JAISCR, 2011, Vol.1, No.1, pp. 7–16

PROJECTIVE NONNEGATIVE MATRIX FACTORIZATION BASED ON α-DIVERGENCE

Zhirong Yang and Erkki Oja
Department of Information and Computer Science
Aalto University School of Science and Technology
P.O.Box 15400, FI-00076, Aalto, Finland

Abstract

The well-known Nonnegative Matrix Factorization (NMF) method can be provided with more flexibility by generalizing the non-normalized Kullback-Leibler divergence to α-divergences. However, the resulting α-NMF method can only achieve mediocre sparsity for the factorizing matrices. We have earlier proposed a variant of NMF, called Projective NMF (PNMF) that has been shown to have superior sparsity over standard NMF. Here we propose to incorporate both merits of α-NMF and PNMF. Our α-PNMF method can produce a much sparser factorizing matrix, which is desired in many scenarios. Theoretically, we provide a rigorous convergence proof that the iterative updates of α-PNMF monotonically decrease the α-divergence between the input matrix and its approximate. Empirically, the advantages of α-PNMF are verified in two application scenarios: (1) it is able to learn highly sparse and localized part-based representations of facial images; (2) it outperforms α-NMF and PNMF for clustering in terms of higher purity and smaller entropy.

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

1 Introduction

Nonnegative learning based on matrix factorization has received a lot of research attention recently. After Lee and Seung [11, 12] presented their *Nonnegative Matrix Factorization* (NMF) algorithms, a multitude of NMF variants have been proposed and applied to many areas such as signal processing, data mining, pattern recognition and gene expression studies [3, 5, 6, 9, 14, 21]. NMF is not only applicable to the feature axis for finding sparse and part-based representations (e.g.[10, 13]), but also to the sample axis, e.g. for finding clusters of data items (e.g. [8, 7, 19]).

The original NMF algorithm minimizes one of two kinds of difference measure between the data matrix and its approximate: the least square error or the non-normalized Kullback-Leibler divergence (or I-divergence). When the latter is used, NMF actually maximizes the Poisson likelihood of the observed data [11]. It was recently pointed out that the divergence minimization can be generalized by using the α-divergence [1], which leads to a family of new algorithms [4, 23]. The convergence proof of NMF with α-divergence is given in [4]. The empirical study by Cichocki et al. shows that the generalized NMF can achieve better performance by using suitable α values.

Projective Nonnegative Matrix Factorization (PNMF) [22] 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 non-overlapping [22]. 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 [19].

In this paper we combine the above two techniques by using α-divergence instead of I-divergence as the error measure in PNMF. We provide a multiplicative optimization algorithm which is theoretically convergent. Experiments are conducted, in which the new algorithm is shown to outperform α-NMF for feature extraction and clustering on a variety of datasets.

Part of the work can be found in our preliminary paper [18]. As an extension, we propose here a novel multiplicative update rule which monotonically decreases the α-divergence between the data matrix and its approximate, without additional normalization or stabilization steps. The new algorithm is more desirable because it makes the objectives at different iterations and with different initial guesses comparable. The proof uses a novel convex function for α-divergence which has not been used in the previous literature on divergence measures. We also provide the multiplicative update rule for the special case $\alpha \to 0$, which completes these algorithms for the entire family of α-divergences.

The rest of the paper is organized as follows. We first briefly review the NMF and PNMF methods in Section 2. In Section 3, we present the α-PNMF objective, its multiplicative optimization algorithm and convergence proof. The experiments are presented in Section 4, and Section 5 concludes the paper.

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 an approximative 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 $\widehat{\mathbf{X}} = \mathbf{WH}$ the approximating matrix. The approximation can be achieved by minimizing two widely used measures: (1) the least square criterion $\varepsilon = \sum_{i,j}\left(X_{ij} - \widehat{X}_{ij}\right)^2$ and (2) the non-normalized *Kullback-Leibler divergence* (or I-divergence)

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

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

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

$$W_{ik}^{\text{new}} = W_{ik}\frac{\left(\mathbf{ZH}^T\right)_{ik}}{\sum_j H_{kj}}, \tag{4}$$

