorization (NTF) is a widely used technique for decomposing a non-negative value tensor into sparse and reasonably interpretable factors. However, NTF performs poorly when the tensor is extremely sparse, which is often the case with real-world data and higher-order tensors. In this paper, we propose Non-negative Multiple Tensor Factorization (NMTF), which factorizes the target tensor and auxiliary tensors simultaneously. Auxiliary data tensors compensate for the sparseness of the target data tensor. The factors of the auxiliary tensors also allow us to examine the target data from several different aspects. We experimentally confirm that NMTF performs better than NTF in terms of reconstructing the given data. Furthermore, we demonstrate that the proposed NMTF can successfully extract spatio-temporal patterns of people's daily life such as leisure, drinking, and shopping activity by analyzing several tensors extracted from online review data sets.**

I. Introduction

As the amount and variety of available data has grown rapidly in recent years, there are many tensors consisting of only non-negative data: e.g. ratings of users $\times$ shops $\times$ times, and social connections between users $\times$ friends $\times$ social networking systems. To analysing such tensor data, we can employ several techniques such as Non-negative Tensor Factorization (NTF) [1] , which is a generalization of Non-negative Matrix Factorization (NMF) [2]. NTF factorizes a target tensor consisting of non-negative values into factor matrices under the non-negative constraints. The non-negativity constraints yield sparse and reasonably interpretable factorization results [3]. NTF has been applied and preforms well in various fields [4]–[7]. An advantage of tensor data over conventionally studied matrix data is its ability to represent observations with various attributes. In fact, tensor factorization techniques have been extended to collaborative filtering [8] and multi relational networks [9] and have been proven effective in these problems.

In spite of these advantages, data sparseness and computational costs have significantly hampered the application of tensor factorization methods to real world problems. This problem becomes more serious as the order of the tensor becomes higher and degrades the predictive performance of NTF. The problem of sparseness has also been a serious problem in matrix factorization [10]. To deal with this problem, only a few solutions have been proposed in the context of multiple data analysis [11], which have been developed based on probabilistic matrix factorization [12]. To compensate for the sparseness of the data matrix, these approaches augment the target matrix by incorporating auxiliary matrices in the analysis. Then the target matrix and the auxiliary matrices are simultaneously factorized. Factorizing multiple matrices simultaneously performs better than factorizing a single target matrix. Employing auxiliary data also seems a promising way to resolve the sparse issue for tensors. However, there have been almost no attempts to validate the idea of factorizing multiple tensors simultaneously. The only exceptions are described in [13], [14], but we cannot apply those methods to multimodal data sets including tensors of different sizes and scales: this is very typical in commercial purchasing records and social network data.

In this paper, we propose a novel tensor factorization method called Non-negative Multiple Tensor Factorization (NMTF), which naturally incorporates auxiliary data tensors into standard tensor factorization. The auxiliary tensor shares indices (axes) with the target tensors, thus NMTF is able to merge information from auxiliary tensors and mitigates the problem of sparseness. Furthermore, factorizing auxiliary tensors simultaneously allows us to examine given data from several different aspects, because we obtain factors from auxiliary tensors as well. We can control the influence of auxiliary tensors during the factorization by employing scaling parameters. We show that the simultaneous tensor decomposition approach can be reformulated into the decomposition of a larger (partially observed) tensor. NMTF is a generalization of NTF and is reduced to NTF in the special case where all scaling parameters are set at zero (no influence). In this paper, we employ the generalized Kullback-Leibler (gKL) divergence as the metric of NMTF, which is widely used in the context of NMF/NTF [15]. Minimizing gKL divergence is known to be equivalent to maximizing the Poisson distribution likelihood, which fits particularly well if the data are discrete values. We employ a three-way tensor in this paper but the extension of NMTF to a N-way tensor is straightforward.

Empirical results showed that NMTF achieved better performance than NTF in a quantitative way. The synthetic data experiments revealed that the performance improvement of NMTF compared with NTF become larger as the target tensor become sparser. We also found that NMTF factorization results are highly interpretable and suggestive for the analysis of real world complex review data. NMTF reveals the detailed user preferences of beers and spatio-temporal patterns of people's daily life such as leisure, drinking, and shopping activity from online review data sets.

II. Related Work

Non-negative Tensor Factorization (NTF) was first proposed in [1], as a generalization of Non-negative Matrix factorization (NMF) [2]. NTF is based on a CANDECOMP/PARAFAC (CP) decomposition [16] and imposes non-negative constraints on tensor and factor matrices. There have

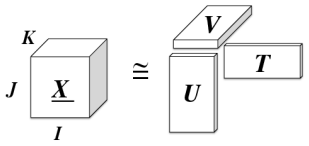


Fig. 1. NTF factorizes a tensor $\underline{\boldsymbol{X}}$ into factor matrices $\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V}$

been a lot of NTF researches concerning sparse constraints [6], [17] and acceleration techniques [7], [18]. As explained, data sparsity becomes a serious problem for large or high order tensors. However, there have been no study concerning the data sparsity in NTF.

As we explained in the introduction, combining multiple data matrices has proved effective in sparse matrix analysis [11], [19]–[21]. Therefore combining multiple tensors is a reasonable idea for solving the sparsity problem. A concatenating method for a collection of tensors is proposed in [13]. A decomposing method for multiple tensors of exactly the same size simultaneously is proposed in [14]. To the best of our knowledge, there are currently no solutions to our sparse data tensor problem.

