Multi-Component Nonnegative Matrix Factorization

Jing Wang[1], **Feng Tian**[1], **Xiao Wang**[2,*], **Hongchuan Yu**[3], **Chang Hong Liu**[4], **Liang Yang**[5]
[1]Faculty of Science and Technology, Bournemouth University, UK
[2]Department of Computer Science and Technology, Tsinghua University, China
[3]National Centre for Computer Animation, Bournemouth University, UK
[4]Department of Psychology, Bournemouth University, UK
[5]School of Information Engineering, Tianjin University of Commerce, China
{jwang, ftian, hyu, liuc}@bournemouth.ac.uk, wangxiao_cv@tju.edu.cn, yangliang@iie.ac.cn

Abstract

Real data are usually complex and contain various components. For example, face images have expressions and genders. Each component mainly reflects one aspect of data and provides information others do not have. Therefore, exploring the semantic information of multiple components as well as the diversity among them is of great benefit to understand data comprehensively and in-depth. However, this cannot be achieved by current nonnegative matrix factorization (NMF)-based methods, despite that NMF has shown remarkable competitiveness in learning parts-based representation of data. To overcome this limitation, we propose a novel multi-component nonnegative matrix factorization (MCNMF). Instead of seeking for only one representation of data, MCNMF learns multiple representations simultaneously, with the help of the Hilbert Schmidt Independence Criterion (HSIC) as a diversity term. HSIC explores the diverse information among the representations, where each representation corresponds to a component. By integrating the multiple representations, a more comprehensive representation is then established. Extensive experimental results on real-world datasets have shown that MCNMF not only achieves more accurate performance over the state-of-the-arts using the aggregated representation, but also interprets data from different aspects with the multiple representations, which is beyond what current NMFs can offer.

1 Introduction

Finding an optimal data representation is a fundamental problem in many data analysis tasks [Tao *et al.*, 2009; Liu *et al.*, 2012]. A good data representation can typically reveal the latent structure of data and facilitate further processes such as clustering, classification and recognition. Nonnegative matrix factorization (NMF) as a fundamental approach for data representation has attracted great attentions because it possesses parts-of-whole interpretations and produces superior practical performance [Lee and Seung, 1999].

Several variants of NMF have been proposed to seek for more effective data representation in the recent years. [Kong *et al.*, 2011] proposed a robust formulation of NMF (RNMF) to deal with large noises by $L_{2,1}$-norm. [Guan *et al.*, 2012] then presented Manhattan NMF (MahNMF) to alleviate heavy tailed Laplacian noise. [Liu *et al.*, 2012] developed a semi-supervised approach called constrained NMF (CNMF) that takes the label information as hard constraints to enforce data with the same label to have the same representations. Under the assumption that the data points nearby have more similar data representations than those far away, [Cai *et al.*, 2011] proposed a graph regularized NMF (GNMF) to model the local manifold structure. Subsequently, [Wang *et al.*, 2016b] proposed a correntropy induced metric based graph regularized NMF (CGNMF) to deal with noises and preserving the intrinsic geometric structure of data simultaneously. [Qian *et al.*, 2016] adopted an approximation of Earth Movers Distance to utilize information of feature correlation. Additionally, LANMF [Liu *et al.*, 2016] presented a large-cone penalty framework to obtain attractive local solutions for NMF. Given a dataset with multiple types of features, [Liu *et al.*, 2013] proposed a multi-view NMF (MultiNMF) which learns a consensus representation shared by multiple features. AMVNMF [Wang *et al.*, 2016a] then extended MultiNMF to semi-supervised setting by enforcing data of same label have the same representations regardless of features.

These NMF-based approaches, which either incorporate regularization terms or prior information for more accurate learning, all tend to treat the features of data as a whole and obtain a single feature representation. However, it is well recognized that real data are complex and consist of components [Changpinyo *et al.*, 2013; Ou *et al.*, 2015]. Taken the Yale dataset[1] as an example, Figure 1 illustrates the face images consisting of multiple components including gender, facial expressions, ethnicity, and lighting direction (under which the images were taken), etc. Since each component mainly represents one subset of features and contains the specific information of the data, current NMFs are unable to distinguish these embedded components thus cannot adequately exploit diverse information among them, which may not lead to satisfactory representations. Hence, it becomes crucial to explore diverse

*Corresponding author

[1]http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

Figure 1: Sample images of the Yale dataset. Each column shows one subject's faces. Images in the same rows contain same components, such as faces with glasses and a neutral expression in row 1; faces without glasses and a happy expression in row 2; faces lit from left and with a neutral expression in row 3.

information from multiple components in order to represent data more comprehensively and accurately.

