Joint Factor Analysis and Latent Clustering

Bo Yang
Dept. ECE, Univ. Minnesota
Minneapolis, MN 55455
Email: yang4173@umn.edu

Xiao Fu
Dept. ECE, Univ. Minnesota
Minneapolis, MN 55455
Email: xfu@umn.edu

Nicholas D. Sidiropoulos
Dept. ECE, Univ. Minnesota
Minneapolis, MN 55455
Email: nikos@umn.edu

***Abstract*—Many real-life datasets exhibit structure in the form of physically meaningful clusters - e.g., news documents can be categorized as sports, politics, entertainment, and so on. Taking these clusters into account together with low-rank structure may yield parsimonious matrix and tensor factorization models and more powerful data analytics. Prior works made use of *data-domain* similarity to improve nonnegative matrix factorization. Here we are instead interested in joint low-rank factorization and *latent-domain* clustering; that is, in clustering the latent reduced-dimension representations of the observed entities. A unified algorithmic framework that can deal with both matrix and tensor factorization and latent clustering is proposed. Numerical results obtained from synthetic and real document data show that the proposed approach can significantly improve factor analysis and clustering accuracy.**

I. Introduction

Factoring a data matrix or tensor (a dataset indexed by more than two indices) into a sum of (typically few) rank-one factors is often referred to as *factor analysis*. Factor analysis finds numerous applications in signal processing, machine learning, and data mining – where it is often used for dimensionality reduction. The singular value decomposition (SVD) is the most widely used tool for dimensionality reduction and latent semantic indexing, but other types of factor analysis have also become popular in recent years – such as nonnegative matrix factorization (NMF). Whereas for matrices constraints such as orthogonality (used in SVD) and nonnegativity are crucial for unique decomposition [1], low-rank structure alone is enough to ensure uniqueness of tensor decomposition, under fairly mild conditions [2]. Such uniqueness is important for latent cluster analysis, for otherwise the latent dimensions may be meaningless from the viewpoint of clustering and interpretation.

Factor analysis lies at the confluence of linear algebra and optimization, and finds numerous applications in quantitative sciences ranging from econometrics and psychometrics to chemometrics. As a result, theoretical aspects such as uniqueness and practical issues such as factorization algorithms have been well-investigated for many commonly used types of factor analysis. Beyond uniqueness (*identifiability*), additional problem-specific regularization and constraints on the model parameters are often very useful to further overdetermine the problem, thereby ensuring more accurate estimates. This observation is not surprising, since real data always contain various types of modeling errors, such as outliers, missing values, and measurement noise, to name a few.

Supported by NSF IIS 1447788.

Different kinds of constraints can be imposed on the loadings and scores (i.e., the coefficients associated with the loadings) when factoring a matrix or a tensor, depending on what is known and meaningful for a given application. For example, sparsity and nonnegativity have been used for social network co-clustering [3]; total variation has been used for hyperspectral unmixing [4]; and data (row) similarity has been used for document clustering [5].

In this paper, we propose to employ another type of prior information for factor analysis. Specifically, our interest lies in imposing cluster structure on the latent representations of the entities associated with the rows (and/or columns) of the matrix or tensor to be factored. If one interprets these reduced-dimension latent representations as the underlying generative model of the observed data, then it is natural to think about spotting similarities and differences in latent space, i.e., in what one could call *latent data mining*. For example, in the Enron email data [6], senders and receivers belong to different working groups, such as the legal group and the executive group. Another example is the Reuters news document dataset, where the news stories can be classified as sports, economy, politics, culture, etc. Prior works have shown that the latent representations of these datasets exhibit cluster structures [6], [7], but this has merely been observed *a posteriori*, instead of being exploited *a priori* to help with factorization.

