Projective Nonnegative Matrix Factorization with α-Divergence

Zhirong Yang and Erkki Oja

Department of Information and Computer Science*
Helsinki University of Technology
P.O. Box 5400, FI-02015, TKK, Espoo, Finland
{zhirong.yang,erkki.oja}@tkk.fi

Abstract. A new matrix factorization algorithm which combines two recently proposed nonnegative learning techniques is presented. Our new algorithm, α-PNMF, inherits the advantages of Projective Nonnegative Matrix Factorization (PNMF) for learning a highly orthogonal factor matrix. When the Kullback-Leibler (KL) divergence is generalized to α-divergence, it gives our method more flexibility in approximation. We provide multiplicative update rules for α-PNMF and present their convergence proof. The resulting algorithm is empirically verified to give a good solution by using a variety of real-world datasets. For feature extraction, α-PNMF is able to learn highly sparse and localized part-based representations of facial images. For clustering, the new method is also advantageous over Nonnegative Matrix Factorization with α-divergence and ordinary PNMF in terms of higher purity and smaller entropy.

1 Introduction

Nonnegative learning based on matrix factorization has received a lot of research attention recently. The first application of *Nonnegative Matrix Factorization* (NMF) [1] was in extracting sparse features of facial images, while recent research also reveals its usefulness in clustering.

However, the original NMF approximation is restricted to least square errors or the Kullback-Leibler divergence between the data matrix and its approximation. It has recently been pointed out that the divergence minimization can be generalized by using the α-divergence [2], which leads to a family of new algorithms [3]. The empirical study by Cichocki et al. shows that the generalized NMF can achieve better performance for various applications by using proper α values.

Projective Nonnegative Matrix Factorization (PNMF) [4] is another variant of NMF. It identifies a nonnegative subspace by integrating the nonnegativity to the PCA objective. PNMF has proven to outperform NMF in feature extraction, where PNMF is able to generate sparser patterns which are more localized and

* Supported by the Academy of Finland in the project *Finnish Centre of Excellence in Adaptive Informatics Research.*

C. Alippi et al. (Eds.): ICANN 2009, Part I, LNCS 5768, pp. 20–29, 2009.

non-overlapping [4]. Clustering results of text data also demonstrate that PNMF is advantageous as it provides better approximation to the binary-valued multi-cluster indicators than NMF.

To achieve both merits of the above methods, we extend the PNMF by using α-divergence instead of KL-divergence as the error measure. We derive the multiplicative update rules for the new learning objective. The convergence of the iterative updates is proven using the Lagrangian approach. Experiments are conducted, in which the new algorithm outperforms α-NMF for extracting sparse and localized part-based representations of facial images. Our method can also achieve better clustering results than α-NMF and ordinary PNMF for a variety of datasets.

2 Related Work

2.1 Nonnegative Matrix Factorization

Given a nonnegative data matrix $\mathbf{X} \in \mathbb{R}_+^{m \times N}$, *Nonnegative Matrix Factorization* (NMF) seeks a decomposition of $\mathbf{X}$ that is of the form:

$$\mathbf{X} \approx \mathbf{WH}, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times N}$ with the rank $r \ll \min(m, N)$.

Denote by $\hat{\mathbf{X}} = \mathbf{WH}$ the approximating matrix. The approximation can be achieved by minimizing two widely used measures: (1) Least Square criterion $\varepsilon = \sum_{i,j} \left(X_{ij} - \hat{X}_{ij} \right)^2$ and (2) *Kullback-Leibler divergence* (KL-divergence)

$$D_{\mathrm{KL}} \left(\mathbf{X} || \hat{\mathbf{X}} \right) = \sum_{i,j} \left(X_{ij} \log \frac{X_{ij}}{\hat{X}_{ij}} - X_{ij} + \hat{X}_{ij} \right). \tag{2}$$

In this paper we focus on the second approximation criterion, which leads to the multiplicative updating rules of the form

$$H_{kj}^{\mathrm{new}} = H_{kj} \frac{\left(\mathbf{W}^T \mathbf{Z} \right)_{kj}}{\sum_i W_{ik}}, \quad W_{ik}^{\mathrm{new}} = W_{ik} \frac{\left(\mathbf{Z}\mathbf{H}^T \right)_{ik}}{\sum_j H_{kj}}, \tag{3}$$

where we use $Z_{ij} = X_{ij} / \hat{X}_{ij}$ for notational brevity.

