Face Recognition based on Linear Projective Non-negative Matrix Factorization with Kullback-Leibler Divergence

Lirui Hu2,3,4*, Siyuan Jia1 and Ning Wu2,4

1College of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi, 541000, China
2College of Electronics and Information Engineering, Beibu Gulf University, Qinzhou, Guangxi, 535000, China
3Key Laboratory of Big Data Resources Utilization, Qinzhou, Guangxi, 535000, China
4Key Laboratory of Advanced Technology for Internet of Things, Qinzhou, Guangxi, 535000, China
*Corresponding author’s e-mail: hulr163@163.com

Abstract. In order to solve the problem of feature extraction and classification of Non-negative Matrix Factorization (NMF) in face recognition, a Linear Projective Non-negative Matrix Factorization method based on Kullback-Leibler divergence (LP-NMF-DIV) is proposed. In LP-NMF-DIV, an objective function of Kullback-Leibler divergence is considered. The Taylor series expansion and the Newton iteration formula are used to find the root formula. The iterative algorithm of basis matrix and linear transformation matrix are derived, and the convergence of the algorithm is proved. The results of experimental show that the algorithm is convergent. In face recognition, compared with some typical NMF methods, the algorithm has high recognition accuracy; when the ranks of the basis matrices are set to different values, the algorithm is stable. The method for LP-NMF-DIV is effective.

1. Introduction
As we all know, the whole of things is composed of parts. Based on this point of view, Non-negative Matrix Factorization (NMF) [1] was constructed. Now NMF has been used in many fields, such as data dimension reduction, machine learning, image retrieval [1], text mining and more, NMF has become a research hotspot nowadays.

The Projective Non-negative Matrix Factorization (P-NMF) [2] was proposed on the basis of NMF. Because it was constructed from the projection angle, only the basis matrix \( W \) was calculated in the algorithm for P-NMF. The computational complexity was lower for one iteration step for P-NMF, as only one matrix had to be computed instead of two for NMF. However, since the algorithm for P-NMF did not converge, we had given two convergent algorithms which were respectively called Convergent Projective Non-negative Matrix Factorization (CP-NMF) [3] and Convergent Projective Non-negative Matrix Factorization with Kullback-Leibler Divergence (CP-NMF-DIV) [4].

From the projection and linear transformation angle, a new NMF was constructed, we called it the Non-negative Matrix Factorization of Linear Projection-Based (LPBNMF) \( \chi = W \Omega \chi \) [5]. In LPBNMF, they given the monotonic convergence algorithm, together with the computed orthogonality and sparsity.
of basis matrix. We also had proposed an easy and effective method called Linear Projective Non-negative Matrix Factorization (LP-NMF) [6], which makes LP-NMF work much easier.

According to the optimization rules of P-NMF, CP-NMF, CP-NMF-DIV, LPBNMF and LP-NMF, the basic matrices were all tend to orthogonal.

However, the Frobenius norm was used as the objective function in LP-NMF. Now, let’s use the Kullback-Leibler divergence as the objective function. Therefore, this article focuses on our method called Linear Projective Non-negative Matrix Factorization method based on Kullback-Leibler divergence (LP-NMF-DIV). The iterative formula of this method is too simple. Compared with NMF and some extended methods, it has higher recognition rate in face recognition.

2. Linear Projective Non-negative Matrix Factorization with Kullback-Leibler Divergence (LP-NMF-DIV)

We construct the objective function according to the definition of Kullback-Leibler divergence. [7]

$$D(X\|QX) = \sum_{ij} X_{ij} \log \frac{X_{ij}}{(QX)_{ij}} - X_{ij} + (QX)_{ij}$$

(1)

In the objective function Eq. (1), $X \in \mathbb{R}_{m \times n}$, $W \in \mathbb{R}^{r \times m}$ and $Q \in \mathbb{R}^{r \times n}$ ($\mathbb{R}^{r \times n}$ A is the set of all m-times-n matrices with nonnegative elements). Generally, the rank of $W$ and $Q$ is much lower than the rank of $X$ (i.e., $r << \min(m, n)$).

The mathematical model of NMF is based on nonlinear projection. However, the basic ideas are as follows: we turn the data $X$ into $QX$ by appropriate linear transformation $Q$, $QX$ is considered as the projection of the sample space $X$ onto a suitable subspace $W$, and the objective function $D(X\|QX)$ is minimized in Eq. (1) to obtain $W$ and $Q$. $W$ is called basis matrix and $Q$ is called linear transformation matrix here.