Some prior work, e.g., [7], [8] made use of *data-domain* clustering to improve the performance of NMF, by imposing latent similarity according to data similarity. This is fundamentally different from *latent-domain* clustering, which has not been explored, to the best of our knowledge. The reason for this is that, due to multilinearity, two entities that are close in latent space may be far in data space and vice versa. In other words, if we compare two entities in the same mode, the other mode acts as a distance-weighting factor that distorts the geometry. To see this clearly, let $\mathbf{X} = \mathbf{A}\mathbf{B}^T$, and consider the squared distance between the first two columns of $\mathbf{X}$, i.e., $||\mathbf{X}(:,1) - \mathbf{X}(:,2)||^2 = ||\mathbf{A}(\mathbf{B}(1,:) - \mathbf{B}(2,:)^T||^2 = (\mathbf{B}(1,:)-\mathbf{B}(2,:))\mathbf{A}^T\mathbf{A}(\mathbf{B}(1,:)-\mathbf{B}(2,:))^T$, where : stands for all values of the respective argument. Notice how the matrix $\mathbf{A}^T\mathbf{A}$ weights the latent-domain distance to produce the data-domain distance, distorting the geometry.

In this work, we aim at taking advantage of the *latent* cluster structures when factoring a matrix or tensor. Unlike [7], [8],

our method does not depend on data similarity, but rather aims to model and learn the latent cluster structure directly. Our formulation balances data fidelity and the latent cluster structure, which can hopefully yield more accurate loading factors as well as cleaner clusters – iterating between factorization and clustering may be mutually beneficial for both tasks. An alternating optimization algorithm that guarantees monotonic decrease of the cost function is proposed. Numerical results using synthetic and real data are presented to showcase the effectiveness of the proposed approach.

II. Problem Formulation

We first consider a simple nonnegative matrix factorization (NMF) model. The development can be generalized to other structured matrix and tensor factorization models in a straightforward manner (cf. Remark 1). The following NMF formulation is commonly adopted in the literature for factoring a data matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$:

$$\begin{aligned} \min_{\mathbf{W} \in \mathbb{R}^{I \times F}, \mathbf{H} \in \mathbb{R}^{F \times J}} \quad & \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 \\ \text{s.t.} \quad & \mathbf{W} \geq \mathbf{0},\ \mathbf{H} \geq \mathbf{0}. \end{aligned} \tag{1}$$

where "$\geq$" denotes the element-wise inequality and $\mathbf{0}$ is an all-zero matrix with proper size. Although NMF is identifiable under certain conditions [1], incorporating additional prior information in the form of regularization can still be very useful in terms of anchoring the solution and fending against modeling errors. We focus on using latent cluster structure-based regularization. To formulate the problem of interest, let us consider the case where $\{\mathbf{X}(i,:)\}_{i=1}^I$ have latent representations drawn from K clusters, i.e., the rows of $\mathbf{W}$ can be divided into K clusters. A penalized NMF formulation would then naturally be as follows:

$$\min_{\mathbf{W},\mathbf{H},\mathbf{S},\mathbf{M}} \ \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \lambda \|\mathbf{W} - \mathbf{S}\mathbf{M}\|_F^2 \tag{2a}$$

$$\text{s.t. } \mathbf{W} \geq \mathbf{0},\ \mathbf{H} \geq \mathbf{0} \tag{2b}$$

$$\|\mathbf{S}(i,:)\|_0 = 1,\ \mathbf{S}(i,k) \in \{0,1\},\ \forall i,k, \tag{2c}$$

where $\|\cdot\|_0$ denotes the ℓ_0 quasi-norm, $\lambda \geq 0$ is a pre-specified regularization parameter, and $\mathbf{S} \in \mathbb{R}^{I \times K}$ and $\mathbf{M} \in \mathbb{R}^{K \times F}$ denote the cluster membership indicator matrix and the centroid matrix, respectively. Specifically, $\mathbf{S}(i,k) = 1$ means that $\mathbf{W}(i,:)$ belongs to cluster k, whose centroid is $\mathbf{M}(k,:)$. Notice that $\min_{\mathbf{S},\mathbf{M}} \ \|\mathbf{W} - \mathbf{S}\mathbf{M}\|_F^2$ together with (2c) is nothing but a K-means problem on the rows of $\mathbf{W}$. Also note that we can interpret the cost in (2) as a log-MAP (*maximum a posteriori*) criterion, where the data fidelity term comes from the conditional likelihood, and the penalty term from the prior density. In our context, the prior is a balanced mixture of uncorrelated Gaussians with different mean vectors (the cluster centroids) and equal variances.