III. Non-negative Tensor Factorization (NTF)

Let us denote a three-way tensor[1] $\underline{\boldsymbol{X}} = \{x_{ijk}\} \in \mathbb{R}^{I \times J \times K}$ with only non-negative values $x_{ijk} \geq 0, \forall\ i, j, k$. Figure 1 illustrates a concept of NTF. NTF factorizes a tensor $\underline{\boldsymbol{X}}$ into three matrices each of which consists of R factors and only contains non-negative values. We denote the three matrices as $\boldsymbol{T} = \{t_{ir}\} \in \mathbb{R}^{I \times R}$, $\boldsymbol{U} = \{u_{jr}\} \in \mathbb{R}^{J \times R}$ and $\boldsymbol{V} = \{v_{kr}\} \in \mathbb{R}^{K \times R}$. The r-th column vectors of each matrix correspond to the r-th factor of the tensor.

Let us denote the reconstructed tensors of $\underline{\boldsymbol{X}}$ with $\boldsymbol{T}$, $\boldsymbol{U}$, $\boldsymbol{V}$ as $\underline{\hat{\boldsymbol{X}}} = \{\hat{x}_{ijk}\} \in \mathbb{R}^{I \times J \times K}$. We define the elements of $\underline{\hat{\boldsymbol{X}}}$ as the sum of the linear products of the three matrices.

$$\hat{x}_{ijk} = \sum_{r=1}^{R} t_{ir} u_{jr} v_{kr}. \tag{1}$$

Let us denote a divergence between two tensors as $\mathcal{D}$. We employ an element-wise divergence d to measure the divergence between $\underline{\boldsymbol{X}}$ and $\underline{\hat{\boldsymbol{X}}}$. $\mathcal{D}(\underline{\boldsymbol{X}}|\underline{\hat{\boldsymbol{X}}})$ is written as:

$$\mathcal{D}(\underline{\boldsymbol{X}}|\underline{\hat{\boldsymbol{X}}}; \Theta) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} d(x_{ijk}|\hat{x}_{ijk}), \tag{2}$$

where $\Theta \triangleq \{U,\ T,\ V\}$. Using the above notations, NTF is formulated as follows:

$$\min_{T,U,V} \mathcal{D}(\underline{\boldsymbol{X}}|\underline{\hat{\boldsymbol{X}}}; \Theta) \text{ subject to } \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V} \geq 0. \tag{3}$$

A. Generalized Kullback-Leibler Divergence

There are a number of candidate divergences for NTF including Euclidean divergence, Itakura-Saito divergence and generalized Kullback-Leibler divergence. Although there is no definitive way of choosing the divergence. To this end, we are interested in discrete value observations such as stars in product reviews. For that purpose, we choose the generalized Kullback-Leibler (gKL) divergence, which is formulated as follows: $d(p|q) = -p \log q + q + p \log p - p$. It is known that minimizing the gKL divergence is equivalent to maximizing the log likelihood of the NTF model if we assume the observations are Poisson distributed, which is a natural choice for a discrete value observation [3]. Please note that our algorithm can be derived for other divergences in an analogous manner.

[1]In this paper, we denote a tensor by a bold-face underlined capital letter, a matrix by a bold-face capital letter, a vector by a bold letter and a scalar by a plain letter.

IV. Proposed Method

A. Non-negative Multiple Tensor Factorization (NMTF)

we propose Non-negative Multiple Tensor Factorization (NMTF), which effectively combines multiple data tensors under a non-negative constraint. NMTF eases the problem of tensor sparsity and further allows us to examine given data from several different aspects. In NMTF, we have a target tensor and a few auxiliary tensors that share one or two indices (axes) with the target tensor. We want to factorize these tensors simultaneously in order to make use of available auxiliary information.

Let us denote the target tensor as $\underline{\boldsymbol{Y}} \in \mathbb{R}^{I_y \times J_y \times K_y}$. We denote the n-th auxiliary tensors as $\underline{\boldsymbol{A}}^{(n)} \in \mathbb{R}^{I_n \times J_n \times K_n}$ $(n = 1, \cdots, N)$. We set $N = 3$ in this paper for the convenience. We define the three factor matrices of $\underline{\boldsymbol{Y}}$ as $\Theta_y = \{\boldsymbol{T}_y,\ \boldsymbol{U}_y,\ \boldsymbol{V}_y\}$ where $\boldsymbol{T}_y \in \mathbb{R}^{I_y \times R}$, $\boldsymbol{U}_y \in \mathbb{R}^{J_y \times R}$, and $\boldsymbol{V}_y \in \mathbb{R}^{K_y \times R}$. Let us define the factor matrices of the n-th auxiliary tensor as $\Theta^{(n)} = \{\boldsymbol{T}^{(n)},\ \boldsymbol{U}^{(n)}, \boldsymbol{V}^{(n)}\}$ where $\boldsymbol{T}^{(n)} \in \mathbb{R}^{I_n \times R}$, $\boldsymbol{U}^{(n)} \in \mathbb{R}^{J_n \times R}$, and $\boldsymbol{V}^{(n)} \in \mathbb{R}^{K_n \times R}$. We define estimated (reconstructed) tensor values based on factor matrix elements as follows:

$$\hat{y}_{ijk} = \sum_{r=1}^{R} t^{y}_{ir} u^{y}_{jr} v^{y}_{kr},\ \hat{a}^{(n)}_{ijk} = \sum_{r=1}^{R} t^{(n)}_{ir} u^{(n)}_{jr} v^{(n)}_{kr}. \tag{4}$$

For simplicity, we assume that $\underline{\boldsymbol{Y}}$ shares first indices with $\underline{\boldsymbol{A}}^{(1)}$, second indices with $\underline{\boldsymbol{A}}^{(2)}$, and third indices with $\underline{\boldsymbol{A}}^{(3)}$, respectively. Note that it is straightforward to extend our results to cases where auxiliary tensors share two indices with the target tensor. According to this assumption, we set $I_1 = I_y,\ J_2 = J_y$ and $I_3 = K_y$. We set $\boldsymbol{T}^{(1)} = \boldsymbol{T}_y,\ \boldsymbol{U}^{(2)} = \boldsymbol{U}_y$ and $\boldsymbol{V}^{(3)} = \boldsymbol{V}_y$.