In fact, the representations learnt through multiple components will also enable us to understand data at a semantic level. For example, when clustering the Yale dataset, current NMFs get only one clustering solution (i.e., all face images of a subject being grouped into one cluster) based on global features of a single representation matrix. With representations of multiple components, multiple clustering solutions can be achieved. For example, one cluster of images can be faces with glasses and another can be faces with a happy expression.

To achieve this, we propose a novel multi-component nonnegative matrix factorization (MCNMF). It captures more comprehensive information and interprets data from different perspectives, by leveraging the multiple components. Instead of factorizing the data matrix into a single basis and representation matrix, MCNMF learns multiple representations based on different basis matrices. The Hilbert-Schmidt Independence Criterion (HSIC) [Gretton *et al.*, 2005] which measures dependence in terms of a kernel dependence measure is applied as a diversity term. With this term, we explicitly coregularize different components to enforce the diversity of the jointly learned representations. An aggregated representation is then established by combining these multiple representations. To solve the objective function of MCNMF, we derive a new iterative updating optimization scheme, with its correctness and convergence being proven as well. Experiments on clustering have demonstrated that MCNMF not only improves the accuracy by the aggregated representation, but also captures different semantic properties of data.

2 Brief Intro to NMF

Given a nonnegative data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$, each data $\mathbf{x}_f$ $(1 \leq f \leq n)$ has m-dimensional features. NMF [Lee and Seung, 1999] aims to find a basis matrix $\mathbf{W} \in \mathbb{R}^{m\times k}$ and a representation matrix $\mathbf{H} \in \mathbb{R}^{k\times n}$, where the product of the two matrices can well approximate the original matrix, represented as $\mathbf{X} \approx \mathbf{WH}$, and k (usually $k \ll m$) denotes the reduced dimension. Formally, NMF solves the following optimization problem to compute the optimal representation matrix $\mathbf{H}^*$:

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} \|\mathbf{X} - \mathbf{WH}\|_F^2. \tag{1}$$

Then the multiplicative algorithm is derived to infer $\mathbf{W}$ and $\mathbf{H}$ [Lee and Seung, 2001]. Obviously, this standard NMF regards the features of data as a whole, so can be considered as a single component approach. Arguably, the semantic aspects of data is much richer than what a single component can capture. To understand data thoroughly and in-depth, we propose our multi-component NMF (MCNMF) in the following.

3 MCNMF

3.1 Objective Function

Assuming $\mathbf{X}$ comes with V components, we use $\mathbf{H}^{(i)} \in \mathbb{R}^{k^{(i)}\times n}$ to denote the representation with $k^{(i)}$-dimensional features that corresponds to the i-th ($i \in \{1, 2, \ldots, V\}$) components, and $\mathbf{W}^{(i)}$ be the corresponding representation matrix of $\mathbf{H}^{(i)}$. Then the product of each $\mathbf{W}^{(i)}\mathbf{H}^{(i)}$ should well approximate $\mathbf{X}$, i.e., $\mathbf{X} \approx \mathbf{W}^{(i)}\mathbf{H}^{(i)}$, from each perspective. To seek for multiple optimal representations $\{\mathbf{H}^{(i)^*}\}_{i=1}^V$, we have the following function:

$$\min_{\mathbf{W}^{(i)}\geq 0, \mathbf{H}^{(i)}\geq 0} \sum_{i=1}^{V} \|\mathbf{X} - \mathbf{W}^{(i)}\mathbf{H}^{(i)}\|_F^2. \tag{2}$$

This will allow us to factorize $\mathbf{X}$ straightforwardly. However, it may fail to explore the diverse information of multiple components effectively as each $\mathbf{H}^{(i)}$ could be very close to or even same as each other.