Direct use of (2) does not yield 'correct' results, due to the scaling ambiguity of matrix/tensor factorization (i.e., $\rho\mathbf{W}(:,f)$ and $\frac{1}{\rho}\mathbf{H}(:,f)$ yield the same $\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2$, $\forall \rho > 0$), but such arbitrary scaling of the columns of $\mathbf{W}$ can have significant impact on the clustering of its rows, obviously. Also note that the K-means part in (2) uses Euclidean distance for clustering, which may not be suitable for certain kinds of data. For example, it has been observed that a better clustering metric for document and web data is correlation or cosine similarity [9], [10]. For data-domain K-means, computing cosine similarity and correlation of the data points can be easily done by normalizing the rows of $\mathbf{X}$ in advance. In our context, however, a naive adoption of the cosine similarity for the clustering part can complicate things, since $\mathbf{W}$ changes in every iteration. To accommodate this, we reformulate the problem as follows.

$$\begin{aligned} \min_{\substack{\mathbf{W},\mathbf{H} \\ \mathbf{S},\mathbf{M},\{d_i\}_{i=1}^I}} \quad & \|\mathbf{X} - \mathbf{D}\mathbf{W}\mathbf{H}\|_F^2 + \lambda \|\mathbf{W} - \mathbf{S}\mathbf{M}\|_F^2 + \eta\|\mathbf{H}\|_F^2 \\ \text{s.t. } & \mathbf{W} \geq \mathbf{0},\ \mathbf{H} \geq \mathbf{0},\ \|\mathbf{W}(i,:)\|_2 = 1,\ \forall i, \\ & \mathbf{D} = \mathrm{Diag}(d_1, \ldots, d_I), \\ & \|\mathbf{S}(i,:)\|_0 = 1,\ \mathbf{S}(i,k) \in \{0,1\},\ \forall i,k, \end{aligned} \tag{3}$$

where $\eta \geq 0$ is a regularization parameter. Introducing the diagonal matrix $\mathbf{D}$ is crucial: It allows us to fix the rows of $\mathbf{W}$ onto the unit 2-norm ball without loss of generality of the factorization model. Notice that we restrict $\mathbf{W}(i,:)$ to have unit-norm for $i = 1, \ldots, I$ so that the Euclidean distance-based K-means clustering on the unit 2-norm ball is equivalent to correlation-based clustering. The last term in the cost function is added to prevent $\mathbf{H}$ from "absorbing" all the energy into it by arbitrarily scaling up its columns; i.e., $\eta\|\mathbf{H}\|_F^2$ has been added to automatically control the scaling ambiguity.

Remark 1 The formulation in (3) is flexible and easily generalizable to other low-rank matrix and tensor factorization models. Consider a three-way tensor $\underline{\mathbf{X}} \in \mathbb{R}^{L \times M \times N}$ with nonnegative loading factors $\mathbf{A} \in \mathbb{R}^{L \times F}$, $\mathbf{B} \in \mathbb{R}^{M \times F}$, $\mathbf{C} \in \mathbb{R}^{N \times F}$. If we know that the rows of $\mathbf{A}$ can be clustered into K groups, we can formulate the *joint nonnegative tensor factorization (NTF) and latent clustering problem* as