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

2.2 Nonnegative Matrix Factorization with α-divergence

The α-divergence [1] 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 was introduced by Cichocki *et al* and called α-NMF [4]:

$$D_\alpha\left(\mathbf{X}||\widehat{\mathbf{X}}\right) = \frac{\sum_{ij}\left(\alpha X_{ij} + (1-\alpha)\widehat{X}_{ij} - X_{ij}^\alpha\widehat{X}_{ij}^{1-\alpha}\right)}{\alpha(1-\alpha)} \tag{5}$$

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

$$H_{kj}^{\text{new}} = H_{kj}\left[\frac{\left(\mathbf{W}^T\widetilde{\mathbf{Z}}\right)_{kj}}{\sum_i W_{ik}}\right]^{\frac{1}{\alpha}}, \tag{6}$$

$$W_{ik}^{\text{new}} = W_{ik}\left[\frac{\left(\widetilde{\mathbf{Z}}\mathbf{H}^T\right)_{ik}}{\sum_j H_{kj}}\right]^{\frac{1}{\alpha}}. \tag{7}$$

α-NMF reduces to the conventional NMF with I-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 [22]

$$\mathbf{X} \approx \mathbf{W}\mathbf{W}^T\mathbf{X}. \tag{8}$$

Denote $\widehat{\mathbf{X}} = \mathbf{W}\mathbf{W}^T\mathbf{X}$ the approximating matrix, $Z_{ij} = X_{ij}/\widehat{X}_{ij}$, and $\mathbf{1}_m$ a column vector of length m and filled with ones. The PNMF multiplicative update rule for I-divergence is given by [22]

$$W'_{ik} = W_{ik}\frac{(\mathbf{A}\mathbf{W})_{ik}}{(\mathbf{B}\mathbf{W})_{ik}} \tag{9}$$

where $\mathbf{A} = \mathbf{Z}\mathbf{X}^T + \mathbf{X}\mathbf{Z}^T$ and $\mathbf{B} = \mathbf{1}_m\mathbf{1}_n^T\mathbf{X}^T + \mathbf{X}\mathbf{1}_n\mathbf{1}_m^T$.

In practice, iterations with only the update rule (9) are sensitive to the initial guess of $\mathbf{W}$ and often have a very zigzag learning path, where the overall scaling of $\mathbf{W}$ fluctuates between odd and even iterations. This is overcome in practice by using an additional normalization step [22]

$$\mathbf{W}^{\text{new}} = \frac{\mathbf{W}'}{\|\mathbf{W}'\|} \tag{10}$$

or a stabilization step [19]

$$\mathbf{W}^{\text{new}} = \mathbf{W}'\sqrt{\frac{\sum_{ij}X_{ij}}{\sum_{ij}\left(\mathbf{W}'\mathbf{W}'^T\mathbf{X}\right)_{ij}}}. \tag{11}$$

The name PNMF comes from another derivation of the approximation scheme (8) 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 [22]. Compared with NMF, PNMF is able to learn a much sparser matrix $\mathbf{W}$ [19, 22, 23]. 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 call 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:

$$\underset{\mathbf{W}\geq\mathbf{0}}{\text{minimize}}\ \mathcal{J}(\mathbf{W}) = D_\alpha(\mathbf{X}||\mathbf{W}\mathbf{W}^T\mathbf{X}). \tag{12}$$

The gradient of the objective with respect to $\mathbf{W}$ is given by

$$\frac{\partial\mathcal{J}(\mathbf{W})}{\partial W_{ik}} = \frac{1}{\alpha}\left[-\left(\widetilde{\mathbf{A}}\mathbf{W}\right)_{ik} + (\mathbf{B}\mathbf{W})_{ik}\right],$$

where $\widetilde{Z}_{ij} = Z_{ij}^\alpha$, $\widetilde{\mathbf{A}} = \widetilde{\mathbf{Z}}\mathbf{X}^T + \mathbf{X}\widetilde{\mathbf{Z}}^T$ and again $\mathbf{B} = \mathbf{1}_m\mathbf{1}_n^T\mathbf{X}^T + \mathbf{X}\mathbf{1}_n\mathbf{1}_m^T$.