The goal of our NMTF is to factorize the target tensor $\underline{\boldsymbol{Y}}$ with the auxiliary tensors $\underline{\boldsymbol{A}}^{(n)}$ $(n = 1, \cdots, N)$ simultaneously. To achieve this goal, we regard tensors $\underline{\boldsymbol{Y}}$ and $\underline{\boldsymbol{A}}^{(n)}$ $(n = 1, \cdots, N)$ as parts of a larger tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{(I_y+I_1+I_2) \times (J_1+J_y+J_3) \times (K_y+K_2+K_1)}$, and perform NTF on this larger tensor (see Fig. 2).

We augment the target tensor $\underline{\boldsymbol{Y}}$ into $\underline{\boldsymbol{X}}$ as follows. First, three axes are augmented by simply concatenating the indices of the target tensor and auxiliary tensors. For example, the size of the first axis is enlarged from I to $I_y + I_2 + I_3$. Given a large empty tensor $\underline{\boldsymbol{X}}$, we place data tensors so as to match all indices as in Fig. 2. Please recall that $I_y = I_1$, $J_y = J_2$, and $K_y = K_3$. As evident from Fig. 2, there are several blocks with no observation, called undefined regions. To encompass

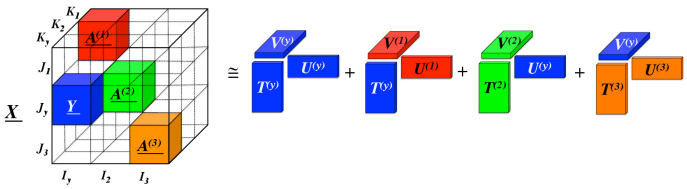


Fig. 2. $\underline{\boldsymbol{Y}}$, $\underline{\boldsymbol{A}}^{(1)}$, $\underline{\boldsymbol{A}}^{(2)}$ and $\underline{\boldsymbol{A}}^{(3)}$ are included in $\underline{\boldsymbol{X}}$. Transparent elements of $\underline{\boldsymbol{X}}$ are undefined regions. Factor matrices $\boldsymbol{U}_y, \boldsymbol{T}_y, \boldsymbol{V}_y$ are shared among the target tensor and additional tensors.

these regions, we introduce the following binary variable ω_{ijk}:

$$\omega_{ijk} = \begin{cases} 1 & x_{ijk} \text{ is defined,} \\ 0 & x_{ijk} \text{ is undefined.} \end{cases} \tag{5}$$

Let $\underline{\boldsymbol{\Omega}} \in \mathbb{R}^{I \times J \times K}$ be a tensor consisting of ω_{ijk}. We set undefined regions in $\underline{\boldsymbol{X}}$ using $\underline{\boldsymbol{\Omega}}$. Let us denote sets Ω_y and $\Omega^{(n)}$ $(n = 1, \cdots, N)$ consisting of indices of $\underline{\boldsymbol{Y}}$ and $\underline{\boldsymbol{A}}^{(n)}$ $(n = 1, \cdots, N)$.

$$\omega_{ijk} = \begin{cases} 1 \text{ if } \{i,j,k\} \in \Omega_y \\ \hat{\eta}^{(n)} \text{ if } \{i,j,k\} \in \Omega^{(n)} \ (n = 1, \cdots, N) \\ 0 \text{ otherwise.} \end{cases} \tag{6}$$

where $\eta^{(n)} \geq 0$ $(n = 1, \cdots, N)$ are scaling parameters of $\underline{\boldsymbol{A}}^{(n)}$ $(n = 1, \cdots N)$. The elements ω_{ijk} are set at 0 if the indices do not indicate the observed values in $\underline{\boldsymbol{Y}}$ or $\underline{\boldsymbol{A}}^{(n)}$ $(n = 1, \cdots, N)$. Let us denote $\hat{\mathcal{D}}$ as the divergence of Non-negative Tensor Factorization with an undefined region.

$$\hat{\mathcal{D}}(\underline{\boldsymbol{X}} | \hat{\underline{\boldsymbol{X}}}; \underline{\boldsymbol{\Omega}}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \omega_{ijk} d(x_{ijk} | \hat{x}_{ijk}) \tag{7}$$

The reconstructed values of the elements $\hat{x}_{ijk}$ are the same as Eq. 1. Then we set T, U, and V as,

$$T = \begin{pmatrix} T_y \\ T^{(2)} \\ T^{(3)} \end{pmatrix}, \ U = \begin{pmatrix} U^{(1)} \\ U_y \\ U^{(3)} \end{pmatrix}, \ V = \begin{pmatrix} V_y \\ V^{(2)} \\ V^{(1)} \end{pmatrix}. \tag{8}$$