2.2 Nonnegative Matrix Factorization with α-Divergence

The α-divergence [2] is a parametric family of divergence functionals, including several well-known divergence measures as special cases. NMF equipped with the following α-divergence as the approximation measure is called α-NMF [3].

$$D_\alpha \left(\mathbf{X} || \hat{\mathbf{X}} \right) = \frac{1}{\alpha(1-\alpha)} \sum_{i=1}^{m} \sum_{j=1}^{N} \left(\alpha X_{ij} + (1-\alpha) \hat{X}_{ij} - X_{ij}^{\alpha} \hat{X}_{ij}^{1-\alpha} \right) \tag{4}$$

The corresponding multiplicative update rules are given by the following, where we define $\tilde{Z}_{ij} = Z_{ij}^{\alpha}$:

$$H_{kj}^{\text{new}} = H_{kj} \left[\frac{\left(\mathbf{W}^T \tilde{\mathbf{Z}}\right)_{kj}}{\sum_i W_{ik}} \right]^{\frac{1}{\alpha}}, \quad W_{ik}^{\text{new}} = W_{ik} \left[\frac{\left(\tilde{\mathbf{Z}} \mathbf{H}^T\right)_{ik}}{\sum_j H_{kj}} \right]^{\frac{1}{\alpha}}. \tag{5}$$

α-NMF reduces to the conventional NMF with KL-divergence when $\alpha \to 1$. Another choice of α characterizes a different learning principle, in the sense that the model distribution is more inclusive ($\alpha \to \infty$) or more exclusive ($\alpha \to -\infty$). Such flexibility enables α-NMF to outperform NMF with α properly selected.

2.3 Projective Nonnegative Matrix Factorization

Replacing $\mathbf{H} = \mathbf{W}^T\mathbf{X}$ in (1), we get the *Projective Nonnegative Matrix Factorization* (PNMF) approximation scheme [4]

$$\mathbf{X} \approx \mathbf{W}\mathbf{W}^T\mathbf{X}. \tag{6}$$

Again, denote $\hat{\mathbf{X}} = \mathbf{W}\mathbf{W}^T\mathbf{X}$ the approximating matrix and $Z_{ij} = X_{ij}/\hat{X}_{ij}$. The PNMF multiplicative update rule for KL-divergence is given by [4]

$$W_{ik}^{\text{new}} = W_{ik} \frac{\left(\mathbf{Z}\mathbf{X}^T\mathbf{W} + \mathbf{X}\mathbf{Z}^T\mathbf{W}\right)_{ik}}{\sum_j \left(\mathbf{W}^T\mathbf{X}\right)_{kj} + \left(\sum_j X_{ij}\right)\left(\sum_b W_{bk}\right)}. \tag{7}$$

The name PNMF comes from another derivation of the approximation scheme (6) where a projection matrix $\mathbf{P}$ in $\mathbf{X} \approx \mathbf{P}\mathbf{X}$ is factorized into $\mathbf{W}\mathbf{W}^T$. This interpretation connects PNMF with the classical *Principal Component Analysis* subspace method except for the nonnegativity constraint [4]. Compared with NMF, PNMF is able to learn a much sparser matrix $\mathbf{W}$. This property is especially desired for extracting part-based representations of data samples or finding cluster indicators.

3 PNMF with α-Divergence

In this section we combine the flexibility of α-NMF and the sparsity of PNMF into a single algorithm. We called the resulting method α-*PNMF* which stands for Projective Nonnegative Matrix Factorization with α-divergence.