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

$$\frac{\partial\mathcal{J}(\mathbf{W})}{\partial W_{ik}} = \Lambda_{ik} \tag{13}$$

and $\Lambda_{ik}W_{ik} = 0$ which indicates $\Lambda_{ik}W_{ik}^{2\alpha} = 0$. Multiplying both sides of (13) by $W_{ik}^{2\alpha}$ leads to $\frac{\partial\mathcal{J}(\mathbf{W})}{\partial W_{ik}}W_{ik}^{2\alpha} = 0$. This suggests a multiplicative update rule:

$$W_{ik}^{\text{new}} = W_{ik}\left[\frac{\left(\widetilde{\mathbf{A}}\mathbf{W}\right)_{ik}}{(\mathbf{B}\mathbf{W})_{ik}}\right]^{\frac{1}{2\alpha}}. \tag{14}$$

for all $\alpha \neq 0$. For the special case $\alpha = 0$, the update rule is given by

$$W_{ik}^{\text{new}} = W_{ik}\exp\left(\frac{1}{2}\frac{\left(\widetilde{\mathbf{A}}^{(0)}\mathbf{W}\right)_{ik}}{(\mathbf{B}\mathbf{W})_{ik}}\right), \tag{15}$$

where $\widetilde{Z}_{ij}^{(0)} = \log Z_{ij}$ and $\widetilde{\mathbf{A}}^{(0)} = \widetilde{\mathbf{Z}}^{(0)}\mathbf{X}^T + \mathbf{X}\widetilde{\mathbf{Z}}^{(0)T}$.

3.2 Convergence Proof

In this Section, we prove that iteratively applying (14) or (15) monotonically decreases the objective function $D_\alpha(\mathbf{X}||\mathbf{W}\mathbf{W}^T\mathbf{X})$.

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 as its tight upper-bound. This is achieved in α-NMF [4] by using the Jensen inequality based on the convex function

$$h(z) = \frac{\alpha + (1-\alpha)z - z^{1-\alpha}}{\alpha(1-\alpha)}. \tag{16}$$

This convex function is however not applicable to the α-PNMF case because it is not decomposable, i.e. not fulfilling $h(xy) \propto h(x)h(y)$ or $h(xy) = h(x) + h(y) + \text{constant}$.

Here we overcome this problem by using a novel convex function

$$g(x,y) = -\frac{x^{\alpha}y^{1-\alpha}}{\alpha(1-\alpha)}. \tag{17}$$

We further introduce

$$f(y) = g\left(X_{ij}, y\right) \tag{18}$$

for notational brevity. Notice that $f(y)$ is convex with respect to y,

$$f(by) = b^{1-\alpha} f(y), \tag{19}$$

$$f(yz) = -\frac{\alpha(1-\alpha)}{X_{ij}^{\alpha}} f(y)f(z). \tag{20}$$

Let $\mathbf{W}$ be the current estimate, $\widehat{\mathbf{X}} = \widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T\mathbf{X}$, and

$$\gamma_{ijk} = \frac{W_{ik}\left(\mathbf{W}^T\mathbf{X}\right)_{kj}}{\sum_l W_{il}\left(\mathbf{W}^T\mathbf{X}\right)_{lj}} = \frac{W_{ik}\left(\mathbf{W}^T\mathbf{X}\right)_{kj}}{\left(\mathbf{W}\mathbf{W}^T\mathbf{X}\right)_{ij}}, \tag{21}$$

$$\beta_{ajk} = \frac{W_{ak}X_{aj}}{\sum_b W_{bk}X_{bj}} = \frac{W_{ak}X_{aj}}{\left(\mathbf{W}^T\mathbf{X}\right)_{kj}}, \tag{22}$$

$$\widetilde{\mathbf{V}} \equiv \widetilde{\mathbf{V}}(\widetilde{\mathbf{W}}, \mathbf{W}),\ \widetilde{V}_{ik} = \widetilde{W}_{ik}^{1-\alpha} W_{ik}^{\alpha} \tag{23}$$

$$S_{ij} = -\frac{1}{\alpha(1-\alpha)} \widetilde{\mathbf{Z}}^T\mathbf{X} \tag{24}$$

Obviously, $\gamma_{ijk} \geq 0$, $\sum_k \gamma_{ijk} = 1$, $\beta_{ajk} \geq 0$, $\sum_a \beta_{ajk} = 1$, $\mathbf{V} \equiv \widetilde{\mathbf{V}}(\mathbf{W}, \mathbf{W})$, and $V_{ik} = W_{ik}$.