2.1. The update rule of basis matrix $W$

For any element $w_{ab}$ of $W$, set $D_{w_{ab}}$ as the part of $D(X\|QX)$, which is relevant to $w_{ab}$ in Eq. (1). By writing $w$ instead of $w_{ab}$ in the expression of $D_{w_{ab}}$, we may get a function $D_{w_{ab}}(w)$. Obviously, the first order derivative of $D_{w_{ab}}(w)$ at $w_{ab}$ is the first order partial derivative of $D(X\|QX)$ with respect to $w_{ab}$. That is

$$D'_{w_{ab}}(w_{ab}) = \frac{\partial D(X\|QX)}{\partial w_{ab}} = \sum_{ij} X_{ij} \log \frac{X_{ij}}{(QX)_{ij}} - X_{ij} + (QX)_{ij} \frac{\partial (QX)_{ij}}{\partial w_{ab}}$$

(2)

The second order derivative expression of $D_{w_{ab}}(w)$ at $w_{ab}$ is

$$D''_{w_{ab}}(w_{ab}) = \sum_{j} \frac{\partial (QX)_{ij}}{\partial w_{ab}} = \sum_{j} X_{ij} (QX)_{ij}$$

and the third order derivative expression of $D_{w_{ab}}(w)$ at $w_{ab}$ is

$$D'''_{w_{ab}}(w_{ab}) = \sum_{j} X_{ij} (QX)_{ij} \frac{\partial (QX)_{ij}}{\partial w_{ab}} = -2 \sum_{j} X_{ij} (QX)_{ij}.$$

Thus, the nth order derivative expression of $D_{w_{ab}}(w)$ at $w_{ab}$ is

$$D^{(n)}_{w_{ab}}(w_{ab}) = (-1)^n (n-1)! \sum_{j} X_{ij} (QX)_{ij}^{(n)}, \ \text{s.t.} \ n \geq 2.$$

The Taylor series expansion of $D_{w_{ab}}(w)$ at $w_{ab}$ is
\[ D_{w_{ab}} (w) = D_{w_{ab}} (w_{ab}) + \frac{1}{2} D'_{w_{ab}} (w_{ab}) (w-w_{ab})^2 + \sum_{n=1}^{\infty} \frac{1}{n!} D^{(n)}_{w_{ab}} (w_{ab}) (w-w_{ab})^n \]  

(3)

At this time, in order to express the calculation time of \( w_{ab} \) in numerical calculation conveniently, we can use \( w^{(i)}_{ab} \) to replace \( w_{ab} \) in \( D_{w_{ab}} (w) \). So, equation

\[ D_{w_{ab}} (w) = D_{w_{ab}} (w^{(i)}_{ab}) + \frac{1}{2} D'_{w_{ab}} (w^{(i)}_{ab}) (w-w^{(i)}_{ab})^2 + \sum_{n=1}^{\infty} \frac{1}{n!} D^{(n)}_{w_{ab}} (w^{(i)}_{ab}) (w-w^{(i)}_{ab})^n \]

(4)
is gotten from Eq. (3).

Now define a function

\[ P_{w_{ab}} (w, w^{(i)}_{ab}) = D_{w_{ab}} (w^{(i)}_{ab}) + \frac{1}{2} \left[ \sum_{j} \frac{X_{ij}}{(WQX)^j} \right] W_{ab} + \sum_{j} (QX)_{ij} \]

\[ \left( w-w^{(i)}_{ab} \right)^2 + \sum_{n=1}^{\infty} \frac{1}{n!} D^{(n)}_{w_{ab}} (w^{(i)}_{ab}) (w-w^{(i)}_{ab})^n \]

(5)

**Theorem 1.** \( P_{w_{ab}} (w, w^{(i)}_{ab}) \) is an auxiliary function of \( D_{w_{ab}} (w) \).

**Proof:** When \( w^{(i)}_{ab} = w \), \( P_{w_{ab}} (w, w^{(i)}_{ab}) = D_{w_{ab}} (w) \). Next we need show that \( P_{w_{ab}} (w, w^{(i)}_{ab}) \geq D_{w_{ab}} (w) \) when \( w^{(i)}_{ab} \neq w \).