$$\begin{aligned} \min_{\substack{\mathbf{A},\mathbf{B},\mathbf{C} \\ \mathbf{S},\mathbf{M},\{d_\ell\}_{\ell=1}^L}} \quad & \left\|\underline{\mathbf{X}}^{(1)} - \mathbf{D}\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\right\|_F^2 + \lambda \|\mathbf{A} - \mathbf{S}\mathbf{M}\|_F^2 \\ & \qquad + \eta\|\mathbf{B}\|_F^2 + \eta\|\mathbf{C}\|_F^2 \\ \text{s.t. } & \mathbf{A},\mathbf{B},\mathbf{C} \geq \mathbf{0},\ \|\mathbf{A}(\ell,:)\|_2 = 1,\ \forall \ell, \\ & \mathbf{D} = \mathrm{Diag}(d_1, \ldots, d_L), \\ & \|\mathbf{S}(i,:)\|_0 = 1,\ \mathbf{S}(\ell,k) \in \{0,1\},\ \forall \ell,k, \end{aligned} \tag{4}$$

where $\mathbf{X}^{(1)} \in \mathbb{R}^{L \times MN}$ is the first-mode matrix unfolding of $\underline{\mathbf{X}}$ [11], [12], "$\odot$" denotes the Khatri-Rao product, $\mathbf{S} \in \mathbb{R}^{L \times K}$ and $\mathbf{M} \in \mathbb{R}^{K \times F}$ are defined as before, and the regularization terms $\|\mathbf{B}\|_F^2$ and $\|\mathbf{C}\|_F^2$ are there to control scaling. If one believes that each loading matrix has a latent cluster structure, clustering-based regularization on $\mathbf{B}$ and $\mathbf{C}$ can also be incorporated.

III. Optimization via Variable Splitting

In this section, we propose a unified algorithmic framework for tackling problems (3) and (4). For ease of exposition, we

use problem (3) as a working example. Generalization to problem (4) is straightforward. Our basic strategy is to alternate between updating $\mathbf{W}$, $\mathbf{H}$, $\mathbf{S}$, $\mathbf{M}$, and $\{d_i\}_{i=1}^I$ one at a time, while fixing the others. The difficulty of implementing this strategy lies in the partial optimizations with respect to (w.r.t.) $\mathbf{W}$ and $\mathbf{S}$, which are nonconvex. For the subproblems w.r.t. $\mathbf{S}$ and $\mathbf{M}$, we propose to use the corresponding (alternating) steps of classical K-means [13]. The partial minimization w.r.t. $\mathbf{W}$ needs more effort, due to the unit row-norm and nonegativity constraints. Here, we propose to employ a variable-splitting strategy. Specifically, we consider the following optimization surrogate:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{H},\mathbf{Z},\mathbf{S},\mathbf{M},\{d_i\}_{i=1}^I} \quad & \|\mathbf{X}-\mathbf{D}\mathbf{W}\mathbf{H}\|_F^2+\lambda\|\mathbf{W}-\mathbf{S}\mathbf{M}\|_F^2 \\ & +\eta\|\mathbf{H}\|_F^2+\mu\|\mathbf{W}-\mathbf{Z}\|_F^2 \\ \text{s.t.} \quad & \mathbf{W}\geq\mathbf{0},\ \mathbf{H}\geq\mathbf{0},\ \|\mathbf{Z}(i,:)\|_2=1,\ \forall i, \\ & \mathbf{D}=\mathrm{Diag}(d_1,\ldots,d_I), \\ & \|\mathbf{S}(i,:)\|_0=1,\ \mathbf{S}(i,k)\in\{0,1\},\ \forall i,k,\end{aligned} \tag{5}$$

where $\mu \geq 0$ and $\mathbf{Z}$ is a slack variable. Note that $\mathbf{Z}$ is introduced to 'split' the effort of dealing with $\mathbf{W}\geq\mathbf{0}$ and $\|\mathbf{W}(i,:)\|_2=1$ in two different subproblems. Notice that when $\mu=+\infty$, (5) is equivalent to (3); in practice, a large μ can be employed to enforce $\mathbf{W}\approx\mathbf{Z}$.

