PROBABILISTIC MATRIX TRI-FACTORIZATION

Jiho Yoo and Seungjin Choi

Department of Computer Science, POSTECH, Korea
{zentasis,seungjin}@postech.ac.kr

ABSTRACT

Nonnegative matrix tri-factorization (NMTF) is a 3-factor decomposition of a nonnegative data matrix, $\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where factor matrices, $\boldsymbol{U}$, $\boldsymbol{S}$, and $\boldsymbol{V}$, are restricted to be nonnegative as well. Motivated by the aspect model used for dyadic data analysis as well as in probabilistic latent semantic analysis (PLSA), we present a probabilistic model with two dependent latent variables for NMTF, referred to as *probabilistic matrix tri-factorization* (PMTF). Each latent variable in the model is associated with the cluster variable for the corresponding object in the dyad, leading the model suited to co-clustering. We develop an EM algorithm to learn the PMTF model, showing its equivalence to multiplicative updates derived by an algebraic approach. We demonstrate the useful behavior of PMTF in a task of document clustering. Moreover, we incorporate the likelihood in the PMTF model into existing information criteria so that the number of clusters can be detected, while the algebraic NMTF cannot.

Index Terms— Co-clustering, document clustering, probabilistic latent semantic indexing, nonnegative matrix factorization

1. INTRODUCTION

Nonnegative matrix factorization (NMF) [11] is a method for multivariate analysis of nonnegative data, seeking a 2-factor decomposition of a nonnegative data matrix, $\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{V}^\top$, where two factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are restricted to be nonnegative. NMF has been successfully applied to a variety of applications, including face detection and recognition, audio and speech processing, text mining, biomedical image analysis, bioinformatics, and so on.

One of prominent applications of NMF, which is of our interest, is document clustering that plays an important role in dyadic data analysis [16, 15]. Dyadic data refers to a domain with two finite sets of objects in which observations are made for pairs with one element from either set [8]. A term-document matrix contains *co-occurrence* frequencies of word-document pairs. When NMF is applied to a term-document matrix, the matrix is decomposed into a product of two factor matrices, where one corresponds to cluster centers and the other is associated with cluster indicator variables [16]. Orthogonality constraints were imposed on factor matrices in the decomposition [5, 2], where a clear link between k-means clustering and NMF is made in such a case. Multiplicative updates for orthogonal NMF were developed, exploiting directly gradient information on Stiefel manifolds [2, 17].

Nonnegative matrix tri-factorization (NMTF) is a 3-factor decomposition, $\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, with nonnegative constraints imposed on factor matrices $\boldsymbol{U}$, $\boldsymbol{S}$, and $\boldsymbol{V}$. Matrix tri-factorization was studied for co-clustering [4] which aims at the simultaneous clustering of rows and columns of a term-document matrix. Block value decomposition (BVD) seeks an approximation of a term-document matrix as a product of row-coefficient matrix, block value matrix, and column-coefficient matrix [13]. Nonnegativity constraints are easily incorporated into BVD. Orthogonality constraints on factor matrices were also considered in nonnegative matrix tri-factorization (NMTF) [5], where the co-clustering was shown to improve the accuracy of document clustering.

The aspect model [8], which is a statistical latent variable model, defines a proper generative model for factor analysis of dyadic data. Probabilistic latent semantic analysis (PLSA) makes use of the aspect model for document clustering. An interesting link between PLSA and NMF was revealed in [6], where the multiplicative updating algorithm for NMF with KL-divergence considered was shown to be equivalent to the EM algorithm for PLSA. In this paper, we present a probabilistic model for NMTF, referred to as *probabilistic matrix tri-factorization* (PMTF) and develop an EM algorithm to learn the PMTF model. As in [6], we show a link between an algebraic approach to NMTF and our probabilistic model. Experiments on several document datasets confirm its comparable performance in a task of document clustering. In addition, we demonstrate that PMTF is able to detect a proper number of clusters, incorporating the likelihood into Akaike information criterion, while algebraic NMTF cannot.

2. PLSA AND NMF

We present a quick review of PLSA [7] where the aspect model is used to model co-occurrence data, as well as a link between PLSA and NMF [6]. Throughout this paper, we consider a term-document matrix $\boldsymbol{X} \in \mathbb{R}_+^{M\times N}$, where observations X_{ij} are co-occurrence frequencies of dyads (w_i, d_j) (i.e., the significance of term (word) w_i in document d_j) for two sets of objects, $\mathcal{W} = \{w_1, \ldots, w_M\}$ and $\mathcal{D} = \{d_1, \ldots, d_N\}$.