Therefore $\hat{\mathcal{D}}$ is rewritten as,

$$\begin{aligned} &\hat{\mathcal{D}}(\underline{\boldsymbol{X}} | \hat{\underline{\boldsymbol{X}}}; \underline{\boldsymbol{\Omega}}) \\ &= \sum_{i=1}^{I_y} \sum_{j=1}^{J_y} \sum_{k=1}^{K_y} d(y_{ijk} | \hat{y}_{ijk}) + \sum_{n=1}^{N} \left[\sum_{i=1}^{I_n} \sum_{j=1}^{J_n} \sum_{k=1}^{K_n} \eta^{(n)} d(a^{(n)}_{ijk} | \hat{a}^{(n)}_{ijk}) \right] \\ &= \mathcal{D}(\underline{\boldsymbol{Y}} | \hat{\underline{\boldsymbol{Y}}}; \Theta_y) + \sum_{n=1}^{N} \mathcal{D}(\eta^{(n)} \underline{\boldsymbol{A}}^{(n)} | \eta^{(n)} \hat{\underline{\boldsymbol{A}}}^{(n)}; \Theta^{(n)}) \\ &= \mathcal{D}(\underline{\boldsymbol{Y}}, \eta^{(n)} \underline{\boldsymbol{A}}^{(n)} | \hat{\underline{\boldsymbol{Y}}}, \ \eta^{(n)} \hat{\underline{\boldsymbol{A}}}^{(n)}; \hat{\Theta}, n = 1, \cdots, N), \end{aligned} \tag{9}$$

where we set factorized matrices as $\hat{\Theta} \triangleq \{\Theta_y, \Theta^{(n)}, \ n = 1 \ldots, N\}$. NMTF minimizes the divergence between the given non-negative tensors $\{\underline{\boldsymbol{Y}}, \ \underline{\boldsymbol{A}}^{(n)}; n = 1, \cdots, N\}$ and those estimated under the non-negative constraints.

$$\begin{aligned} &\min_{\Theta} \mathcal{D}(\underline{\boldsymbol{Y}}, \eta^{(n)} \underline{\boldsymbol{A}}^{(n)} | \hat{\underline{\boldsymbol{Y}}}, \ \eta^{(n)} \hat{\underline{\boldsymbol{A}}}^{(n)}; \hat{\Theta}, n = 1, \cdots, N) \\ &\text{subject to } \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V} \geq 0. \end{aligned} \tag{10}$$

Note that if the scaling parameters $\eta^{(1)} = \eta^{(2)} = \eta^{(3)} = 0$, NMTF is reduced to the original NTF. And if the non-negative constraints are removed, we can regard this model as a multiple tensor version of CP factorization.

B. Multiplicative Update Rules

In this section, we derive multiplicative update rules for NMTF, similar to those of NTF. Remember that we set $t^{(1)}_{ir} = t^{y}_{ir}$, $u^{(2)}_{jr} = u^{y}_{jr}$ and $v^{(3)}_{kr} = v^{y}_{kr}$. We derive multiplicative update rules for t^{y}_{ir}, u^{y}_{jr}, and v^{y}_{kr}, as below:

$$\begin{aligned} &t^{y}_{ir}{}^{(\text{new})} = \\ &t^{y}_{ir} \frac{\sum_{j=1}^{J_y} \sum_{k=1}^{K_y} \left[\frac{y_{ijk}}{\hat{y}_{ijk}} u^{y}_{jr} v^{y}_{kr} \right] + \eta^{(1)} \sum_{j=1}^{J_1} \sum_{k=1}^{K_1} \left[\frac{a^{(1)}_{ijk}}{\hat{a}^{(1)}_{ijk}} u^{(1)}_{jr} v^{(1)}_{kr} \right]}{\sum_{j=1}^{J_y} \sum_{k=1}^{K_y} u^{y}_{jr} v^{y}_{kr} + \eta^{(1)} \sum_{j=1}^{J_1} \sum_{k=1}^{K_1} u^{(1)}_{jr} v^{(1)}_{kr}}, \\ &u^{y}_{jr}{}^{(\text{new})} = \\ &u^{y}_{jr} \frac{\sum_{i=1}^{I_y} \sum_{k=1}^{K_y} \left[\frac{y_{ijk}}{\hat{y}_{ijk}} t^{y}_{ir} v^{y}_{kr} \right] + \eta^{(2)} \sum_{i=1}^{I_2} \sum_{k=1}^{K_2} \left[\frac{a^{(2)}_{ijk}}{\hat{a}^{(2)}_{ijk}} t^{(2)}_{ir} v^{(2)}_{kr} \right]}{\sum_{i=1}^{I_y} \sum_{k=1}^{K_y} t^{y}_{ir} v^{y}_{kr} + \eta^{(2)} \sum_{i=1}^{I_2} \sum_{k=1}^{K_2} t^{(2)}_{ir} v^{(2)}_{kr}}, \\ &v^{y}_{kr}{}^{(\text{new})} = \\ &v^{y}_{kr} \frac{\sum_{i=1}^{I_y} \sum_{j=1}^{J_y} \left[\frac{y_{ijk}}{\hat{y}_{ijk}} t^{y}_{ir} u^{y}_{jr} \right] + \eta^{(3)} \sum_{j=1}^{J_3} \sum_{k=1}^{K_3} \left[\frac{a^{(3)}_{ijk}}{\hat{a}^{(3)}_{ijk}} t^{(3)}_{ir} u^{(3)}_{jr} \right]}{\sum_{i=1}^{I_y} \sum_{j=1}^{J_y} t^{y}_{ir} u^{y}_{jr} + \eta^{(3)} \sum_{i=1}^{I_3} \sum_{j=1}^{J_3} t^{(3)}_{ir} u^{(3)}_{jr}}. \end{aligned} \tag{11}$$

We can derive multiplicative update rules for other parameters $t^{(n)}_{ir}$, $u^{(n)}_{jr}$, and $v^{(n)}_{kr}$ in a similar way. Note that this algorithm is stable. Let us focus on the update of t^{y}_{ir}. The updated t^{y}_{ir} is always non-negative because all elements in the r.h.s. of the update equations are non-negative.