Problem (5) can be handled as follows. First, $\mathbf{W}$ can be updated by solving

$$\begin{aligned}\mathbf{W}:=\arg\min_{\mathbf{W}\geq\mathbf{0}} \quad & \|\mathbf{X}-\mathbf{D}\mathbf{W}\mathbf{H}\|_F^2+\lambda\|\mathbf{W}-\mathbf{S}\mathbf{M}\|_F^2 \\ & +\mu\|\mathbf{W}-\mathbf{Z}\|_F^2,\end{aligned}$$

which can be easily converted to a nonnegative least squares (NLS) problem, and solved to optimality. The update of $\mathbf{H}$, i.e.,

$$\mathbf{H}:=\arg\min_{\mathbf{H}\geq\mathbf{0}} \ \|\mathbf{X}-\mathbf{D}\mathbf{W}\mathbf{H}\|_F^2+\eta\|\mathbf{H}\|_F^2,$$

is also an NLS problem. The subproblem w.r.t. d_i for $i=1,\ldots,I$ can be written as

$$d_i:=\arg\min_{d_i} \ \|\mathbf{X}(i,:)-d_i\mathbf{W}(i,:)\mathbf{H}\|_2^2,$$

which admits a simple closed-form solution, i.e., $d_i=\mathbf{X}(i,:)\mathbf{b}_i/(\mathbf{b}_i^T\mathbf{b}_i)$, where $\mathbf{b}_i^T=\mathbf{W}(i,:)\mathbf{H}$. To update $\mathbf{Z}$, we aim at solving the following projection problem:

$$\mathbf{Z}:=\arg\min_{\|\mathbf{Z}(:,f)\|_2=1,\ \forall f}\|\mathbf{Z}-\mathbf{W}\|_F^2,$$

which can be done by normalization: $\mathbf{Z}(:,f):=\frac{\mathbf{W}(:,f)}{\|\mathbf{W}(:,f)\|_2}$, for $f=1,\ldots,F$. Finally, the update w.r.t. $\mathbf{M}$ and $\mathbf{S}$ can be carried out by applying the K-means algorithm. We update each of $\mathbf{W}$, $\mathbf{H}$, $\mathbf{Z}$, $\{d_i\}_{i=1}^I$ and $(\mathbf{S},\mathbf{M})$ cyclically. Since each update does not increase the cost function, the value of the cost function will eventually converge. We should mention that, when applying the described algorithmic structure to solving Problem (4), the only change is that an additional partial minimization using NLS is needed, since the tensor model has three latent factors.

IV. Numerical Results

In this section, we use both synthetic and real data to showcase the effectiveness of the proposed approach. We generate a three-way tensor $\underline{\mathbf{X}}\in\mathbb{R}^{L\times M\times N}$ with $L=M=N=30$ and loading factors $\mathbf{A}\in\mathbb{R}^{L\times F}$, $\mathbf{B}\in\mathbb{R}^{M\times F}$, $\mathbf{C}\in\mathbb{R}^{N\times F}$. To obtain $\mathbf{A}$ with a cluster structure on its rows, we first generate a matrix $\tilde{\mathbf{A}}(i,:)$ for $i=1,\ldots,I$ from K structures by letting $\tilde{\mathbf{A}}(i,:)=\mathbf{M}(k,:)+10^{-2}\mathbf{N}(i,:)$ if $\mathrm{mod}(i,K)=k$, where the elements of $\mathbf{N}$ are drawn from the zero-mean i.i.d. normal distribution, and $\mathbf{M}(k,:)=2\mathbf{e}_k^T+\mathbf{1}^T$ for $k=1,\ldots,K$, in which $\mathbf{e}_k$ denotes the unit vector with the kth elements being one. By the above, the rows of $\tilde{\mathbf{A}}$ randomly scatter around the rows of $\mathbf{M}$. We then let $\mathbf{A}=\mathbf{D}\tilde{\mathbf{A}}$, where $\mathbf{D}$ is a diagonal matrix whose diagonal elements are uniformly distributed between zero and three. Note that we deliberately multiply $\mathbf{D}$ to $\tilde{\mathbf{A}}$ so that the rows of $\mathbf{A}$ belonging to the same clusters could have different scalings, which is usually the case in practice. $\mathbf{B}$ and $\mathbf{C}$ are randomly drawn from an i.i.d. uniform distribution between zero and one. Nonnegative noise following the same uniform distribution is added to the obtained tensor. To create more severe modeling error so that the situation is more realistic, we finally replace eight slabs (i.e., $\underline{\mathbf{X}}(:,:,i)$'s) with elements uniformly distributed between zero and one; these slabs mimic outlying data that are commonly seen in data analytics.