PLSA makes use of a statistical latent class model (known as aspect model), for factor analysis of dyadic data. The generative model used in PLSA is given by

$$p(w_i, d_j) = \sum_z p(w_i|z)p(d_j|z)p(z), \tag{1}$$

where $z \in \{1, \ldots, K\}$ is the latent class variable. Given the latent variable z, random variables w and d are conditionally independent. The complete-data likelihood is given by

$$p(\mathcal{X}, y) = \prod_i \prod_j p(w_i, d_j, z)^{C_{ij}}, \tag{2}$$

where C_{ij} are empirical counts for dyads (w_i, d_j) in $\mathcal{X} = \{(w_i, d_j)\}$. EM algorithm to learn the model (1) was developed in [7], where E-step is given by

$$p(z|w_i, d_j) = \frac{p(z)p(w_i|z)p(d_j|z)}{\sum_{z'} p(z')p(w_i|z')p(d_j|z')}, \tag{3}$$

and M-step re-estimates parameters

$$p(w_i|z) = \frac{\sum_j C_{ij} p(z|w_i, d_j)}{\sum_i \sum_j C_{ij} p(z|w_i, d_j)}, \tag{4}$$

$$p(d_j|z) = \frac{\sum_i C_{ij} p(z|w_i, d_j)}{\sum_i \sum_j C_{ij} p(z|w_i, d_j)}, \tag{5}$$

$$p(z) = \frac{\sum_i \sum_j C_{ij} p(z|w_i, d_j)}{\sum_i \sum_j C_{ij}}. \tag{6}$$

On the other hand, NMF seeks a 2-factor decomposition of $\boldsymbol{X}$ that is of the form

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{V}^\top, \tag{7}$$

where factor matrices, $\boldsymbol{U} \in \mathbb{R}_+^{M \times K}$ and $\boldsymbol{V} \in \mathbb{R}_+^{N \times K}$, are restricted to be nonnegative. Considering KL-divergence between $\boldsymbol{X}$ and $\boldsymbol{U}\boldsymbol{V}^\top$, as a discrepancy measure, multiplicative updates [12] for factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are given by

$$U_{ij} \leftarrow U_{ij} \frac{\sum_k (X_{ij}/[\boldsymbol{U}\boldsymbol{V}^\top]_{ik}) V_{kj}}{\sum_k V_{kj}}, \tag{8}$$

$$V_{ij} \leftarrow V_{ij} \frac{\sum_k (X_{ki}/[\boldsymbol{U}\boldsymbol{V}^\top]_{ki}) U_{kj}}{\sum_k U_{kj}}. \tag{9}$$

Without loss of generality, assume that $\sum_i \sum_j X_{ij} = 1$. We define scaling matrices $\boldsymbol{D}_U \equiv \text{diag}\left(\boldsymbol{1}^\top \boldsymbol{U}\right)$ and $\boldsymbol{D}_V \equiv \text{diag}\left(\boldsymbol{1}^\top \boldsymbol{V}\right)$, where $\boldsymbol{1} = [1, \ldots, 1]^\top$. Then the factorization (7) can be rewritten as

$$\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{D}_U^{-1})(\boldsymbol{D}_U \boldsymbol{D}_V)(\boldsymbol{V}\boldsymbol{D}_V^{-1})^\top. \tag{10}$$

Comparing (10) with the factorization (1), one can see that entries of $(\boldsymbol{U}\boldsymbol{D}_U^{-1})$ correspond to $p(w_i|z)$, elements of $(\boldsymbol{V}\boldsymbol{D}_V^{-1})$ are associated with $p(d_j|z)$, and $\boldsymbol{D} \equiv \boldsymbol{D}_U \boldsymbol{D}_V$ corresponds to cluster prior $p(z)$. It was shown in [6] that the EM algorithm for PLSA is equivalent to multiplicative updates for NMF with KL-divergence error measure.

3. NONNEGATIVE MATRIX TRI-FACTORIZATION

Nonnegative matrix tri-factorization (NMTF) is a 3-factor decomposition of a nonnegative dyadic data matrix $\boldsymbol{X} \in \mathbb{R}_+^{M \times N}$ that takes the form

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top, \tag{11}$$

where $\boldsymbol{U} \in \mathbb{R}_+^{M \times K}$, $\boldsymbol{S} \in \mathbb{R}_+^{K \times R}$, and $\boldsymbol{V} \in \mathbb{R}_+^{R \times N}$ are constrained to be nonnegative matrices.