The only concern regarding computational stability is the case where some u and v take the value 0, then the r.h.s. of the equation becomes $\frac{0}{0}$. But this will never happens if we handle the data and the model initialization appropriately. First, ensure that $\sum_{j=1}^{J_y} \sum_{k=1}^{K_y} y_{i,j,k} > 0$ and $\sum_{j=1}^{J_1} \sum_{k=1}^{K_1} a_{i,j,k} > 0$ for all i. This is easily accomplished by excluding the empty index i' that violates these conditions. Second, initialize the model so that $\forall\ \hat{y} > 0$ and $\forall\ \hat{a} > 0$ hold. If $y_{i,j,k} = 0$ or $a_{i,j,k} = 0$, then we can simply skip the summation for the corresponding elements and speeding up computational times. Details of deriving update rules and a proof of convergence at local minima are to be published.

C. NMTF as Probabilistic Generative Model

We mentioned that a generalized Kullback-Leibler divergence is equal to a negative log likelihood of Poisson distribution. NMTF could be interpreted as a probabilistic generative model. Let us denote a Poisson distribution with a parameter ζ as $p(\xi|\zeta) = \text{Poisson}(\xi|\zeta)$. From the definition, a negative log likelihood of the Poisson distribution is approximately equal to a generalized Kullback-Leibler divergence as below: $-\log p(\xi|\zeta) = d_{gKL}(\xi|\zeta) + \text{const.}$ We denote a log likelihood of a tensor $\underline{\boldsymbol{X}}$ as:

$$\log p(\underline{\boldsymbol{X}} | \hat{\underline{\boldsymbol{X}}}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \log p(x_{ijk} | \hat{x}_{ijk}). \tag{12}$$

Finally, the probabilistic generative model of NMTF can be written as:

$$\log p(\underline{\boldsymbol{X}}|\hat{\underline{\boldsymbol{X}}}) = \log p(\underline{\boldsymbol{Y}}|\hat{\underline{\boldsymbol{Y}}}) + \sum_{n=1}^{N} \log p(\hat{\eta}^{(n)}\underline{\boldsymbol{A}}^{(n)}|\hat{\eta}^{(n)}\hat{\underline{\boldsymbol{A}}}^{(n)})$$
$$= -\mathcal{D}(\underline{\boldsymbol{Y}}, \eta^{(n)}\underline{\boldsymbol{A}}^{(n)}|\hat{\underline{\boldsymbol{Y}}},\ \eta^{(n)}\hat{\underline{\boldsymbol{A}}}^{(n)}; \hat{\Theta}, n = 1, \cdots, N) + \text{const.}\ . \quad (13)$$

Eq. (13) indicates that minimizing the divergence in NMTF (Eq. (10)) is approximately equivalent to maximizing the Poisson log likelihood of the probabilistic model.

V. Experiments

A. Evaluation Measure

In our experiments, almost all the tensor elements are equal to zero. Thus, we focus on the predictive log likelihood for the non-zero elements to evaluate the performance of the factorization results. We split the elements of the target tensor for 5-fold cross validation. We employ the average log likelihood of non-zero elements in the test sets. A higher average log likelihood results in the better modeling performance. We define the average log likelihood as: $\frac{1}{M}\sum_{m=1}^{M}\log p(x_m|\theta)$, where M is the number of non-zero elements in the test sets and θ is the estimated parameter of a model.

B. Synthetic Data Experiment

In this experiment we evaluate the performance of NTF and NMTF in terms of the average test log likelihoods on a synthetic data set. Data sets are stochastically sampled from the probabilistic model. We set the sizes of the tensors at $I_y = J_y = K_y = 100$ and $I^{(n)} = J^{(n)} = K^{(n)} = 100$ ($n = 1, \cdots, N$), respectively. The number of factors is set at $R = 5,\ 10,$ and 20. We set the sparseness of the auxiliary tensors at 90% (i.e., 90% of all the tensor elements have zero values). We set the scaling parameters at $\eta^{(1)} = \eta^{(2)} = \eta^{(3)} = 1$.

In the first experiment, we evaluated the model performance on the sparseness of the target tensor. We examined cases where the sparseness of the target tensor was $90\%,\ 99\%,\ 99.9\%$ and 99.99%. In the second experiment, we also evaluated the effect of using auxiliary tensors. We examined NMTF with different numbers of auxiliary tensors, N. The scaling parameters were determined by cross-validation. The sparseness of $\underline{\boldsymbol{Y}}$ is set at 99%. Note that NMTF reduces to NTF when no auxiliary tensors are available ($N = 0$).

Table I shows the results of the first experiment evaluating the effect of different data sparseness. The numbers of bases R was set at $5, 10,$ and 20, respectively. The average log likelihoods for test sets worsens as the data sparseness increases. We confirmed that NMTF significantly outperforms the original NTF especially when the target tensor is extremely sparse. Table II also shows the results of the second experiment evaluating the effects of the number of available auxiliary tensors. It is evident that the average log likelihoods improve as the number of auxiliary tensors increase. This improvement indicates that NMTF successfully integrates the auxiliary tensor and target tensor information into factors.

TABLE I. Synthetic data experiment : The average test log likelihoods for different sparseness of the target tensor $\underline{\boldsymbol{Y}}$.