3.1 Multiplicative Update Rule

α-PNMF solves the following optimization problem:

$$\underset{\mathbf{W} \geq \mathbf{0}}{\text{minimize}} \ \mathcal{J}(\mathbf{W}) = D_{\alpha}(\mathbf{X} || \mathbf{W}\mathbf{W}^T\mathbf{X}). \tag{8}$$

The derivative of the objective with respect to $\mathbf{W}$ is

$$\frac{\partial \mathcal{J}(\mathbf{W})}{\partial W_{ik}} = \frac{1}{\alpha}\Big[-\left(\tilde{\mathbf{Z}}\mathbf{X}^T\mathbf{W} + \mathbf{X}\tilde{\mathbf{Z}}^T\mathbf{W}\right)_{ik} + \sum_j \left(\mathbf{W}^T\mathbf{X}\right)_{kj} + \left(\sum_j X_{ij}\right)\left(\sum_b W_{bk}\right)\Big] \tag{9}$$

Denote Λ_{ik} the Lagrangian multipliers associated with the constraint $W_{ik} \geq 0$. The Karush-Kuhn-Tucker (KKT) conditions require

$$\frac{\partial \mathcal{J}(\mathbf{W})}{\partial W_{ik}} = \Lambda_{ik} \tag{10}$$

and $\Lambda_{ik}W_{ik} = 0$ which indicates $\Lambda_{ik}W_{ik}^{\alpha} = 0$. Multiplying both sides of (10) by W_{ik}^{α} leads to $\frac{\partial \mathcal{J}(\mathbf{W})}{\partial W_{ik}}W_{ik}^{\alpha} = 0$. This suggests a multiplicative update rule by writing $\tilde{Z}_{ij} = Z_{ij}^{\alpha}$:

$$W'_{ik} = W_{ik}\left[\frac{\left(\tilde{\mathbf{Z}}\mathbf{X}^T\mathbf{W} + \mathbf{X}\tilde{\mathbf{Z}}^T\mathbf{W}\right)_{ik}}{\sum_j \left(\mathbf{W}^T\mathbf{X}\right)_{kj} + \left(\sum_j X_{ij}\right)\left(\sum_a W_{ak}\right)}\right]^{\frac{1}{\alpha}}. \tag{11}$$

3.2 Convergence Proof

The convergence of NMF and most of its variants, including α-NMF, to a local minimum of the cost function is analyzed by using an auxiliary function [3]. It is however difficult to directly construct such a function for α-PNMF because of the auto-association induced by $\mathbf{W}$ and its transpose. Here we overcome this problem by applying the Lagrangian technique to decouple the auto-association.

With the constraint $\mathbf{H} = \mathbf{W}^T\mathbf{X}$, one can write the Lagrangian objective function as

$$\mathcal{L}(\mathbf{W}, \mathbf{H}) = D_{\alpha}(\mathbf{X}||\mathbf{W}\mathbf{H}) + \mathrm{Tr}\left(\boldsymbol{\Psi}^T\left(\mathbf{H} - \mathbf{W}^T\mathbf{X}\right)\right), \tag{12}$$

by introducing multipliers Ψ_{ik}. Following [3], we apply Jensen's inequality using the convex function $f(z) = [\alpha + (1-\alpha)z - z^{1-\alpha}]/(\alpha(1-\alpha))$ for $\alpha \geq 0$, which leads to

$$f\left(\sum_k W_{ik}H_{kj}\right) \leq \sum_k \zeta_{ijk} f\left(\frac{W_{ik}H_{kj}}{X_{ij}\zeta_{ijk}}\right), \tag{13}$$

with $\zeta_{ijk} = \frac{W_{ik}H_{kj}}{\sum_l W_{il}H_{lj}}$. After some manipulation, one can find that $\mathcal{L}(\mathbf{W}, \mathbf{H})$ is upper-bounded by the auxiliary function

$$G(\mathbf{W}', \mathbf{W}) = \frac{1}{\alpha} \sum_{i,j,k} X_{ij} \zeta_{ijk} \left[\alpha + (1-\alpha) \frac{W'_{ik} H_{kj}}{X_{ij} \zeta_{ijk}} - \left(\frac{W'_{ik} H_{kj}}{X_{ij} \zeta_{ijk}} \right)^{1-\alpha} \right] + \mathrm{Tr}\left(\boldsymbol{\Psi}^T \left(\mathbf{H} - \mathbf{W}^T \mathbf{X} \right) \right) + \sum_{ik} \left(\mathbf{X} \tilde{\mathbf{Z}}^T \mathbf{W} \right)_{ik} \left(W'_{ik} - W_{ik} - W_{ik} \log \frac{W'_{ik}}{W_{ik}} \right). \tag{14}$$

The last line of eq. (14) is a tight upper-bound of zero. To see this, one can insert $y = W'_{ik}/W_{ik}$ into the inequality $y \geq 1 + \log y$ for $y \geq 0$, where the equality holds if and only if $y = 1$. This additional bounding aims to add the same term $(\mathbf{X}\tilde{\mathbf{Z}}^T\mathbf{W})_{ik}$ to both numerator and denominator of the resulting multiplicative update rule and thus maintains the nonnegativity of $\mathbf{W}$.