Recently matrix tri-factorization draws extensive attention due to its usefulness in co-clustering dyadic data such as co-occurrence matrix, rating matrix, and proximity matrix. Block value decomposition [13] exploits the latent block structure captured by $\boldsymbol{S}$ for co-clustering. The factorization (11) with orthogonality constraints ($\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$) is also developed [5].

We consider KL-divergence of the model from the data as an error measure,

$$\mathcal{J} = \sum_{i=1}^{M} \sum_{j=1}^{N} \left\{ X_{ij} \log \frac{X_{ij}}{[\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top]_{ij}} - X_{ij} + [\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top]_{ij} \right\}, \tag{12}$$

where $[\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top]_{ij} = \sum_a \sum_b U_{ia} S_{ab} V_{jb}$. Derivatives with respect to each factor matrix are given by

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial U_{ij}} &= \sum_{a=1}^{N} \left\{ [\boldsymbol{V}\boldsymbol{S}^\top]_{aj} - \frac{X_{ia}}{[\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top]_{ia}} [\boldsymbol{V}\boldsymbol{S}^\top]_{aj} \right\}, \\
\frac{\partial \mathcal{J}}{\partial V_{ij}} &= \sum_{a=1}^{M} \left\{ [\boldsymbol{U}\boldsymbol{S}]_{aj} - \frac{X_{ai}}{[\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top]_{ai}} [\boldsymbol{U}\boldsymbol{S}]_{aj} \right\}, \\
\frac{\partial \mathcal{J}}{\partial S_{ij}} &= \sum_{a=1}^{M} \sum_{b=1}^{N} \left\{ U_{ai} V_{bj} - \frac{X_{ab}}{[\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top]_{ab}} U_{ai} V_{bj} \right\}.
\end{aligned}$$

Applying the techniques [3], one can easily derive multiplicative updates that are of the form:

$$U_{ij} \leftarrow U_{ij} \frac{\sum_{a=1}^{N} (X_{ia}/[\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top]_{ia}) [\boldsymbol{V}\boldsymbol{S}^\top]_{aj}}{\sum_{a=1}^{N} [\boldsymbol{V}\boldsymbol{S}^\top]_{aj}}, \tag{13}$$

$$V_{ij} \leftarrow V_{ij} \frac{\sum_{a=1}^{M} (X_{ai}/[\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top]_{ai}) [\boldsymbol{U}\boldsymbol{S}]_{aj}}{\sum_{a=1}^{N} [\boldsymbol{U}\boldsymbol{S}]_{aj}}, \tag{14}$$

$$S_{ij} \leftarrow S_{ij} \frac{\sum_{a=1}^{M} \sum_{b=1}^{N} (X_{ab}/[\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top]_{ab}) U_{ai} V_{bj}}{\sum_{a=1}^{M} \sum_{b=1}^{N} U_{ai} V_{bj}}. \tag{15}$$

4. PROBABILISTIC MATRIX TRI-FACTORIZATION

4.1. Probabilistic model

We consider a term-document matrix but our model and algorithm work for other dyadic matrices as well. Introducing two latent variables, y_l and z_k, which are associated with cluster variables for terms (words) and documents, respectively. The term-document joint distribution is factorized as

$$\begin{aligned}
p(w_i, d_j) &= \sum_{l=1}^{R} \sum_{k=1}^{K} p(w_i, d_j | y_l, z_k) p(y_l, z_k) \\
&= \sum_{l=1}^{R} \sum_{k=1}^{K} p(w_i|y_l) p(d_j|z_k) p(y_l, z_k),
\end{aligned} \tag{16}$$

where $p(y_l, z_k)$ is the joint prior probability for term cluster y_l and document cluster z_k. Relating (16) to the 3-factor decomposition (11), marginal distributions $p(w_i|y_l)$ and $p(d_j|z_k)$ are associated with $\boldsymbol{U}\boldsymbol{D}_U^{-1}$ and $\boldsymbol{V}\boldsymbol{D}_V^{-1}$, respectively. Then by the following factorization

$$\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{D}_U^{-1})(\boldsymbol{D}_U \boldsymbol{S} \boldsymbol{D}_V)(\boldsymbol{V}\boldsymbol{D}_V^{-1})^\top,$$

we can see that the joint probability $p(y_l, z_k)$ is represented by each element of $\boldsymbol{D}_U \boldsymbol{S} \boldsymbol{D}_V$.