For any data, $\mathbf{x}_f$, it comes with a pair of components, i and j. $\mathbf{x}_f$'s latent distinct information of each component cannot be fully explored unless its representations of two components, i.e., $\mathbf{h}_f^{(i)}$ and $\mathbf{h}_f^{(j)}$, are enforced to be independent to each other. Given n data vectors, we assume that each ith component is drawn from $\mathcal{X}$ space and the jth component from $\mathcal{Y}$ space. Then, in essence, we aim to learn a mapping function G of their representations from $S := \{(\mathbf{h}_1^{(i)}, \mathbf{h}_1^{(j)}), (\mathbf{h}_2^{(i)}, \mathbf{h}_2^{(j)}), \ldots, (\mathbf{h}_n^{(i)}, \mathbf{h}_n^{(j)})\} \subseteq \mathcal{X} \times \mathcal{Y}$, i.e., G: $\mathcal{X} \to \mathcal{Y}$, to minimize the dependence between data representations in the $\mathcal{X}$ and $\mathcal{Y}$.

To do so, we employ the Hilbert-Schmidt Independence Criterion (HSIC) due to its simplicity and neat theoretical properties such as exponential convergence. HSIC computes the square of the norm of the cross-covariance operator over the domain $\mathcal{X} \times \mathcal{Y}$ in Hilbert Space. As an effective measure of dependence, the HSIC has been applied to several machine learning tasks recently [Song *et al.*, 2007; Zhang and Zhou, 2010; Niu *et al.*, 2010]. Mathmatically, an empirical estimate of the HSIC [Gretton *et al.*, 2005] is defined as

$$\mathrm{HSIC}(\mathbf{H}^{(i)}, \mathbf{H}^{(j)}) = (n-1)^{-2} tr(\mathbf{R}\mathbf{K}^{(i)}\mathbf{R}\mathbf{K}^{(j)}), \tag{3}$$

where $\mathbf{K}^{(i)}$ and $\mathbf{K}^{(j)}$ are the centered Gram matrices of kernel functions defined over $\mathbf{H}^{(i)}$ and $\mathbf{H}^{(j)}$. $\mathbf{R} = \mathbf{I} - \frac{1}{n}\mathbf{e}\mathbf{e}^T$, where $\mathbf{I}$ is an identity matrix and $\mathbf{e}$ is an all-one column vector.

Thus, to explore the diverse information from more components, we extend (3) and combine it with (2) to produce the following function:

$$\min_{\mathbf{W}^{(i)}\geq 0,\mathbf{H}^{(i)}\geq 0}\sum_{i=1}^{V}\|\mathbf{X}-\mathbf{W}^{(i)}\mathbf{H}^{(i)}\|_F^2+\alpha\sum_{j\neq i}\mathrm{HSIC}(\mathbf{H}^{(i)},\mathbf{H}^{(j)}), \tag{4}$$

where α is the parameter of the diversity regularization term. The first term represents the error between $\mathbf{X}$ and the product of the basis and representation matrices in different components. The second term ensures that any two of V representations be diverse to each other.

Here, we use the inner product kernel for HSIC, i.e., $\mathbf{K}^{(i)} = {\mathbf{H}^{(i)}}^T\mathbf{H}^{(i)}$. For notational convenience, we ignore the scaling factor $(n-1)^{-2}$ of HSIC, and rewrite (4) to form the final objective function as

$$\min_{\mathbf{W}^{(i)}\geq 0,\mathbf{H}^{(i)}\geq 0}\sum_{i=1}^{V}\|\mathbf{X}-\mathbf{W}^{(i)}\mathbf{H}^{(i)}\|_F^2+\alpha\sum_{j\neq i}tr(\mathbf{R}\mathbf{K}^{(i)}\mathbf{R}\mathbf{K}^{(j)}). \tag{5}$$

After obtaining the optimal representation ${\mathbf{H}^{(i)}}^*$ of each component, the final aggregated representation $\mathbf{H}^*$ can be obtained by combining all ${\mathbf{H}^{(i)}}^*$, i.e., $\mathbf{H}^* = [{\mathbf{H}^{(1)}}^*, {\mathbf{H}^{(2)}}^*, \ldots, {\mathbf{H}^{(V)}}^*] \in \mathbb{R}^{\sum_{i=1}^{V}k^{(i)}\times n}$.