Setting $\partial G / \partial \mathbf{W}' = 0$, we get

$$\left(\frac{W'_{ik}}{W_{ik}} \right)^{\alpha} = \frac{\left(\tilde{\mathbf{Z}} \mathbf{H}^T \right)_{ik} + \left(\mathbf{X} \tilde{\mathbf{Z}}^T \mathbf{W} \right)_{ik}}{\sum_j H_{kj} - \alpha \left(\mathbf{X} \boldsymbol{\Psi}^T \right)_{ik} + \left(\mathbf{X} \tilde{\mathbf{Z}}^T \mathbf{W} \right)_{ik}}. \tag{15}$$

Next we solve $\boldsymbol{\Psi}$ by using the KKT conditions. From

$$\frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{H})}{\partial H_{kj}} = \frac{1}{\alpha} \left(\sum_i W_{ik} - \left(\mathbf{W}^T \tilde{\mathbf{Z}} \right)_{kj} \right) + \Psi_{kj} = 0 \tag{16}$$

we get

$$\alpha \left(\mathbf{X} \boldsymbol{\Psi}^T \right)_{ik} = \left(\mathbf{X} \tilde{\mathbf{Z}}^T \mathbf{W} \right)_{ik} - \left(\sum_j X_{ij} \right) \left(\sum_b W_{bk} \right) \tag{17}$$

Inserting (17) and $\mathbf{H} = \mathbf{W}^T\mathbf{X}$ into (15), one obtains the multiplicative update rule (11). This concludes our proof of the following result:

Theorem 1. *$D_\alpha(\mathbf{X}||\mathbf{W}\mathbf{W}^T\mathbf{X})$ is non-increasing under the multiplicative updates using (11).*

3.3 Stabilization

The multiplicative updates can start from any initial guess of $\mathbf{W}$. However, we find some initial values may lead to a very zigzag convergence path. The overall scaling of $\mathbf{W}$ greatly fluctuates between odd and even iterations.

We propose to overcome this problem by introducing one more parameter ρ. The modified objective becomes to minimize $\tilde{\mathcal{J}}(\rho, \mathbf{W}) = D_\alpha(\mathbf{X}||\rho\mathbf{W}\mathbf{W}^T\mathbf{X})$. Fixing $\mathbf{W}$, the global optimal ρ^* can be solved by setting the derivative of $\tilde{\mathcal{J}}(\rho, \mathbf{W})$ with respect to ρ to zero, which results in

$$\rho^* = \left(\frac{\sum_{ij} \hat{X}_{ij} \tilde{Z}_{ij}}{\sum_{ij} \hat{X}_{ij}} \right)^{\frac{1}{\alpha}} \tag{18}$$

Next, fixing ρ the optimal $\mathbf{W}$ given its current estimate can be found by inserting ρ^* in the denominator of (11). Equivalently, one can apply the original multiplicative update rule and then compute

$$W_{ik}^{\text{new}} = W'_{ik} \left(\frac{\sum_{ij} \hat{X}_{ij} \tilde{Z}_{ij}}{\sum_{ij} \hat{X}_{ij}} \right)^{\frac{1}{2\alpha}} \tag{19}$$

with re-calculated $\hat{\mathbf{X}}$ and $\tilde{\mathbf{Z}}$.

If $\mathbf{W}\mathbf{W}^T\mathbf{X}$ approximates $\mathbf{X}$ well, all the $\tilde{Z}_{ij}$ approach one and so does ρ^*. The modified objective is thus equivalent to the original one. Therefore ρ serves as an intermediate variable that stabilizes and speeds up the algorithm especially in early iterations.