Sparseness	NTF	NMTF
90%	-1.72 ± 0.01	-1.97 ± 0.02
99%	-4.16 ± 0.28	-4.44 ± 0.19
99.9%	-56.65 ± 49.61	$\mathbf{-6.12 \pm 0.59}$
99.99%	-273.65 ± 237.86	$\mathbf{-8.18 \pm 3.80}$

(a) $R = 5$

Sparseness	NTF	NMTF
90%	-3.23 ± 0.05	-3.48 ± 0.06
99%	-8.74 ± 0.38	-8.22 ± 0.22
99.9%	-92.34 ± 23.32	$\mathbf{-12.53 \pm 0.75}$
99.99%	-628.25 ± 514.51	$\mathbf{-13.61 \pm 5.24}$

(b) $R = 10$

Sparseness	NTF	NMTF
90%	-6.20 ± 0.05	-6.33 ± 0.06
99%	-18.82 ± 0.48	$\mathbf{-14.65 \pm 0.12}$
99.9%	-250.02 ± 43.43	$\mathbf{-25.44 \pm 1.25}$
99.99%	-628.25 ± 514.51	$\mathbf{-13.00 \pm 4.92}$

(c) $R = 20$

TABLE II. Synthetic data experiment: The average test log likelihoods of the target tensor $\underline{\boldsymbol{Y}}$ for different numbers of auxiliary tensors.

Training set	NMTF ($R = 5$)	NMTF ($R = 10$)	NMTF ($R = 20$)
$\underline{\boldsymbol{Y}}$(= NTF)	-56.65 ± 49.61	-92.34 ± 23.32	-250.02 ± 43.43
$\underline{\boldsymbol{Y}} + \underline{\boldsymbol{A}}^{(1)}$	-44.47 ± 2.91	-77.10 ± 16.64	-168.08 ± 7.16
$\underline{\boldsymbol{Y}} + \sum_n^2 \underline{\boldsymbol{A}}^{(n)}$	-6.49 ± 0.52	-13.17 ± 0.98	-26.08 ± 1.79
$\underline{\boldsymbol{Y}} + \sum_n^3 \underline{\boldsymbol{A}}^{(n)}$	$\mathbf{-6.12 \pm 0.59}$	$\mathbf{-12.53 \pm 0.75}$	$\mathbf{-25.44 \pm 1.25}$

TABLE III. Real-world data sets: The average test log likelihoods on the target tensor $\underline{\boldsymbol{Y}}$.

	Yelp		MovieLens	
R	NTF	NMTF	NTF	NMTF
10	-22.30 ± 0.39	$\mathbf{-17.72 \pm 0.32}$	-23.96 ± 0.09	$\mathbf{-23.74 \pm 0.12}$
20	-30.69 ± 0.69	$\mathbf{-18.38 \pm 0.30}$	-27.16 ± 0.30	$\mathbf{-24.31 \pm 0.12}$
50	-62.97 ± 1.94	$\mathbf{-21.00 \pm 0.37}$	-39.54 ± 0.44	$\mathbf{-26.68 \pm 0.37}$

C. Real Data Experiments

In this section, we evaluate NMTF and NTF with three public data sets provided by Yelp and MovieLens[2]. These data sets include the user's reviews of, for example, places, and movies with various auxiliary data such as time-stamp, geolocation, and check-in counts.

The Yelp data set is a collection of real-world reviews about numerous business places (e.g. restaurants, department stores, etc) in the greater Phoenix area, Arizona. We construct a target tensor $\underline{\boldsymbol{Y}}$ containing $1,228$ users, $1,860$ business places, and 7 days of weeks. An element y_{ijk} represents the user i_j who reviewed the business places j_y on the day k_y. We prepared two auxiliary tensors for the Yelp data set. The first auxiliary tensor $\underline{\boldsymbol{A}}^{(1)}$ consists of users $1,228$ user, 235 business categories, and $92,052$ words. An element $x_{ijk}^{(1)}$ denotes the term-frequency of the word $k^{(1)}$ about the business category j^1 in reviews by user i_y. The second auxiliary tensor $\underline{\boldsymbol{A}}^{(2)}$ contains 63 geolocation grids and $1,860$ business places, and $(24 * 7) = 186$ hours of weeks. We assume an approximately 10km $\times$ 10km geolocation grid. An element $x_{ijk}^{(2)}$ represents the check-ins count of at the business place j_y within the geolocation grid $i^{(2)}$ on the hour-time $k^{(2)}$. We eliminate

[2]http://www.yelp.com, http://www.movielens.org

business places and users whose total numbers of reviews with fewer than 30. $\underline{\boldsymbol{Y}}$ is 99.9963% sparse and $\underline{\boldsymbol{A}}^{(1)}$ is 99.9729% sparse, and $\underline{\boldsymbol{A}}^{(2)}$ is 99.4265% sparse. The numbers of bases are determined at $R = 10, 20$, and 50 in preliminary experiments.

The MovieLens data set is a famous collection of commercial movie reviews. We construct a target tensor $\underline{\boldsymbol{Y}}$ containing $101,970$ movies, $2,113$ users, and 590 weeks. The element y_{ijk} represents the rating user j_y rating the movie i_y posted in the week k_y. An auxiliary tensor $\underline{\boldsymbol{A}}^{(1)}$ consists of $101,970$ movies, 186 locations of movies, and 20 genres. An element $x^{(1)}_{ijk}$ is set at 1 if a location $j^{(1)}$ is included in meta-data about the shooting location of a movie i_y, and also the genre $k^{(1)}$ is tagged to the movie. $\underline{\boldsymbol{Y}}$ is 99.9933% sparse and $\underline{\boldsymbol{A}}^{(1)}$ is 99.9320% sparse. The numbers of bases are set at $R = 10, 20$, and 50 by preliminary experiments.