Because \( W \geq 0, Q \geq 0, X \geq 0 \) and \( W_{ab} = w^{(i)}_{ab} \),

\[ \left[ \sum_{j} \frac{X_{ij}}{(WQX)^j} \right] W_{ab} + \sum_{j} (QX)_{ij} = D'_{w_{ab}} (w^{(i)}_{ab}) + \left[ \frac{1}{w_{ab}} \right] (w-w^{(i)}_{ab}) \geq D'_{w_{ab}} (w^{(i)}_{ab}) \cdot \]

Then

\[ P_{w_{ab}} (w, w^{(i)}_{ab}) \geq D_{w_{ab}} (w) \]

Therefore, it can be known from reference [7] that \( P_{w_{ab}} (w, w^{(i)}_{ab}) \) is an auxiliary function of \( D_{w_{ab}} (w) \).

**Theorem 2.** \( D_{w_{ab}} (w) \) is non-increasing under the update

\[ w^{(i+1)}_{ab} = \arg \min_{w} P_{w_{ab}} (w, w^{(i)}_{ab}) \]

**Proof:** Because

\[ D_{w_{ab}} (w^{(i+1)}_{ab}) \leq P_{w_{ab}} (w^{(i+1)}_{ab}, w^{(i)}_{ab}) \leq P_{w_{ab}} (w^{(i)}_{ab}, w^{(i)}_{ab}) = D_{w_{ab}} (w^{(i)}_{ab}) \]

\( D_{w_{ab}} (w) \) is non-increasing.

Using the definition of auxiliary function and Theorem 2, as long as get the local minimum of \( P_{w_{ab}} (w, w^{(i)}_{ab}) \), we can get the local minimum of \( D_{w_{ab}} (w) \). But how to get the local minimum of \( D_{w_{ab}} (w) \)? The method is to calculate the first order derivative of \( P_{w_{ab}} (w, w^{(i)}_{ab}) \) with respect to \( w \).

\[ P'_{w_{ab}} (w, w^{(i)}_{ab}) = D'_{w_{ab}} (w^{(i)}_{ab}) + \frac{1}{2} \left[ \sum_{j} \frac{X_{ij}}{(WQX)^j} \right] W_{ab} + \sum_{j} (QX)_{ij} \]

\[ (w-w^{(i)}_{ab})^2 + \sum_{n=1}^{\infty} \frac{1}{n!} D^{(n)}_{w_{ab}} (w^{(i)}_{ab}) (w-w^{(i)}_{ab})^n \]

(6)

In order to calculate the root of the equation

\[ P'_{w_{ab}} (w, w^{(i)}_{ab}) = 0 \]

(7)

Newton iterative formula is used to solve the root of Eq. (7) because the function \( P'_{w_{ab}} (w, w^{(i)}_{ab}) \) is a Taylor series expansion with the respect to \( w \) from Eq. (6), and the result as follows:

\[ w = w^{(i)}_{ab} - \frac{P'_{w_{ab}} (w^{(i)}_{ab}, w^{(i)}_{ab})}{P''_{w_{ab}} (w^{(i)}_{ab}, w^{(i)}_{ab})} \]

(8)

In Eq. (8),
\[
P_{ab}^w (w_{ab}^{(i)}, w_{ab}^{(i)}) = D_{ab}^w (w_{ab}^{(i)})
\]
and
\[
P_{ab}^w (w_{ab}^{(i)}, w_{ab}^{(i)}) = \left(\frac{\sum_j X_{abj}^w (QX_{abj}^e) W_{ab} + \sum_j (QX_{abj})_{ij}}{w_{ab}^{(i)}}\right).
\]

Eq. (8) can be simplified by using Eq. (2), and the result is
\[
w = w_{ab}^{(i)} + \left(\frac{\sum_j X_{abj}^w (QX_{abj}^e) W_{ab} + \sum_j (QX_{abj})_{ij}}{w_{ab}^{(i)}}\right).
\] (9)

So, the iterative rule of \( w_{ab} \) expressed in matrix form is
\[
w^{(i+1)} = w^{(i)} \left(\frac{1}{(W_{ab})^2}\right) W_{ab} + 1_{m \times 1} (1_{n \times n}(QX_{abj}))
\] (10)
where \( w^{(i)} = W \), \( 1_{m \times 1} \) and \( 1_{n \times n} \) respectively are \( m \)-by-1 and \( 1 \)-by-\( n \) matrices whose elements are all the number 1, and “.*”, “/” and “^2” respectively denote element-wise multiplication, division and square.