4 Experiments

Suppose the nonnegative matrix $\mathbf{X} \in \mathbb{R}_+^{m \times N}$ is composed of N data samples $\mathbf{x}_j \in \mathbb{R}_+^m$, $j = 1, \ldots, N$. Basically, α-PNMF can be applied on this matrix in two different ways. One employs the approximation scheme $\mathbf{X} \approx \mathbf{W}\mathbf{W}^T\mathbf{X}$ and performs *feature extraction* by projecting each sample into a nonnegative subspace. The other approach approximates the transposed matrix $\mathbf{X}^T$ by $\mathbf{W}\mathbf{W}^T\mathbf{X}^T$ where $\mathbf{W} \in \mathbb{R}_+^{N \times r}$. The latter approach can be used for *clustering* where the elements of $\mathbf{W}$ indicate the membership of each sample to the r clusters.

4.1 Feature Extraction

We have used the FERET database of facial images [5] as the training data set. After face segmentation, 2,409 frontal images (poses "fa" and "fb") of 867 subjects were stored in the database for the experiments. All face boxes were normalized to the size of 32×32 and then reshaped to a 1024-dimensional vector by column-wise concatenation. Thus we obtained a 1024×2409 nonnegative data matrix, whose elements are re-scaled into the region [0,1] by dividing with their maximum. For good visualization, we empirically set $r = 25$ in the feature extraction experiments.

After training, the basis vectors are stored in the columns of $\mathbf{W}$ in α-NMF and α-PNMF. The basis vectors have same dimensionality with the image samples and thus can be visualized as *basis images*. In order to encode the features of different facial parts, it is expected to find some localized and non-overlapping patterns in the basis images. The resulting basis images using $\alpha = 0.5$ (Hellinger divergence), $\alpha = 1$ (KL-divergence) and $\alpha = 2$ (χ^2-divergence) are shown in Figure 1. Both methods can identify some facial parts such as eyebrows and lips. In comparison, α-PNMF is able to generate much sparser basis images with more part-based visual patterns.

Notice that two non-negative vectors are orthogonal if and only if they do not have the same non-zero dimensions. Therefore we can quantify the sparsity of the basis vectors by measuring their orthogonalities with the τ measurement:

$$\tau = 1 - \frac{\|\mathbf{R} - \mathbf{I}\|_F}{(r(r-1))}, \tag{20}$$

where $\|\cdot\|_F$ is the Frobenius matrix norm and R_{st} the normalized inner product between two basis vectors $\mathbf{w}_s$ and $\mathbf{w}_t$:

$$R_{st} = \frac{\mathbf{w}_s^T \mathbf{w}_t}{\|\mathbf{w}_s\| \|\mathbf{w}_t\|}. \tag{21}$$

Larger τ's indicate higher orthogonality and τ reaches 1 when the columns of $\mathbf{W}$ are completely orthogonal. The orthogonalities using the two compared methods are displayed under the respective basis image plots in Figure 1. All τ values in the right are larger than their left counterparts, which confirms that α-PNMF is able to extract a sparser transformation matrix $\mathbf{W}$. It is worth to notice that α-PNMF achieves the high sparseness without the explicit orthogonality constraint compared with some other exiting methods such as [6].

4.2 Clustering

We have used a variety of datasets, most of which are frequently used in machine learning and information retrieval research. Table 1 summarizes the characteristics of the datasets. The descriptions of these datasets are as follows:

- *Iris*, *Ecoli5*, *WDBC*, and *Pima*, which are taken from the UCI data repository with respective datasets Iris, Ecoli, Breast Cancer Wisconsin (Prognostic), and Pima Indians Diabetes. The *Ecoli5* dataset contains only samples of the five largest classes in the original Ecoli database.
- *AMLALL* gene expression database [7]. This dataset contains acute lymphoblastic leukemia (ALL) that has B and T cell subtypes, and acute myelogenous leukemia (AML) that occurs more commonly in adults than in children. The data matrix consists of 38 bone marrow samples (19 ALL-B, 8 ALL-T and 11 AML) with 5000 genes as their dimensions.
- *ORL* database of facial images [8]. There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions and facial details. In our experiments, we down-sampled the images to size 46×56 and rescaled the gray-scale values to $[0, 1]$.

The number of clusters r is generally set to the number of classes. This work focuses on cases where $r > 2$, as there exist closed form approximations for the two-way clustering solution (see e.g. [9]). We thus set r equal to five times of the number of classes for *WDBC* and *Pima*.