To obtain a posterior probability of documents, we have to calculate the column of the $\boldsymbol{D}_U \boldsymbol{S} \boldsymbol{D}_V$ to get a prior probability for document cluster, as

$$\begin{aligned}
p(z_k|d_j) &\propto p(d_j|z_k) p(z_k) \\
&= p(d_j|z_k) \sum_a p(y_l, z_k) \\
&= \left[\text{diag}(\boldsymbol{1}^\top \boldsymbol{D}_U \boldsymbol{S} \boldsymbol{D}_V)(\boldsymbol{D}_V^{-1} \boldsymbol{V}^\top) \right]_{kj} \\
&= \left[\text{diag}(\boldsymbol{1}^\top \boldsymbol{D}_U \boldsymbol{S}) \boldsymbol{D}_V \boldsymbol{D}_V^{-1} \boldsymbol{V}^\top \right]_{kj} \\
&= \left[\text{diag}(\boldsymbol{1}^\top \boldsymbol{D}_U \boldsymbol{S}) \boldsymbol{V}^\top \right]_{kj}.
\end{aligned}$$

Therefore, we should normalize the encoding matrix by using the matrix diag$(\mathbf{1}^\top \boldsymbol{D}_U \boldsymbol{S})$. We assign document d_j to cluster k^* if

$$k^* = \arg\max_k \left[\boldsymbol{V} \operatorname{diag}\left(\mathbf{1}^\top \boldsymbol{D}_U \boldsymbol{S}\right)\right]_{jk}. \tag{17}$$

4.2. EM algorithm

In order to develop an EM algorithm to learn our model (16), we first consider the complete-data distribution

$$p(\mathcal{X}, y, z) = \prod_{i=1}^{M} \prod_{j=1}^{N} p(w_i, d_j, y, z)^{C_{ij}},$$

where C_{ij} is the empirical counts for dyad (w_i, d_j) in $\mathcal{X} = \{(w_i, d_j)\}$. In E-step we calculate the expected complete-data log-likelihood $\mathbb{E}\,\mathcal{L}_c$ that has the form

$$\mathbb{E}\,\mathcal{L}_c = \sum_{l,k,i,j} C_{ij} p(y_l, z_k|w_i, d_j) \log[p(w_i|y_l)p(d_j|z_k)p(y_l, z_k)],$$

where the posterior distribution over latent variables is computed as

$$p(y_l, z_k|w_i, d_j) = \frac{p(w_i|y_l)p(d_j|z_k)p(y_l, z_k)}{\sum_l \sum_k p(w_i|y_l)p(d_j|z_k)p(y_l, z_k)}. \tag{18}$$

In M-step we re-estimate parameters $p(w_i|l)$, $p(d_j|z_k)$, and $p(y_l, z_k)$ such that updated parameters (denoted by $\tilde{p}$) are computed as

$$\tilde{p}(w_i|y_l) = \frac{\sum_k \sum_j C_{ij} p(y_l, z_k|w_i, d_j)}{\sum_k \sum_i \sum_j C_{ij} p(y_l, z_k|w_i, d_j)}, \tag{19}$$

$$\tilde{p}(d_j|z_k) = \frac{\sum_l \sum_i C_{ij} p(y_l, z_k|w_i, d_j)}{\sum_l \sum_i \sum_j C_{ij} p(y_l, z_k|w_i, d_j)}, \tag{20}$$

$$\tilde{p}(y_l, z_k) = \frac{\sum_i \sum_j C_{ij} p(y_l, z_k|w_i, d_j)}{\sum_i \sum_j C_{ij}}. \tag{21}$$

We show the equivalence between NMTF multiplicative updates (13)-(15) and EM algorithm (19)-(21), as in [6] where the relation between PLSI and NMF was shown. We define $T = \sum_i \sum_j C_{ij}$. Then, combining (19) and (21) leads to

$$\begin{aligned} \tilde{p}(w_i|y_l) &= \frac{\sum_k \sum_j C_{ij} p(y_l, z_k|w_i, d_j)}{\sum_k \tilde{p}(y_l, z_k) T} \\ &= \frac{\sum_j \left(\frac{C_{ij}}{R}\right) p(w_i|y_l) \frac{\sum_k p(y_l, z_k) p(d_j|z_k)}{p(w_i, d_j)}}{\sum_k \tilde{p}(y_l, z_k)}. \end{aligned}$$

Taking into account $X_{ij} = \frac{C_{ij}}{T}$ and $\sum_a [\boldsymbol{V}\boldsymbol{S}^\top]_{aj} = \tilde{p}(y_j)$, one can easily see the equivalence between (19) and (13). In the same manner, remaining equations can be also shown to be equivalent.