3.2 Optimization

The optimization problem in (5) is not convex in both variables $\mathbf{W}^{(i)}$ and $\mathbf{H}^{(i)}$, so it is infeasible to find the global minimum. In addition, as the matrix $\mathbf{R}$ contains negative values, it is technically challenging to solve (5) directly. Here we propose an algorithm that separates the optimization of (5) to two subproblems and optimizes them iteratively, which guarantees each subproblem converges to the local minima.

$\mathbf{W}^{(i)}$-subproblem: Updating $\mathbf{W}^{(i)}$ with $\mathbf{H}^{(i)}$ fixed in (5) leads to a standard NMF formulation [Lee and Seung, 2001], so the updating rule for $\mathbf{W}^{(i)}$ is

$$\mathbf{W}^{(i)} \leftarrow \mathbf{W}^{(i)} \odot \frac{(\mathbf{X}{\mathbf{H}^{(i)}}^T)}{(\mathbf{W}^{(i)}\mathbf{H}^{(i)}{\mathbf{H}^{(i)}}^T)}. \tag{6}$$

$\mathbf{H}^{(i)}$-subproblem: When updating $\mathbf{H}^{(i)}$ with $\mathbf{W}^{(i)}$ in (5) fixed, we need to solve the following function:

$$\min_{\mathbf{H}^{(i)}\geq 0} J(\mathbf{H}^{(i)}) = \|\mathbf{X}-\mathbf{W}^{(i)}\mathbf{H}^{(i)}\|_F^2+\alpha\sum_{j=1,j\neq i}^{V}tr(\mathbf{R}\mathbf{K}^{(i)}\mathbf{R}\mathbf{K}^{(j)}) \tag{7}$$

We then introduce a Lagrange multiplier matrix $\boldsymbol{\eta} = [\eta_{pq}] \in \mathbb{R}^{k\times n}$ for the nonnegative constraint on $\mathbf{H}^{(i)}$. Utilizing $\|\mathbf{A}\|_F^2 = tr(\mathbf{A}^T\mathbf{A})$, we obtain the following function:

$$\begin{aligned}\min_{\mathbf{H}^{(i)}\geq 0} J'(\mathbf{H}^{(i)}) &= tr(\mathbf{X}\mathbf{X}^T) - 2tr(\mathbf{X}{\mathbf{H}^{(i)}}^T{\mathbf{W}^{(i)}}^T)\\ &+ tr(\mathbf{W}^{(i)}\mathbf{H}^{(i)}{\mathbf{H}^{(i)}}^T{\mathbf{W}^{(i)}}^T)\\ &+ \alpha\sum_{j=1,j\neq i}^{V}tr(\mathbf{R}{\mathbf{H}^{(i)}}^T\mathbf{H}^{(i)}\mathbf{R}\mathbf{K}^{(j)}) + tr(\boldsymbol{\eta}\mathbf{H}^{(i)}).\end{aligned} \tag{8}$$

Setting the derivative of $J'(\mathbf{H}^{(i)})$ to be 0 with respect to $\mathbf{H}^{(i)}$, we have

$$\boldsymbol{\eta} = {\mathbf{W}^{(i)}}^T\mathbf{X} - {\mathbf{W}^{(i)}}^T\mathbf{W}^{(i)}\mathbf{H}^{(i)} - \alpha\mathbf{H}^{(i)}\mathbf{R}\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R}. \tag{9}$$

Following the Karush-Kuhn-Tucker (KKT) condition for the nonnegativity of $\mathbf{H}^{(i)}$, we have the following equation:

$$\begin{aligned}&({\mathbf{W}^{(i)}}^T\mathbf{X} - {\mathbf{W}^{(i)}}^T\mathbf{W}^{(i)}\mathbf{H}^{(i)}\\ &- \alpha\mathbf{H}^{(i)}\mathbf{R}\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R})_{pq}H_{pq}^{(i)} = 0.\end{aligned} \tag{10}$$