Suppose there is ground truth data that labels the samples by one of q classes. We have used the *purity* and *entropy* measures to quantify the performance of the compared clustering algorithms:

$$\text{purity} = \frac{1}{N} \sum_{k=1}^{r} \max_{1 \le l \le q} n_k^l, \quad \text{entropy} = -\frac{1}{n \log_2 q} \sum_{k=1}^{r} \sum_{l=1}^{q} n_k^l \log_2 \frac{n_k^l}{n_k},$$

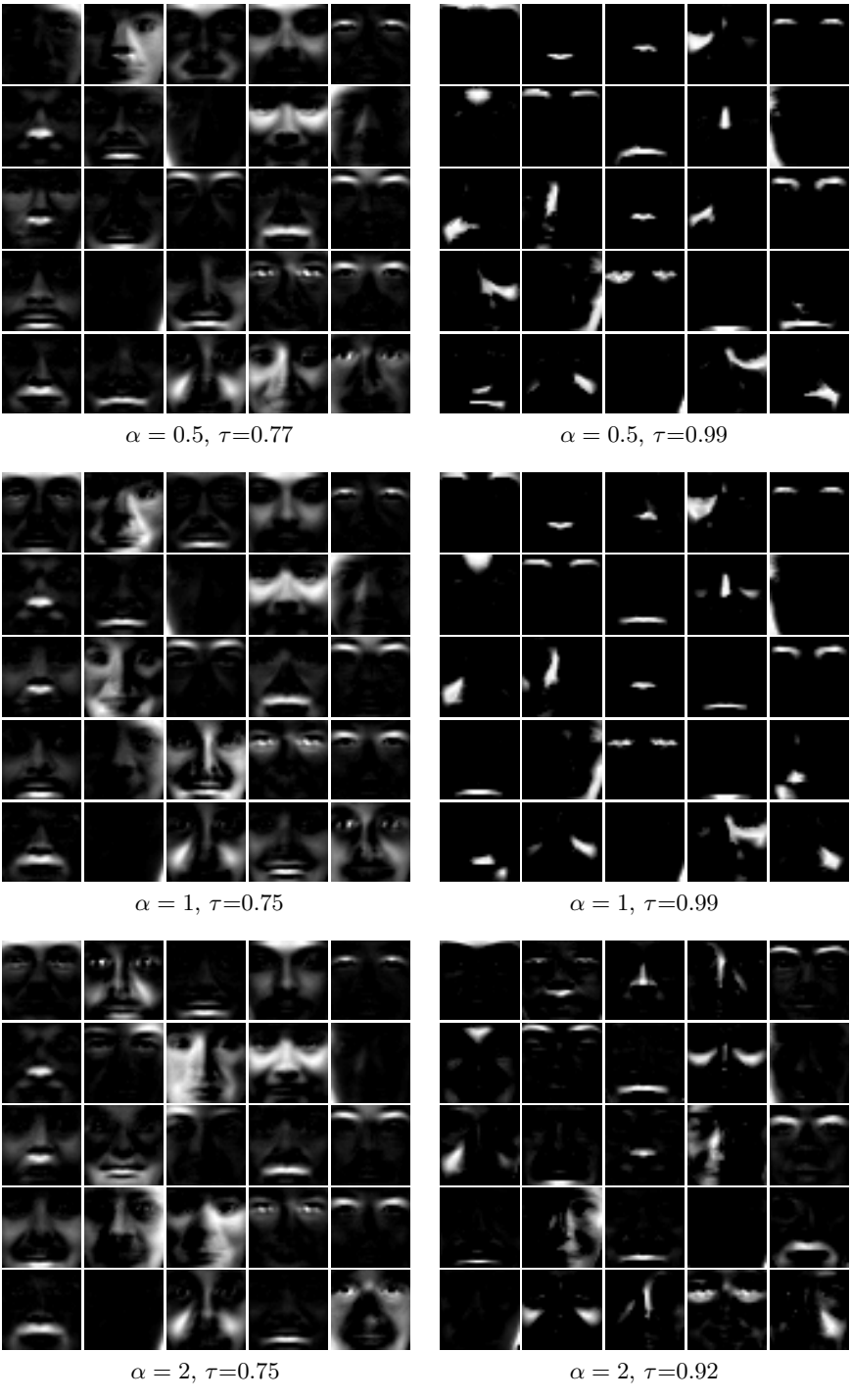

Fig. 1. The basis images of (left) α-NMF and (right) α-PNMF

Table 1. Dataset descriptions

datasets	#samples	#dimensions	#classes	r
Iris	150	4	3	3
Ecoli5	327	7	5	5
WDBC	569	30	2	10
Pima	768	8	2	10
AMLALL	38	5000	3	3
ORL	400	2576	40	40