We apply the tensor version of the formulation in (5) to factor the synthesized tensors for $F=K=5$. We run 100 independent trials with different randomly generated tensors. Fig. 1 shows the averaged mean-squared-error (MSE) between $\mathbf{A}$ and its estimate $\hat{\mathbf{A}}$ under various λ when μ and η are fixed to 100 and 10^{-3}, respectively. The plain NTF algorithm (without latent clustering) is employed as a baseline. Here

$$\mathrm{MSE}=\min_{\substack{\boldsymbol{\pi}\in\Pi,\\ c_1,\ldots,c_F\in\{\pm1\}}}\frac{1}{F}\sum_{f=1}^F\left\|\frac{\mathbf{A}(:,f)}{\|\mathbf{A}(:,f)\|_2}-c_f\frac{\hat{\mathbf{A}}(:,\pi_f)}{\|\hat{\mathbf{A}}(:,\pi_f)\|_2}\right\|_2^2,$$

where π_f denotes the matched index with f and c_f is introduced for fixing the inherent scaling ambiguity. One can see that the proposed approach consistently yields lower MSEs for $\hat{\mathbf{A}}$ than plain NTF, for all λ, and the difference is 4 dB for $\lambda\geq 5\times10^3$. Similar results can be seen in Fig. 2, where we fix $\lambda=6\times10^3$ and vary F. The advantage of the proposed method is even more obvious when dealing with real data. Here, we present experimental results using the Reuters document corpus[1]. We use a subset of the full corpus as in [7], which contains 8,213 documents from 41 clusters. Following standard pre-processing, the stop words are removed, each document is represented as a term-frequency-inverse-document-frequency (tf-idf) vector, and *normalized cut weighting* is applied; see [14], [15] for details. We apply the formulation in (5) to the pre-processed data, and we use the obtained $\mathbf{S}$ to indicate the cluster labels of the documents. A regularized NMF-based approach, namely, *locally consistent*

[1]Available online: http://www.daviddlewis.com/resources/testcollections/reuters21578/

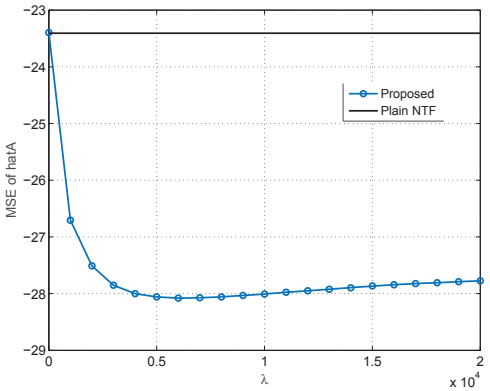

Fig. 1. The MSEs of the algorithms under various λ; $F = K = 5$.