In the derivation below we also employ the following inequality for any symmetric real matrix $\mathbf{M}$ independent of $\widetilde{\mathbf{W}}$ [8]:

$$\frac{1}{2}\mathrm{Tr}\left(\widetilde{\mathbf{W}}^T\mathbf{M}\widetilde{\mathbf{W}}\right) \leq \sum_{ik} \frac{\widetilde{W}_{ik}^2}{2W_{ik}} \left(\mathbf{M}\mathbf{W}\right)_{ik}, \tag{25}$$

where the equality holds if and only if $\widetilde{\mathbf{W}} = \mathbf{W}$.

We can then apply the Jensen inequality twice to obtain the upper bound of $\mathcal{J}_1(\widetilde{\mathbf{W}}) \equiv -\sum_{ij} \frac{X_{ij}^{\alpha}\widehat{X}_{ij}^{1-\alpha}}{\alpha(1-\alpha)}$ by G_1 (see Figure 1).

The gradient of G_1 with respect to W_{ik} using the chain rule is:

$$\frac{\partial G_1}{\partial \widetilde{W}_{ik}} = \sum_{al} \frac{\partial G_1}{\partial V_{al}} \frac{\partial V_{al}}{\partial \widetilde{W}_{ik}} = \frac{\partial G_1}{\partial V_{ik}} \frac{\partial V_{ik}}{\partial \widetilde{W}_{ik}} \tag{37}$$

$$= -\frac{1}{\alpha}\left(\frac{W_{ik}}{\widetilde{W}_{ik}}\right)^{2\alpha-1} \left(\widetilde{\mathbf{A}}\mathbf{W}\right)_{ik} \tag{38}$$

Recall $\mathbf{B} = \mathbf{1}_m\mathbf{1}_n^T\mathbf{X}^T + \mathbf{X}\mathbf{1}_n\mathbf{1}_m^T$. We have

$$\mathcal{J}_2(\widetilde{\mathbf{W}}) \equiv \sum_{ij} \frac{1}{\alpha}\left(\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T\mathbf{X}\right)_{ij} \tag{39}$$

$$= \frac{1}{2\alpha}\mathrm{Tr}\left[\widetilde{\mathbf{W}}^T\mathbf{B}\widetilde{\mathbf{W}}\right] \tag{40}$$

$$\leq \frac{1}{\alpha}\sum_{ik} \frac{\widetilde{W}_{ik}^2}{2W_{ik}} \left(\mathbf{B}\mathbf{W}\right)_{ik} \equiv G_2(\widetilde{\mathbf{W}}, \mathbf{W}) \tag{41}$$

Therefore,

$$\mathcal{J}(\widetilde{\mathbf{W}}) = \sum_{ij} \frac{X_{ij}}{1-\alpha} + \mathcal{J}_1(\widetilde{\mathbf{W}}) + \mathcal{J}_2(\widetilde{\mathbf{W}}) \tag{42}$$

$$\leq \sum_{ij} \frac{X_{ij}}{1-\alpha} + G_1(\widetilde{\mathbf{W}}, \mathbf{W}) + G_2(\widetilde{\mathbf{W}}, \mathbf{W}) \tag{43}$$

$$\equiv G(\widetilde{\mathbf{W}}, \mathbf{W}) \tag{44}$$

Minimization over $\widetilde{\mathbf{W}}$ is implemented by setting $\frac{\partial G}{\partial \widetilde{W}_{ik}} = 0$:

$$-\frac{1}{\alpha}\left(\frac{W_{ik}}{\widetilde{W}_{ik}}\right)^{2\alpha-1} \left(\widetilde{\mathbf{A}}\mathbf{W}\right)_{ik} + \frac{1}{\alpha}\frac{\widetilde{W}_{ik}}{W_{ik}} \left(\mathbf{B}\mathbf{W}\right)_{ik} = 0. \tag{45}$$

The factor $1/\alpha$ cancels when $\alpha \neq 0$, which leads to the update rule (14). When $\alpha \to 0$, we can apply L'Hôpital's rule to the both sides of (45) and obtain

$$\frac{\widetilde{W}_{ik}}{W_{ik}}\left[2\log\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)\left(\mathbf{B}\mathbf{W}\right)_{ik} - \left(\widetilde{\mathbf{A}}^{(0)}\mathbf{W}\right)_{ik}\right] = 0. \tag{46}$$