Table 2. Clustering (a) purities and (b) entropies using α-NMF, PNMF and α-PNMF. The best result for each dataset is highlighted with boldface font.

(a)

	α-NMF			PNMF	α-PNMF		
datasets	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$	-	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
Iris	0.83	0.85	0.84	0.95	0.95	0.95	**0.97**
Ecoli5	0.62	0.65	0.67	0.72	0.72	0.72	**0.73**
WDBC	0.70	0.70	0.72	0.87	0.86	0.87	**0.88**
Pima	0.65	0.65	0.65	0.65	0.67	0.65	**0.67**
AMLALL	0.95	0.92	0.92	0.95	**0.97**	0.95	0.92
ORL	0.47	0.47	0.47	0.75	0.76	0.75	**0.80**

(b)

	α-NMF			PNMF	α-PNMF		
datasets	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$	-	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
Iris	0.34	0.33	0.33	0.15	0.15	0.15	**0.12**
Ecoli5	0.46	0.58	0.50	0.40	0.40	0.40	**0.40**
WDBC	0.39	0.38	0.37	0.16	0.17	0.16	**0.14**
Pima	0.92	0.90	0.90	0.91	0.90	0.91	**0.89**
AMLALL	0.16	0.21	0.21	0.16	**0.08**	0.16	0.21
ORL	0.35	0.34	0.35	0.14	0.14	0.14	**0.12**

where n_k^l is the number of samples in the cluster k that belong to original class l and $n_k = \sum_l n_k^l$. A larger purity value and a smaller entropy indicate better clustering performance.

The resulting purities and entropies are shown in Table 2, respectively. α-PNMF performs the best for all selected datasets. Recall that when $\alpha = 1$ the proposed method reduces to PNMF and thus returns results identical to the latter. Nevertheless, α-PNMF can outperform PNMF by adjusting the α value. When $\alpha = 0.5$, the new method achieves the highest purity and lowest entropy for the gene expression dataset *AMLALL*. For the other five datasets, one can set $\alpha = 2$ and obtain the best clustering result using α-PNMF. In addition, one can see that Nonnegative Matrix Factorization with α-divergence works poorly in our clustering experiments, much worse than the other methods. This

is probably because α-NMF has to estimate many more parameters than those using projective factorization. α-NMF is therefore prone to falling into bad local optima.

5 Conclusions

We have presented a new variant of NMF by introducing the α-divergence into the PNMF algorithm. Our α-PNMF algorithm theoretically converges to a local minimum. The resulting factor matrix is of high sparsity or orthogonality, which is desired for part-based feature extraction and multi-way clustering. Experimental results with various datasets indicate that the proposed algorithm can be considered as a promising replacement for α-NMF and PNMF for feature extraction and clustering.

References

1. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
2. Amari, S.: Differential-geometrical methods in statistics. Lecture Notes in Statistics, vol. 28. Springer, New York (1985)
3. Cichocki, A., Lee, H., Kim, Y.D., Choi, S.: Non-negative matrix factorization with α-divergence. Pattern Recognition Letters 29, 1433–1440 (2008)
4. Yuan, Z., Oja, E.: Projective nonnegative matrix factorization for image compression and feature extraction. In: Kalviainen, H., Parkkinen, J., Kaarna, A. (eds.) SCIA 2005. LNCS, vol. 3540, pp. 333–342. Springer, Heidelberg (2005)
5. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Analysis and Machine Intelligence 22, 1090–1104 (2000)
6. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 126–135 (2006)
7. Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences 101(12), 4164–4169 (2004)
8. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. In: Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota FL, December 1994, pp. 138–142 (1994)
9. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)