Because $\mathbf{R}$ contains negative values, we decompose $\mathbf{R}$ into two nonnegative parts for ensuring $\mathbf{H}^{(i)} \geq 0$ in each iteration:

$$\mathbf{R} = \mathbf{R}^+ - \mathbf{R}^-, \tag{11}$$

where $\mathbf{R}_{pq}^+ = (|\mathbf{R}_{pq}|+\mathbf{R}_{pq})/2$ and $\mathbf{R}_{pq}^- = (|\mathbf{R}_{pq}|-\mathbf{R}_{pq})/2$. Substituting (11) into (10), we obtain

$$\begin{aligned}&({\mathbf{W}^{(i)}}^T\mathbf{X} - {\mathbf{W}^{(i)}}^T\mathbf{W}^{(i)}\mathbf{H}^{(i)}\\ &+ \alpha\mathbf{H}^{(i)}(\mathbf{R}^+\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R}^- + \mathbf{R}^-\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R}^+)\\ &- \alpha\mathbf{H}^{(i)}(\mathbf{R}^-\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R}^- + \mathbf{R}^+\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R}^+))_{pq}H_{pq}^{(i)} = 0.\end{aligned} \tag{12}$$

This is the fixed point equation whose solution must satisfy at convergence. Denote $\mathbf{R}_a = \mathbf{R}^+\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R}^-$, $\mathbf{R}_b = \mathbf{R}^-\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R}^+$, $\mathbf{R}_c = \mathbf{R}^-\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R}^-$, $\mathbf{R}_d = \mathbf{R}^+\sum_{j=1,j\neq i}^{V}\mathbf{K}^{(j)}\mathbf{R}^+$, then given an initial value of $\mathbf{H}^{(i)}$, the successive update of $\mathbf{H}^{(i)}$ is:

$$\mathbf{H}^{(i)} \leftarrow \mathbf{H}^{(i)} \odot \sqrt{\frac{{\mathbf{W}^{(i)}}^T\mathbf{X} + \alpha\mathbf{H}^{(i)}(\mathbf{R}_a+\mathbf{R}_b)}{{\mathbf{W}^{(i)}}^T\mathbf{W}^{(i)}\mathbf{H} + \alpha\mathbf{H}^{(i)}(\mathbf{R}_c+\mathbf{R}_d)}}. \tag{13}$$

The correctness of the updating rule (13) can be guaranteed by the following theorem.

Theorem 1. If the updating rule of $\mathbf{H}^{(i)}$ converges, then the final solution satisfies the KKT optimality condition.

Proof of Theorem 1. At convergence, $\mathbf{H}^{\infty} = \mathbf{H}^{t+1} = \mathbf{H}^t = \mathbf{H}$, where t denotes the t-th iteration, i.e.,

$$\mathbf{H}^{(i)} = \mathbf{H}^{(i)} \odot \sqrt{\frac{{\mathbf{W}^{(i)}}^T\mathbf{X} + \alpha\mathbf{H}^{(i)}(\mathbf{R}_a+\mathbf{R}_b)}{{\mathbf{W}^{(i)}}^T\mathbf{W}^{(i)}\mathbf{H} + \alpha\mathbf{H}^{(i)}(\mathbf{R}_c+\mathbf{R}_d)}} \tag{14}$$

Then for each $H_{pq}^{(i)}$, we have

$$\begin{aligned}&({\mathbf{W}^{(i)}}^T\mathbf{X} - {\mathbf{W}^{(i)}}^T\mathbf{W}^{(i)}\mathbf{H}^{(i)} + \alpha\mathbf{H}^{(i)}(\mathbf{R}_a+\mathbf{R}_b)\\ &- \alpha\mathbf{H}^{(i)}(\mathbf{R}_c+\mathbf{R}_d))_{pq}(\mathbf{H}^{(i)})_{pq}^2 = 0.\end{aligned} \tag{15}$$

which is equivalent to (12). □

We can now prove the convergence of the updating rule, by making use of an auxiliary function as in [Lee and Seung, 2001]. The definition of the auxiliary function is as follows:

Table 1: Clustering results ((mean ± standard deviation)%) on the four datasets (bold numbers represent the best results)