5. NUMERICAL EXPERIMENTS

5.1. Document clustering

We applied the proposed PMTF algorithm to the task of document clustering. We also tested NMF and NMTF with KL-divergence for comparison. Four standard document datasets were used: CSTR[5], NG20-binary, NG20-multi5, and NG20-multi10[13], which has 4, 2, 5, and 10 clusters, respectively. Clustering accuracy was used as a performance measure. We measured the averaged accuracies over 10 trials with different initial conditions for each algorithm and dataset (Table 1). All the algorithms showed similar performances, and none of the algorithms are significantly better than the others. The PMTF algorithm works well for the document clustering task, comparable to the conventional NMF algorithm with KL-divergence.

Table 1. Clustering accuracies of three algorithms: NMF, NMTF, and PMTF algorithm, averaged over 10 trials.

	NMF	NMTF	PMTF
CSTR	0.7877	0.7814	0.7940
NG20-binary	0.9111	0.9029	0.9092
NG20-multi5	0.6731	0.6781	0.6654
NG20-multi10	0.4079	0.3984	0.4007

5.2. Estimating the number of clusters

In the practical situations, the number of clusters of given data is usually not known in advance. To determine the number of clusters of the document data, the most simple approach is to examine the resulting divergence between the data and model. This is based on the assumption that the divergence becomes smaller when we select appropriate number of clusters. However, the model with larger number of clusters usually over-fitted to the data, resulting smaller divergence than the divergence with the correct number of clusters. To prevent this and select correct number of clusters, we have to add some value on the divergences from the models with larger number of clusters. This kind of penalization is not a straight-forward task because the resulting optimal cluster numbers can vary with the amount of penalization.

In PMTF, we can calculate the likelihood of the fitted model, so we can directly apply well-established statistical theory of model selection. In this case, the amount of penalization can be determined based on the statistical theory. Akaike information criterion (AIC) [1] and Bayesian information criterion (BIC) [14] are the standard model selection methods based on the statistical theory. AIC compares the following quantities over the models,

$$\text{AIC} = \log \mathcal{L}_{\mathcal{M}} - k_{\mathcal{M}},$$

where $\log \mathcal{L}_{\mathcal{M}}$ is the log-likelihood of the model $\mathcal{M}$ and $k_{\mathcal{M}}$ is the number of free parameters in the model. In the PMTF, $\log \mathcal{L}_{\mathcal{M}}$ is calculated as

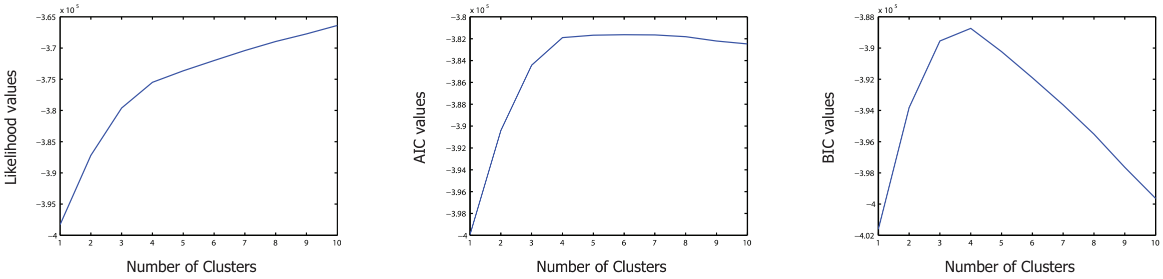
$$\begin{aligned} \log \mathcal{L}_{\mathcal{M}} &= \sum_i \sum_j C_{ij} \log p(w_i, d_j) \\ &= \sum_i \sum_j C_{ij} \log \left[\sum_l \sum_k p(w_i|y_l) p(d_j|z_k) p(y_l, z_k) \right], \end{aligned}$$

and $k_{\mathcal{M}} = MK + KR + RN$. On the other hand, BIC makes use of different penalization

$$\text{BIC} = \log \mathcal{L}_{\mathcal{M}} - \frac{1}{2} k_{\mathcal{M}} \log T,$$

where T is the number of data points. In the case of PMTF, $T = \sum_i \sum_j C_{ij}$. BIC gives heavier penalty on the complex model than AIC in usual cases (when $T > e^2 \approx 7.38$). The number of clusters with maximum AIC or BIC value is selected as a true number of clusters.