D. Results

1) Quantitative analysis: Before closely examining the learned bases from the real-world data, we conduct numerical evaluations. The average test log likelihoods for the the Yelp data set and the MovieLens data set are shown in Table III. The scaling parameters are set at $\eta_1 = 0.1, \eta_2 = 0.001$ on the Yelp data set and $\eta_1 = 0.01$ on the MovieLens data set by 5-fold cross-validation. As with the synthetic data experiments, NMTF performed better than the original NTF with real-world sparse tensor data.

2) Qualitative analysis of Yelp data set: We constructed three tensors from the Yelp data set including an auxiliary tensor of geolocations and check-in counts. The number of bases are set at $R = 50$.

Fig. 3 show the extracted factors $\boldsymbol{U}_y$, $\boldsymbol{V}_y$, $\boldsymbol{T}^{(1)}$, $\boldsymbol{V}^{(1)}$, $\boldsymbol{T}^{(2)}$, and $\boldsymbol{V}^{(2)}$ of bases learned from the Yelp data set. The colors in the figures corresponds to the colors of the factors in Fig. 2. The blue bars in the top left show the 10 highest values in $\boldsymbol{U}_y$ with corresponding the names of business places. The green and pink lines in the top right show the estimated check-in counts per hour on each day of the week, namely $\boldsymbol{V}^{(2)}$. The green lines indicate weekday responses while the pink lines indicate those of weekends. The red bars placed in the middle left present the 10 highest values in $\boldsymbol{T}^{(1)}$ with corresponding business categories. The red bars in the middle right show the 10 highest values in $\boldsymbol{V}^{(1)}$ with corresponding words. The blue circles in the bottom left show the locations in Phoenix city, Arizona in I_y, and the circle size indicates the estimated values in $\boldsymbol{U}_y$.

Fig. 3a shows decomposed factors for a specific learned base ($r = 1$). Zoo, museum, and a few parks are selected as representative business places. The estimated check-in counts hit peaks in the mornings on weekends, which is a reasonable pattern for these business places. Top categories and words are also easily interpretable in this base. Moreover, it is easy to see that the selected business places are located across the map. From this evidence, we can imagine that dining and drinking behavior on weekdays can be extracted in this basis.

Fig. 3b shows another extracted basis ($r = 2$). Restaurants, bars and pubs are the top-valued business places. The check-in pattern is vary different from the previous basis: there are many check-in counts on weekday evenings and weekend mornings. The selected categories and words match these business places. The business places are densely located in the central area of Phoenix. We see that the going out and drinking patterns of users on weekdays are mainly extracted in this factor.

Fig. 3c shows yet another basis ($r = 3$). IKEA, Apple store, a toy store, a gardening store, and a pet shop achieve higher values for the place factor. Users checked into these business places most frequently at 5 p.m. on Sundays. The 10 highest categories and words include "shopping", "food", "store", and "IKEA". A place with a very large location weight is situated in the outer area of the city, but many other business places are located in the central downtown area. Our understanding of this basis is that it shows daily purchasing behavior downtown and occasional shopping on the periphery.

Finally, we present an interesting and completely different pattern in Fig. 3d ($r = 4$). The listed business places serve typical Asian foods such as sushi and noodles. The check-in counts do not very greatly between weekdays and weekends. The categories and words are also related to Asian foods such as "Sushi Bar", "Japanese", "roll", and "fish". The patterns of users enjoying these typical foods are extracted in this basis. NMTF revealed expressive patterns with geolocation and check-in counts in this experiment.

VI. Conclusion

In this paper, we proposed a novel tensor factorization technique called Non-negative Multiple Tensor Factorization (NMTF). We formulated NMTF as a generalization of NTF and NMTF includes NTF as a special choice of scaling parameters. We adopted the generalized Kullback-Leibler divergence as the distance metric between tensors and we derived a method for parameter estimation based on the multiplicative update rule. We evaluated NMTF and the original NTF on both synthetic and real-world data sets. The performance of NMTF was quantitatively better than that of NTF. We also confirmed that NMTF successfully extracted informative and understandable factors from multiple tensors.

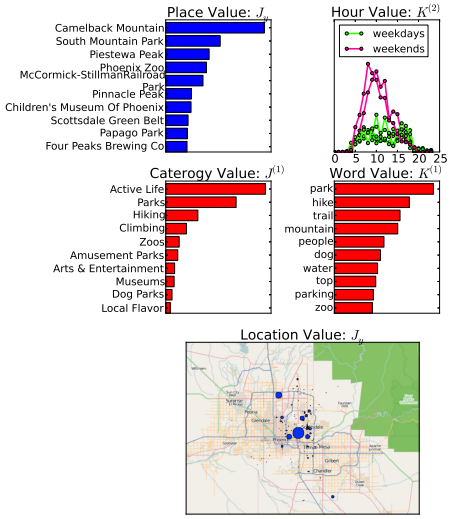
In this paper, we considered a case where each auxiliary tensor shares only one axis with the target tensor. One natural extension of this work is to allow the tensor to the share multiple axes. Decomposing different orders of tensors should allow us to model more complex and rich data set, for example, factorizing matrices and three-way tensors or factorize more higher 4 or 5-way tensors simultaneously. Other choices of divergence such as Euclidean distance are also important topics to investigate. Finally, we remark on the choice of the number of factor bases R. Determining appropriate R is an unsolved issue in NTF including the proposed NMTF. Though there are no definitive ways, this is also an important task to investigate.

Acknowledgment: This work was partially supported by JSPS KAKENHI 25870192.