Notice that the sequence of W_{ik} remains positive given a positive initialization. Thus we can safely remove the factor $W_{ik}/\widetilde{W}_{ik}$, resulting the update rule (15) for $\alpha \to 0$ or the inverse I-divergence. In summary,

$$\mathcal{J}(\mathbf{W}^{\text{new}}) = G(\mathbf{W}^{\text{new}}, \mathbf{W}^{\text{new}}) \tag{47}$$

$$\leq G(\mathbf{W}^{\text{new}}, \mathbf{W}) \tag{48}$$

$$\leq G(\mathbf{W}, \mathbf{W}) = \mathcal{J}(\mathbf{W}), \tag{49}$$

where the first inequality comes from the upper bound and the second by the minimization. Iteratively applying (14) thus monotonically decreases $D_{\alpha}(\mathbf{X}||\mathbf{W}\mathbf{W}^T\mathbf{X})$. □

Remark 1:

Theoretically, convergent update rules for PNMF based on the non-normalized KL-divergence

$$\mathcal{J}_1(\widetilde{\mathbf{W}}) \equiv \sum_{ij} -\frac{X_{ij}^{\alpha}\widehat{X}_{ij}^{1-\alpha}}{\alpha(1-\alpha)} = \sum_{ij} g\left(X_{ij},\widehat{X}_{ij}\right) = \sum_{ij} f\left(\widehat{X}_{ij}\right) \tag{26}$$

$$= \sum_{ij} f\left(\sum_k \widetilde{W}_{ik}\left(\widetilde{\mathbf{W}}^T\mathbf{X}\right)_{kj}\right) = \sum_{ij} f\left(\sum_k \gamma_{ijk}\frac{\widetilde{W}_{ik}\left(\widetilde{\mathbf{W}}^T\mathbf{X}\right)_{kj}}{\gamma_{ijk}}\right) \tag{27}$$

$$\leq \sum_{ij}\sum_k \gamma_{ijk} f\left(\frac{\widetilde{W}_{ik}\left(\widetilde{\mathbf{W}}^T\mathbf{X}\right)_{kj}}{\gamma_{ijk}}\right) = \sum_{ij}\sum_k \gamma_{ijk} f\left(\frac{W_{ik}}{\gamma_{ijk}}\frac{\widetilde{W}_{ik}}{W_{ik}}\left(\widetilde{\mathbf{W}}^T\mathbf{X}\right)_{kj}\right) \tag{28}$$

$$= -\sum_{ij}\sum_k \gamma_{ijk}\frac{\alpha(1-\alpha)}{X_{ij}^{\alpha}}\left[\frac{W_{ik}}{\gamma_{ijk}}\right]^{1-\alpha} f\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right) f\left(\left(\widetilde{\mathbf{W}}^T\mathbf{X}\right)_{kj}\right) \tag{29}$$

$$= -\sum_{ij}\sum_k \gamma_{ijk}\frac{\alpha(1-\alpha)}{X_{ij}^{\alpha}}\left[\frac{W_{ik}}{\gamma_{ijk}}\right]^{1-\alpha} f\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right) f\left(\beta_{ajk}\frac{\sum_a \widetilde{W}_{ak}X_{aj}}{\beta_{ajk}}\right) \tag{30}$$

$$\leq -\sum_{ij}\sum_k \gamma_{ijk}\frac{\alpha(1-\alpha)}{X_{ij}^{\alpha}}\left[\frac{W_{ik}}{\gamma_{ijk}}\right]^{1-\alpha} f\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)\sum_a \beta_{ajk} f\left(\frac{\widetilde{W}_{ak}X_{aj}}{\beta_{ajk}}\right) \tag{31}$$

$$= -\sum_{ijak} \gamma_{ijk}\frac{\alpha(1-\alpha)}{X_{ij}^{\alpha}}\left[\frac{W_{ik}}{\gamma_{ijk}}\right]^{1-\alpha} f\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right)\beta_{ajk} f\left(\frac{W_{ak}}{\beta_{ajk}}\frac{\widetilde{W}_{ak}}{W_{ak}}X_{aj}\right) \tag{32}$$