	Metric	NMF	RNMF	GNMF	Cauchy NMF	LANMF	MCNMF
Yale	AC	40.48 ±3.25	38.55 ±2.76	41.58±2.54	41.45 ±4.26	39.76 ±2.70	**46.42±1.95**
	NMI	46.35 ±2.15	43.98 ±2.46	46.30 ±1.66	49.57 ±2.88	45.52 ±**1.25**	**49.65**±1.66
	purity	42.91 ±2.16	41.33 ±3.36	42.67±2.73	43.39 ±3.02	42.18 ±1.64	**47.15±1.45**
ORL	AC	54.90±3.44	54.20 ±2.11	59.60±2.50	56.45±2.86	52.40 ±2.31	**62.95±1.20**
	NMI	76.22±1.34	75.33±1.04	77.80±1.12	74.80±1.23	73.11±1.79	**79.39±1.10**
	purity	62.20±2.08	59.75±1.51	64.55±1.59	60.45±1.85	57.80±1.55	**66.20±1.47**
COIL20	AC	62.49±1.56	59.57±4.83	68.39±2.62	63.49±4.34	63.17 ±3.98	**69.61±1.30**
	NMI	74.35±1.49	73.24±2.21	77.30±1.54	76.34±2.08	76.63±2.51	**78.84±0.81**
	purity	66.40±1.37	64.29±3.75	69.83±2.47	67.28±2.06	67.68±3.51	**70.06±1.25**
Notting-Hill	AC	68.04±3.68	74.08 ±2.90	75.88 ±3.40	64.65±4.93	72.64 ±4.17	**77.54±1.69**
	NMI	60.27±3.50	64.74±**2.55**	62.97±2.93	56.29±2.32	64.94±3.56	**66.63** ±3.33
	purity	72.93±5.38	78.39±**2.25**	77.19±2.85	70.25±3.93	78.67±3.67	**79.49**±2.79

Definition 1. A function $G(\mathbf{Q}, \mathbf{Q}')$ is an auxiliary function of the function $J(\mathbf{Q})$ if $G(\mathbf{Q}, \mathbf{Q}') \geq J(\mathbf{Q})$ and $G(\mathbf{Q}, \mathbf{Q}) = J(\mathbf{Q})$ for any $\mathbf{Q}, \mathbf{Q}'$.

The auxiliary function gives rise to the following lemma [Lee and Seung, 2001]:

Lemma 1. If G is an auxiliary function of J, then J is non-increasing under the update rule $\mathbf{Q}^{t+1} = \arg\min_{\mathbf{Q}} G(\mathbf{Q}, \mathbf{Q}^t)$.

Under the constraint in (11), we now have the specific form of the auxiliary function $G(\mathbf{H}^{(i)}, \mathbf{H}^{(i)'})$ for the objective function $J(\mathbf{H}^{(i)})$ in (7) based on Lemma 2.

Lemma 2. The function

$$
\begin{aligned}
G(\mathbf{H}^{(i)}, \mathbf{H}^{(i)'}) = & -2\sum_{pq}(\mathbf{W}^{(i)^T}\mathbf{X})_{pq}\mathbf{H}^{(i)'}_{pq}(1+\log\frac{\mathbf{H}^{(i)'}_{pq}}{\mathbf{H}^{(i)}_{pq}}) \\
& + \sum_{pq}\frac{(\mathbf{W}^{(i)^T}\mathbf{W}^{(i)}\mathbf{H}^{(i)'})_{pq}\mathbf{H}^{(i)^2}_{pq}}{\mathbf{H}^{(i)'}_{pq}} \\
& - \sum_{pqk}(\mathbf{R}_a+\mathbf{R}_b)_{jk}\mathbf{H}^{(i)'}_{pq}\mathbf{H}^{(i)'}_{pk}(1+\log\frac{\mathbf{H}^{(i)}_{pq}\mathbf{H}^{(i)}_{pk}}{\mathbf{H}^{(i)'}_{pq}\mathbf{H}^{(i)'}_{pk}}) \\
& + \sum_{pq}\frac{(\mathbf{H}^{(i)'}(\mathbf{R}_c+\mathbf{R}_d))_{pq}\mathbf{H}^{(i)^2}_{pq}}{\mathbf{H}^{(i)'}_{pq}}
\end{aligned} \tag{16}
$$

is an auxiliary function for $J(\mathbf{H}^{(i)})$ in (7).