We demonstrate the estimation of the number of clusters using AIC and BIC for the CSTR dataset (Fig. 1). The true number of document clusters of CSTR dataset is four. We tried different number of clusters from one to ten, and compute AIC and BIC values for each case. Number of term clusters was set to be equal to the number of document clusters for each case. Ten trials were done for each case, and the average of final likelihood value is used. As a result, AIC still cannot find the true number of clusters because of too small amount of penalty. However, maximum BIC clearly indicates the true number of clusters.



(a) (b) (c)

Fig. 1. An exemplary behavior with the number of clusters varying, on CSTR dataset is shown: (a) the likelihood which increases as the number of clusters increases; (b) AIC values where the curve starts to be flat when the number of clusters equals 4; (c) BIC values where the peak is achieved at 4. Both AIC and BIC correctly identify the true of number of clusters (which is 4 in our test).

6. CONCLUSIONS

We have presented a probabilistic model for NMTF, referred to as probabilistic matrix tri-factorization (PMTF), developing an EM algorithm to learn the model. Motivated by PLSA and NMF, we have shown that the EM algorithm for PMTF is equivalent to multiplicative updates for NMTF that are derived in an algebraic way. More specifically, they are equivalent each other at stationary points. Experiments on document data sets confirmed that NMF, NMTF, and PMTF work well in a task of document clustering. However, PMTF is capable of detecting the number of clusters by incorporating the likelihood into AIC or BIC, while algebraic NMF or NMTF is not. As a future work, we are working on multilinear generalization of PMTF, developing a probabilistic model counterpart of nonnegative Tucker decomposition [9, 10], in order to tackle polyadic data.

Acknowledgments: This work was supported by Korea Research Foundation Grant (KRF-2008-313-D00939), KOSEF WCU Program (Project No. R31-2008-000-10100-0), and Microsoft Research Asia.

7. REFERENCES

[1] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[2] S. Choi, “Algorithms for orthogonal nonnegative matrix factorization,” in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Hong Kong, 2008.

[3] A. Cichocki, H. Lee, Y. D. Kim, and S. Choi, “Nonnegative matrix factorization with α-divergence,” *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1433–1440, 2008.

[4] I. S. Dhillon, S. Mallela, and D. S. Modha, “Information-theoretic co-clustering,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2003, pp. 89–98.

[5] C. Ding, T. Li, W. Peng, and H. Park, “Orthogonal nonnegative matrix tri-factorizations for clustering,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Philadelphia, PA, 2006.

[6] E. Gaussier and C. Goutte, “Relation between PLSA and NMF and implications,” in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil, 2005.

[7] T. Hofmann, “Probablistic latent semantic analysis,” in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999.

[8] T. Hofmann, J. Puzicha, and M. I. Jordan, “Learning from dyadic data,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 11. MIT Press, 1999.

[9] Y. D. Kim and S. Choi, “Nonnegative Tucker decomposition,” in *Proceedings of the IEEE CVPR-2007 Workshop on Component Analysis Methods*, Minneapolis, Minnesota, 2007.

[10] Y. D. Kim, A. Cichocki, and S. Choi, “Nonnegative Tucker decomposition with alpha-divergence,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, 2008.

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

[12] ——, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13. MIT Press, 2001.

[13] B. Long, Z. Zhang, and P. S. Yu, “Co-clustering by block value decomposition,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Chicago, IL, 2005.

[14] G. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[15] F. Shahnaz, M. Berry, P. Pauca, and R. Plemmons, “Document clustering using nonnegative matrix factorization,” *Information Processing and Management*, vol. 42, pp. 373–386, 2006.

[16] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, Toronto, Canada, 2003.

[17] J. Yoo and S. Choi, “Orthogonal nonnegative matrix factorization: Multiplicative updates on Stiefel manifolds,” in *Proceedings of the 9th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, Daejeon, Korea, 2008.