$$= \sum_{ijak} \gamma_{ijk}\beta_{ajk}\left[\frac{\alpha(1-\alpha)}{X_{ij}^{\alpha}}\right]^2\left[\frac{W_{ik}W_{ak}}{\gamma_{ijk}\beta_{ajk}}\right]^{1-\alpha} f(X_{aj}) f\left(\frac{\widetilde{W}_{ik}}{W_{ik}}\right) f\left(\frac{\widetilde{W}_{ak}}{W_{ak}}\right) \tag{33}$$

$$= \sum_{aik}\left[\widetilde{W}_{ik}^{1-\alpha}W_{ik}^{\alpha}\right]\left[\widetilde{W}_{ak}^{1-\alpha}W_{ak}^{\alpha}\right]\left[-\frac{1}{\alpha(1-\alpha)}\sum_j Z_{ij}^{\alpha}X_{aj}\right] \tag{34}$$

$$= \sum_{aik} V_{ik}V_{ak}S_{ai} = \mathrm{Tr}\left(\widetilde{\mathbf{V}}^T\mathbf{S}\widetilde{\mathbf{V}}\right) = \frac{1}{2}\mathrm{Tr}\left[\widetilde{\mathbf{V}}^T\left(\mathbf{S}+\mathbf{S}^T\right)\widetilde{\mathbf{V}}\right] \tag{35}$$

$$\leq -\frac{1}{\alpha(1-\alpha)}\sum_{ik}\frac{\widetilde{V}_{ik}^2}{2V_{ik}}\left(\widetilde{\mathbf{A}}\mathbf{V}\right)_{ik} \equiv G_1(\widetilde{\mathbf{V}},\mathbf{V}). \tag{36}$$

Figure 1. Upper-bounding $\mathcal{J}_1(\widetilde{\mathbf{W}}) \equiv -\sum_{ij}\frac{X_{ij}^{\alpha}\widehat{X}_{ij}^{1-\alpha}}{\alpha(1-\alpha)}$.

are unresolved in the previous PNMF literature [19, 22, 23]. This is now given by our proof as a special case ($\alpha \to 1$):

$$W_{ik}^{\text{new}} = W_{ik} \sqrt{\frac{(\mathbf{Z}\mathbf{X}^T\mathbf{W} + \mathbf{X}\mathbf{Z}^T\mathbf{W})_{ik}}{\sum_j (\mathbf{W}^T\mathbf{X})_{kj} + (\sum_j X_{ij})(\sum_b W_{bk})}}. \tag{50}$$

Remark 2:

We have previouly proposed an algorithm that iterates the following two steps [18]:

$$W'_{ik} = W_{ik} \left[\frac{\left(\widetilde{\mathbf{A}}\mathbf{W}\right)_{ik}}{(\mathbf{B}\mathbf{W})_{ik}} \right]^{\frac{1}{\alpha}}, \tag{51}$$

$$W_{ik}^{\text{new}} = W'_{ik} \left(\frac{\sum_{ij} \widehat{X}_{ij} \widetilde{Z}_{ij}}{\sum_{ij} \widehat{X}_{ij}} \right)^{\frac{1}{2\alpha}} \tag{52}$$

The update rule (51) is obtained by turning α-PNMF into a constrained α-NMF with $\mathbf{H} = \mathbf{W}^T\mathbf{X}$. It guarantees the Lagrangian objective decreases in each iteration. However, the definition of such a function varies across different iterations and also across different starting values because the Lagrangian multipliers solved by the K.K.T. conditions are determined by the current $\mathbf{W}$. The resulting objectives are therefore not comparable, which hinders monitoring its convergence and prevents improvement by multiple runs using different initial guesses. By constrast, the update rule (14) assures the monotonic decrease of the original α-PNMF objective whose definition does not depend on the iterations and starting $\mathbf{W}$ values. Therefore, one may easily monitor the convergence, rerun the algorithm several times and select the solution with the best objective.

The new multiplicative algorithm also overcomes another shortcoming of the previous one. The update rule (51) is sensitive to the overall scaling of $\mathbf{W}$ and results in zigzag learning paths. Therefore it must be accompanied with a stabilization step (52) with re-calculated $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{Z}}$. However, the proof of the consistence of this additional update rule with the original objective $D_\alpha(\mathbf{W}\mathbf{W}^T\mathbf{X})$ is still lacking. In contrast, the new algorithm using (14) does not require any additional normalization or stabilization steps, which facilitates its theoretical analysis.