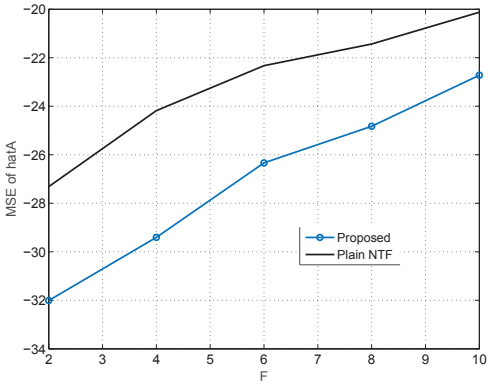

Fig. 2. The MSEs of the algorithms under various F; $F = K$.

concept factorization (LCCF) [7] is employed as the baseline. LCCF is considered state-of-the-art for clustering the Reuters corpus; it makes use of data-domain similarity to enforce latent similarity, and it demonstrates superior performance compared to other algorithms on several document clustering tasks. We apply the proposed approach and LCCF on the Reuters data for various K (number of clusters) and use $F = K$ for both methods. For each K, we perform 50 Monte-Carlo trials by randomly selecting K clusters out of the total 41 clusters, and report the performance by comparing the results with the ground truth. Performance is measured by a commonly used metric called *clustering accuracy*, whose detailed definition can be found in [7]. Simply speaking, the clustering accuracy ranges from 0 to 1, and higher accuracies indicate better performances. Table I presents the results averaged from the 50 trials. The proposed approach consistently exhibits higher accuracy than LCCF, and for $K \geq 7$ we get about 10% improvement, which is significant.

V. Conclusion

We proposed a simultaneous factor analysis and latent clustering framework, which can be applied for latent data mining of matrix and tensor data. The idea is to make use of the cluster structure in the latent space to help dimensionality reduction, and vice-versa. A variable-splitting based alternating optimization algorithm was derived to deal with the proposed problem formulations. Both simulations and real data experiments on the Reuters documents corpus showed that the proposed approach is effective in enhancing the performance of NMF and NTF in critical situations, e.g., when the data contains severe modeling errors. The proposed framework can also be potentially combined with other factor analysis models, which will be discussed in the follow-up journal version.

TABLE I
Clustering accuracies of the algorithms on the Reuters text corpus under various number of clusters.

K	2	3	4	5	6
LCCF [7]	0.888	0.839	0.811	0.74	0.742
Proposed	**0.930**	**0.858**	**0.851**	**0.781**	**0.777**
K	7	8	9	10	–
LCCF [7]	0.713	0.682	0.649	0.633	–
Proposed	**0.770**	**0.751**	**0.744**	**0.686**	–

References

[1] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 211–224, 2014.

[2] N. D. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of n-way arrays," *Journal of chemometrics*, vol. 14, no. 3, pp. 229–239, 2000.

[3] E. E. Papalexakis, N. D. Sidiropoulos, and R. Bro, "From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 493–506, 2013.

[4] A. Zymnis, S. Kim, J. Skaf, M. Parente, and S. Boyd, "Hyperspectral image unmixing via alternating projected subgradients," in *Proc. Asilomar 2007*, pp. 1164–1168, Nov. 2007.

[5] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, 2011.

[6] B. W. Bader, R. A. Harshman, and T. G. Kolda, "Temporal analysis of social networks using three-way dedicom," *Sandia National Laboratories TR SAND2006-2161*, vol. 119, 2006.

[7] D. Cai, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 6, pp. 902–913, 2011.

[8] S. Gao, I. W. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely–laplacian sparse coding for image classification," in *Proc. IEEE CVPR 2010*, pp. 3555–3561, 2010.

[9] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pp. 58–64, 2000.

[10] A. Huang, "Similarity measures for text document clustering," in *Proc. the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pp. 49–56, 2008.

[11] R. Bro, "PARAFAC. tutorial and applications," *Chemometrics and intelligent laboratory systems*, vol. 38, no. 2, pp. 149–171, 1997.

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

[13] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[14] W. Xu and Y. Gong, "Document clustering by concept factorization," in *Proc. 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 202–209, 2004.

[15] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.