Proof of Lemma 2. We find upper bounds for each of the two positive terms by the following lemma [Ding *et al.*, 2010],

Lemma 3. For any nonnegative matrices $\mathbf{S} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{g\times g}$, $\mathbf{F} \in \mathbb{R}^{n\times g}$ and $\mathbf{F}' \in \mathbb{R}^{n\times g}$, with $\mathbf{S}$ and $\mathbf{B}$ being symmetric, then the following inequality holds

$$
tr(\mathbf{F}^T\mathbf{S}\mathbf{F}\mathbf{B}) \leq \sum_{i=1}^{n}\sum_{p=1}^{g}(\mathbf{S}\mathbf{F}'\mathbf{B})\frac{\mathbf{F}^2{}_{ip}}{\mathbf{F}'{}_{ip}}. \tag{17}
$$

Then, we have following inequations:

$$
tr(\mathbf{W}^{(i)^T}\mathbf{W}^{(i)}\mathbf{H}^{(i)}\mathbf{H}^{(i)^T}) \leq \sum_{pq}\frac{(\mathbf{W}^{(i)^T}\mathbf{W}^{(i)}\mathbf{H}'_i)_{pq}(\mathbf{H}^{(i)})^2_{pq}}{(\mathbf{H}'_i)_{pq}}, \tag{18}
$$

$$
tr(\mathbf{H}(\mathbf{R}_c+\mathbf{R}_d)\mathbf{H}^T) \leq \sum_{pq}\frac{(\mathbf{H}'(\mathbf{R}_c+\mathbf{R}_d))_{pq}\mathbf{H}^2_{pq}}{\mathbf{H}'_{pq}}. \tag{19}
$$

To obtain lower bounds for the remaining terms, we use the inequality $z > 1 + \log z, \forall z > 0$ [Ding *et al.*, 2010] and have

$$
\begin{aligned}
& tr(\mathbf{W}^T\mathbf{X}\mathbf{H}^T) \\
& \geq \sum_{pq}(\mathbf{W}^T\mathbf{X})_{pq}\mathbf{H}'_{pq}(1+\log\frac{\mathbf{H}_{pq}}{\mathbf{H}'_{pq}}),
\end{aligned} \tag{20}
$$

$$
\begin{aligned}
& tr(\mathbf{H}(\mathbf{R}_a+\mathbf{R}_b)\mathbf{H}^T) \\
& \geq \sum_{pqk}(\mathbf{R}_a+\mathbf{R}_b)_{jk}\mathbf{H}'_{pq}\mathbf{H}'_{pk}(1+\log\frac{\mathbf{H}_{pq}\mathbf{H}_{pk}}{\mathbf{H}'_{pq}\mathbf{H}'_{pk}}).
\end{aligned} \tag{21}
$$

Collecting all bounds, we have the final auxiliary function in Lemma 2. □

Based on the lemmas 1 and 2, we can prove the convergence of the updating rule (13).

Theorem 2. The optimization problem (7) is non-increasing under the iterative updating rule (13).

Proof of Theorem 2. Lemma 2 provides a specific form $G(\mathbf{H}, \mathbf{H}')$ of the auxiliary function for $J(\mathbf{H})$ in the problem (7). We can have the solution for $\min_{\mathbf{H}} G(\mathbf{H}, \mathbf{H}')$ by the following KKT condition

$$
\begin{aligned}
\frac{\partial G(\mathbf{H}, \mathbf{H}')}{\partial \mathbf{H}_{pq}} = & -2(\mathbf{W}^T\mathbf{X})_{pq}\frac{\mathbf{H}'_{pq}}{\mathbf{H}_{pq}} + 2\frac{(\mathbf{W}^T\mathbf{W}\mathbf{H}')_{pq}\mathbf{H}_{pq}}{\mathbf{H}'_{pq}} \\
& - 2\frac{(\mathbf{H}'(\mathbf{R}_a+\mathbf{R}_b))_{pq}\mathbf{H}'_{pq}}{\mathbf{H}_{pq}} + 2\frac{(\mathbf{H}'(\mathbf{R}_c+\mathbf{R}_d))_{pq}\mathbf{H}_{pq}}{\mathbf{H}'_{pq}} = 0,
\end{aligned} \tag{22}
$$

which gives rise to the updating rule in (13). Following Lemma 1, under this updating rule the objective function values of $J(\mathbf{H})$ in (7) will be non-increasing. □

Bulk of the computation depends on the matrix multiplication in the updating rules (13) and (6), the complexity of updating $\mathbf{W}^{(i)}$ and $\mathbf{H}^{(i)}$ is $\mathcal{O}(mnk^{(i)})$ and $\mathcal{O}(\sum_{j=1,j\neq i}^{V}(k^{(j)}n^2 + nk^{(j)^2}))$, respectively. So the overall computation of MCNMF is $\mathcal{O}(\sum_{i=1}^{V}(\sum_{j=1,j\neq i}^{V} k^{(j)}n^2 + k^{(i)}mn))$.