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. Firstly, 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 second approach approximates the transposed matrix $\mathbf{X}^T$ by $\mathbf{W}\mathbf{W}^T\mathbf{X}^T$ where $\mathbf{W} \in \mathbb{R}_+^{N\times r}$, where α-PNMF can be used for *clustering*, with the elements of $\mathbf{W}$ now indicating the membership of each sample in the r clusters. We conduct benchmark experiments on both cases.

4.1 Feature Extraction

We have used the FERET database of facial images [15] as the training data set. After the 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), $\alpha = 1$ (I-divergence) and $\alpha = 2$ (χ^2-divergence) are shown in Figure 2. 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 [20]:

$$\tau = 1 - \frac{\|\mathbf{R} - \mathbf{I}\|_F}{(r(r-1))}, \tag{53}$$

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

$$R_{st} = \frac{\mathbf{w}_s^T \mathbf{w}_t}{\|\mathbf{w}_s\|\|\mathbf{w}_t\|}. \tag{54}$$

Larger τ's indicate higher orthogonality and τ reaches 1 when the columns of **W** are completely orthogonal. The numerical values for the orthogonalities τ using the two compared methods are given under the respective basis image plots in Figure 2. All τ values in the right are considerably larger than their left counterparts, which confirms that α-PNMF is able to extract a sparser transformation matrix **W**.

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 [2]. 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 [16]. 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. [17]). We thus set r equal to five times 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:

$$\text{purity} = \frac{1}{N}\sum_{k=1}^{r} \max_{1\le l\le q} n_k^l, \tag{55}$$

$$\text{entropy} = -\frac{1}{n\log_2 q}\sum_{k=1}^{r}\sum_{l=1}^{q} n_k^l \log_2 \frac{n_k^l}{n_k}, \tag{56}$$

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 of the cost function. The resulting factor matrix is of high sparsity or orthogonality, which is desired for part-based feature extraction and data clustering. Experimental results with various datasets indicate that the proposed algorithm can be considered as a promising replacement for both α-NMF and PNMF.