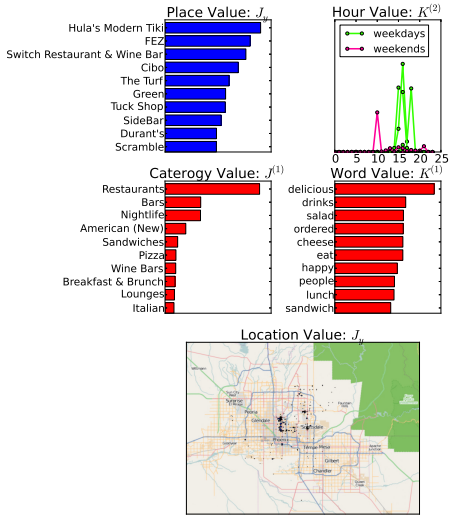
References

[1] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. ICML*, 2005.

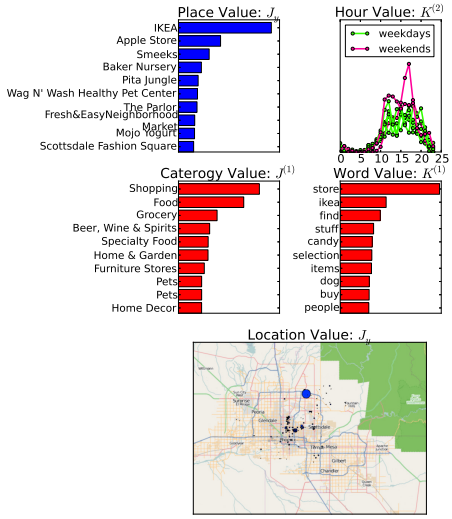
[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.



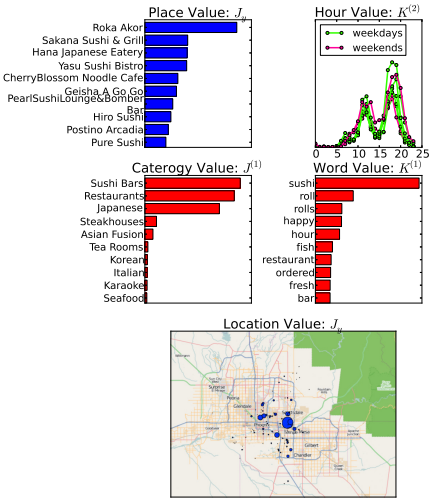
(a) Amusement and park activity factor on weekend morning (r=1).



(b) Bar and restaurant factor on weekday night and weekend morning (r=2).



(c) Shopping at various stores factor on weekend evening (r=3).



(d) Asian restaurant factor on both weekdays and weekends (r=4).

Fig. 3. Factors corresponds to four bases extracted from the Yelp data set.

[3] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

[4] D. FitzGerald, M. Cranitch, and E. Coyle, "Sound source separation using shifted non-negative tensor factorisation," in *Proc. ICASSP*, 2006.

[5] T. Hazan, S. Polak, and A. Shashua, "Sparse image coding using a 3D non-negative tensor factorization," in *Proc. ICCV*, 2005.

[6] E. C. Chi and T. G. Kolda, "On tensors, sparsity, and nonnegative factorizations," *SIAM Journal on Matrix Analysis and Applications*, vol. 33, no. 4, pp. 1272–1299, 2012.

[7] Q. Zhang, M. W. Berry, B. T. Lamb, and T. Samuel, "A parallel nonnegative tensor factorization algorithm for mining global climate data," in *Proc. ICCS*, 2009.

[8] W. Chu and Z. Ghahramani, "Probabilistic models for incomplete multi-dimensional arrays," in *Proc. AISTATS*, 2009.

[9] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. ICML*, 2011.

[10] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer Society*, vol. 42, no. 8, pp. 30–37, 2009.

[11] H. Ma, H. Yang, M. R. Lyu, and I. King, "SoRec: Social recommendation using probabilistic matrix factorization," in *Proc. CIKM*, 2008.

[12] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. NIPS*, 2008.

[13] K. Y. Yılmaz, A. T. Cemgil, and U. Simsekli, "Generalised coupled tensor factorisation," in *Proc. NIPS*, 2011.

[14] T. Yokota, A. Cichocki, and Y. Yamashita, "Linked PARAFAC/CP tensor decomposition and its fast implementation for multi-block tensor analysis," in *Proc. ICONIP*, 2012.

[15] S. Zafeiriou and M. Petrou, "Nonnegative tensor factorization as an alternative Csiszar–Tusnady procedure: algorithms, convergence, probabilistic interpretations and novel probabilistic tensor latent variable analysis algorithms," *Data Mining and Knowledge Discovery*, vol. 22, no. 3, pp. 419–466, 2011.

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

[17] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons, and S. Amari, "Novel multi-layer non-negative tensor factorization with sparsity constraints," in *Proc. ICANNGA*, 2007.

[18] J. Antikainen, J. Havel, R. Josth, A. Herout, P. Zemcik, and M. Hauta-Kasari, "Nonnegative tensor factorization accelerated using GPGPU," *IEEE Trans.*, vol. 22, no. 7, pp. 1135–1141, 2011.

[19] S. K. Gupta, D. Phung, B. Adams, T. Tran, and S. Venkatesh, "Non-negative shared subspace learning and its application to social media retrieval," in *Proc. SIGKDD*, 2010.

[20] G. Bouchard, S. Guo, and D. Yin, "Convex collective matrix factorization," in *Proc. AISTATS*, 2013.

[21] K. Takeuchi, K. Ishiguro, A. Kimura, and H. Sawada, "Non-negative multiple matrix factorization," in *Proc. IJCAI*, 2013.