Because the Newton iterative formula for solving roots is convergent, we can use this iterative rule Eq. (10) and make the auxiliary function \( P_{ab}^w (w, w_{ab}) \) locally minimum, so the objective function \( D_{ab}^w (w) \) will be the local minimum. If we update all elements of \( W \) with formula (10), we can get the local minimum value of objective function \( D(X|WQX) \). Therefore, the algorithm converges.

The Eq. (10) is the iterative calculation formula of the basis matrix \( W \).

2.2. The update rule of linear transformation matrix Q

In the same way, the iterative rule for linear transformation matrix \( Q \) is
\[
Q^{(i+1)} = Q^{(i)} \left(\frac{1}{(W_{ab})^2}\right) W_{ab} + 1_{1 \times n} (1_{1 \times 1}(QX_{abj}))
\] (11)
where \( Q^{(i)} = Q \), \( 1_{1 \times n} \) and \( 1_{1 \times 1} \) respectively are 1-by-\( m \) and 1-by-\( n \) matrices whose elements are all the number 1, and “.*”, “/” and “^2” respectively denote element-wise multiplication, division and square too.

2.3. The steps of Algorithm

We can obtain an algorithm to calculate the basis matrix \( W \) and the linear transformation matrix \( Q \) by using Eq. (10) and Eq. (11). The specific steps of the algorithm are as follows:

**Step 1**: read the given data to matrix \( X \); initialize \( W \) and \( Q \) with non-negative data;

**Step 2**: update \( W \) with Eq. (10);

**Step 3**: update \( Q \) with Eq. (11);

**Step 4**: repeat step2 and Step3 until algorithm converges.

\( W^n \) and \( W^{n+1} \) are respectively used to denote the \( n \)-th and \( n+1 \)-th iterative result of the basis matrix \( W \).

\( Q^n \) and \( Q^{n+1} \) are respectively used to denote the \( n \)-th and \( n+1 \)-th iterative result of the linear transformation matrix \( Q \). In order to make the algorithm converge, it must satisfy
\[
D(W^n | W^n) < \epsilon 1, \text{ s.t. } \epsilon 1 > 0
\] (12)
and
\[
D(Q^n | Q^n) < \epsilon 2, \text{ s.t. } \epsilon 2 > 0
\] (13)
In inequality (12) and (13), $\varepsilon_1$ and $\varepsilon_2$ are two arbitrarily small positive numbers. $D$ stands for Kullback-Leibler divergence. Here, we call $\varepsilon_1$ and $\varepsilon_2$ convergent precision.

3. Experimental results and analysis

In the following experiments, $X$ consists of the first five images of each person in the Olivetti Research Laboratory (ORL) database, with a total of 200 data. In order to reduce the amount of computation and improve the speed of computation, each image is compressed to quarter.

3.1. Algorithm convergence and the basis matrix of LP-NMF-DIV

In this experiment, we set the rank of the basis matrix $W$ to 80 and the convergent precision 0.01 for $\varepsilon_1$ and $\varepsilon_2$, and $W$ and $Q$ are initialized randomly by non-negative data. After 2637 iterations, the convergence accuracy of the algorithm is obtained. In Figure 1, we can see that the algorithm is convergent. At the same time, Figure 2 is an image of the basis matrix. It is very sparse, which shows that by optimizing the objective function $D(X\|WQX)$, the basis matrix $W$ tends to be orthogonal.

![Figure 1. Curve of objective function value changing with iteration steps](image1)

![Figure 2. The image of Basis matrix](image2)

3.2. Face recognition

In the learning stage, $X$ consists of the first five images of each person in the ORL data set, with a total of 200 data. In order to reduce the amount of computation and improve the speed of computation, each image is compressed to quarter. We initialize randomly $W$ and $Q$ with non-negative data. After the algorithm converges, we get the basis matrix $W$, matrix $Q$, and $QX$.

In the test phase of pattern recognition, the remaining five images of each person in ORL face image database are taken as the test data, and each image is compressed to quarter of the original image.

In the following experiment, we use the $QX$ as the template library, use the $Q$ obtained in the learning phase to calculate the feature vectors $QX$ of the test images $X$, and follow the nearest neighbor rule for face recognition. Next, we compare this method with the methods of NMF [7], LNMF [8],
NMFOS [9], ONMF [10], DNMF [11] and LP-NMF [6]. When the ranks (i.e., the feature subspace dimensions) of the basis matrix are set 20 to 160, the results are shown in Figure 3.