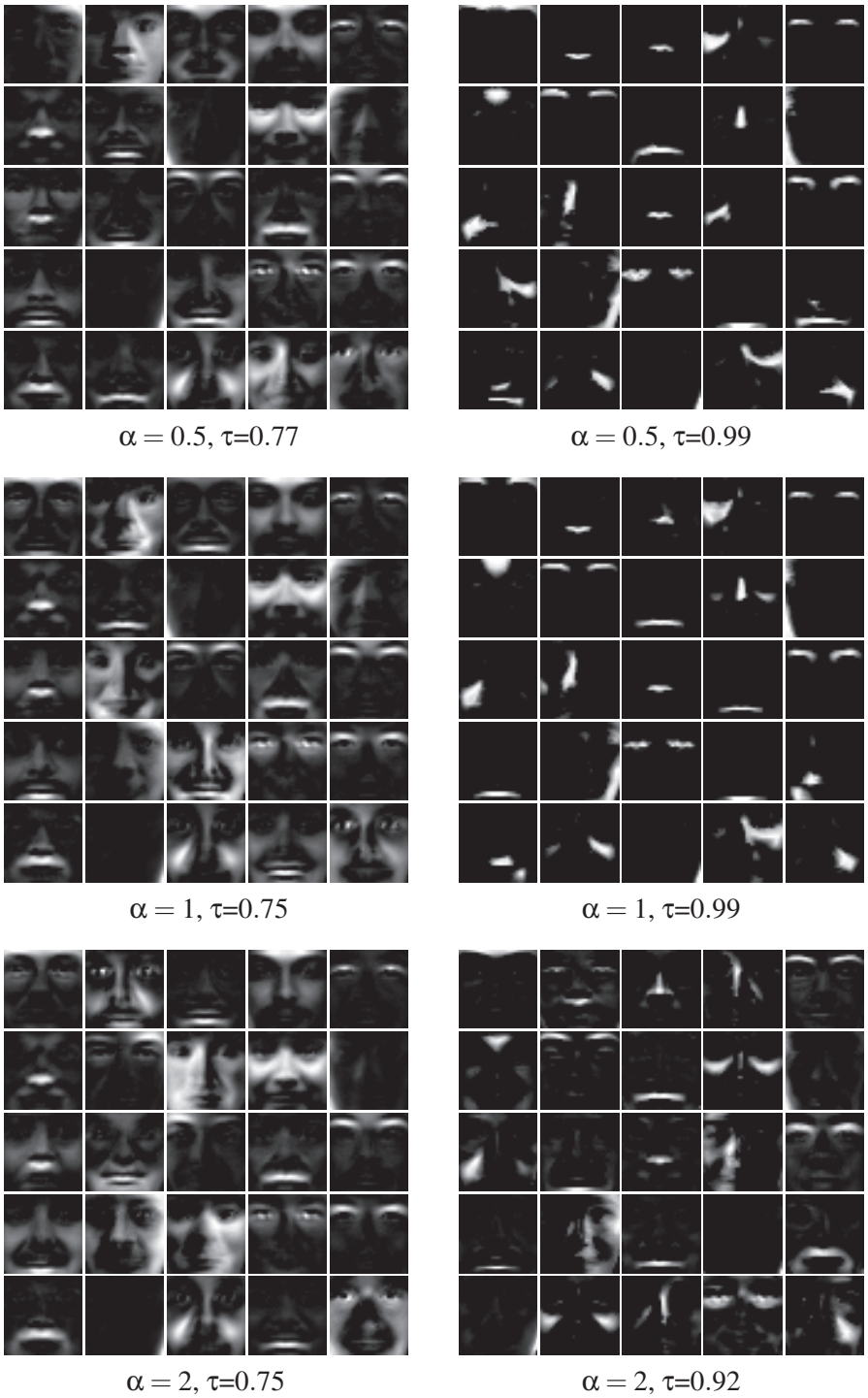

Figure 2. The basis images of (left) α-NMF and (right) α-PNMF.

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**

References

[1] S. Amari. *Differential-geometrical methods in statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1985.

[2] J.-Ph. Brunet, P. Tamayo, T.R. Golub, and J.P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[3] Seungjin Choi. Algorithms for orthogonal nonnegative matrix factorization. In *Proceedings of IEEE International Joint Conference on Neural Networks*, pages 1828–1832, 2008.

[4] A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi. Nonnegative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29:1433–1440, 2008.

[5] A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis*. John Wiley, 2009.

[6] I.S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with bregman divergences. In *Advances in Neural Information Processing Systems*, volume 18, pages 283–290, 2006.

[7] Chris Ding, Tao Li, and M.I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2010.

[8] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2006.

[9] K. Drakakis, S. Rickard, R. de Fréin, and A. Cichocki. Analysis of financial data using nonnegative matrix factorization. *International Mathematical Forum*, 3:1853–1870, 2008.

[10] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

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

[12] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.

[13] W. Liu, N. Zheng, and X. Lu. Non-negative matrix factorization for visual coding. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, volume 3, pages 293–296, 2003.

[14] A. Pascual-Montano, J.M. Carazo, Kieko Kochi, Dietrich Lehmann, and R. D. Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):403–415, 2006.

[15] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, October 2000.

[16] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, pages 138–142, Sarasota FL, December 1994.

[17] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

[18] Zhirong Yang and Erkki Oja. Projective nonnegative matrix factorization with α-divergence. In *Proceedings of 19th International Conference on Artificial Neural Networks (ICANN)*, pages 20–29, Limassol, Cyprus, 2009. Springer.

[19] Zhirong Yang and Erkki Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transaction on Neural Networks*, 2010. In press.

[20] Zhirong Yang, Zhijian Yuan, and Jorma Laaksonen. Projective non-negative matrix factorization with applications to facial image processing. *International Journal on Pattern Recognition and Artificial Intelligence*, 21(8):1353–1362, December 2007.

[21] S.S. Young, P. Fogel, and D. Hawkins. Clustering scotch whiskies using non-negative matrix factorization. *Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association*, 14(1):11–13, 2006.

[22] Zhijian Yuan and Erkki Oja. Projective nonnegative matrix factorization for image compression and feature extraction. In *Proc. of 14th Scandinavian Conference on Image Analysis (SCIA 2005)*, pages 333–342, Joensuu, Finland, June 2005.

[23] Zhijian Yuan and Erkki Oja. A family of modified projective nonnegative matrix factorization algorithms. In *Proceedings of the 9th International Symposium on Signal Processing and Its Applications (ISSPA)*, pages 1–4, 2007.