From the Figure 3, the results show that the recognition accuracy of LP-NMF-DIV method is significantly higher than that of NMF and ONMF methods. The reason is that the basis matrix $W$ is forced to tend to be orthogonal by the objective function for LP-NMF-DIV in Eq. (1). Therefore, the basis matrix tends to be more orthogonal in LP-NMF-DIV than in NMF. The discriminative ability of feature vectors $QX$ of LP-NMF-DIV is better. When the rank of base matrix is greater than or equal to 60, the recognition accuracy of LP-NMF-DIV is slightly higher than that of LNMF or NMFOS. Because of the approximately orthogonal constraints for the basis matrices in the objective functions for LNMF and NMFOS, the feature vectors also works well in its discriminative power. However, feature vectors $QX$ have a better discriminative ability to LP-NMF-DIV. Compared with LP-NMF-DIV, the recognition accuracy of LP-NMF is closer.

In addition, since the orthogonality and the sparseness of the basis matrix for LP-NMF-DIV are always better, the recognition accuracy is stable in most cases where the ranks of the basis matrix are set with different values. Thus the recognition accuracy is less affected by the number of the rank of basis matrix.

4. Conclusion
In the development of NMF, we briefly introduce the research methods of NMF, LPBNMF, P-NMF and LP-NMF. Based on LP-NMF with Kullback-Leibler divergence, we propose a new method called LP-NMF-DIV.

In LP-NMF-DIV, the iterative formula and algorithm steps are given, and the convergence of the algorithm is proved. Some properties of LP-NMF-DIV are proved to be better in the convergence of algorithm, and orthogonality and sparsity of the basis matrix. Compared with the typical NMF method, LP-NMF-DIV method has higher recognition accuracy in face recognition.

Acknowledgment
This work was supported by Science and Technology Development Program of Chinese Guangxi Province (Grant No. AC17195057), Natural Science Foundation of Chinese Guangxi Province (Grant
No. 2016GXNSFAA380102), High-level Scientific Research and Cultivation Project of Qinzhou University, China (Grant No. 2016PY-GJ03), Key Laboratory of Advanced Technology for Internet of Things, Qinzhou, China (Grant No. IOT2017A02), and School Grade Research Project of Qinzhou University, China (Grant No. 2017KYQD118).

References
[1] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, vol. 401, no. 6755, pp. 788−791, 1999.
[2] Z. R. Yang, Z. J. Yuan, J. Laaksonen, Projective nonnegative matrix factorization with applications to facial image processing, International Journal of Pattern Recognition and Artificial Intelligence, vol. 21, no. 8, pp. 1353−1362, 2007.
[3] L. R. Hu, J. G. Wu, L. Wang, Method for Convergent Projective Non-negative Matrix Factorization, Journal of Information and Computational Science, vol. 10, no. 6, pp. 1847−1854, 2013.
[4] L. R. Hu, L. Dai, J. G. Wu, Convergent Projective Non-negative Matrix Factorization with Kullback–Leibler Divergence, Pattern Recognition Letters, vol. 36, pp. 15−21, 2014.
[5] L. Li, Y. J. Zhang, Linear projection-based non-negative matrix factorization, Acta Automatica Sinica, vol. 36, no.1, pp. 23−39, 2010.
[6] L. R. Hu, J. G. Wu, L. Wang, Linear Projective Non-negative Matrix Factorization, ICIC Express Letters, vol. 7, no. 8, pp. 2285−2291, 2013.
[7] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, Advances in Neural Information Processing Systems, MIT Press, pp. 556−562, 2001.
[8] S. Z. Li, X. W. Hou, H. J. Zhang, et al., Learning spatially localized, parts-based representation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1−6, 2001.
[9] Z. Li, X. Wu, H. Peng, Non-negative matrix factorization on orthogonal subspace, Pattern Recognition Letters, vol. 31, no. 9, pp. 905−911, 2010.
[10] J. Yoo, S. Choi, Orthogonal non-negative matrix factorization for co-clustering: multiplicative updates on stiefel manifolds, Information Processing and Management, vol. 46, no. 5, pp. 559-570, 2010.
[11] S. Nikitidis, A. Tefas, I. Pitas, Recent advances in discriminant non-negative matrix factorization, in: Proceedings of the 1st Asian Conference on Pattern Recognition, pp. 1-6, 2011.