4 Experiments

4.1 Dataset

We carried several experiments on the following benchmark datasets to show the effectiveness of MCNMF. The Yale contains 11 facial images for each of 15 subjects. Sample images are shown in Figure 1. For each subject, its face images are either in different facial expressions (such as happy or sad), or configurations (such as with or without glasses). The ORL[2] dataset consists of 400 facial images belonging to 40 different subjects. Similar to the Yale dataset, the images were taken with various lighting and facial expressions. The Notting-Hill [Cao *et al.*, 2015] is a video face dataset, which is derived from the movie "Notting Hill". The faces of 5 main casts were used, including 4660 faces in 76 tracks. The COIL20 image library[3] is composed of 1440 images for 20 objects. The 72 images of each object were captured by a fixed camera at a pose intervals of 5 degree. For this dataset, we regard the different poses and shapes as components.

4.2 Experiment Setup

We first compared MCNMF against the standard NMF [Lee and Seung, 2001] to verify the effectiveness of exploring diverse information from multi-components, and then with the state-of-the-arts: RNMF [Kong *et al.*, 2011], GNMF [Cai *et al.*, 2011], Cauchy NMF [Liutkus *et al.*, 2015] and LANMF [Liu *et al.*, 2016]. For each compared method, the parameters were set according to the parameter settings in original papers. For MCNMF, we varied the regularization parameter α within [0.01, 0.05] with 0.01 interval and fixed the number of components $V = 3$ (more discussion in next subsection). In addition, we set each $k^{(i)}$ equals to number of clusters according to the groundtruth of each dataset. The dimensions of obtained optimal representations $\mathbf{H}^*$ for all the compared methods were all set to be $k = \sum_{i=1}^{V} k^{(i)}$ for fair comparison.

4.3 Performance Analysis

Clustering result. We applied k-means to the obtained representations $\mathbf{H}^*$ for clustering. Since k-means is sensitive to the initial values, we repeated the clustering process 50 times to give the average performance. Moreover, since all the compared methods converge to local minimum, we ran each method 10 times to avoid randomness. Similar to the work [Cao *et al.*, 2015], we adopted three widely used evaluation metrics accuracy (AC) [Liu *et al.*, 2012], normalized mutual information (NMI) [Liu *et al.*, 2012] and purity [Ding *et al.*, 2006] to assess the quality of the results with a comprehensive evaluation. The final average clustering results along with standard deviations are reported in Table 1. As we can see, MCNMF outperforms the other methods against all metrics and gets the lowest standard deviations on 9 of 12 results, which demonstrate the robustness of MCNMF. Besides, it can be noticed that GNMF performs the second best in terms of AC, but not for other metrics. Especially, MCNMF outperforms GNMF with a large margin: 4.84% and 3.35% on the Yale and ORL, respectively. This is probably because that the images in both of the two datasets have more components, such as different lighting and expressions. Obviously, richer information has been explored and obtained for comprehensive representations, which brings significant improvements.

Component study. We closely examined the learned representations for each component to analyze their latent semantics. In particular, we took the Yale dataset as an example. Like the previous experiment setting, we fixed the number of components V to 3, and applied k-means on the representation $\mathbf{H}^{(i)}$ of each component to cluster the data into 3 clusters. The result is shown in Figure 2. We can see that the each representation has effectively captured some distinct information (such as unhappy or surprised expressions) which is reflected by a corresponding cluster. This result enables the understanding of the data from various perspectives in a semantic level, which would be hardly achievable by current NMF-based methods as they cannot identify components. Also, note that from $\mathbf{H}^{(2)}$ and $\mathbf{H}^{(3)}$, there is a common cluster: the right-lit faces. This is reasonable that although multiple representations usually describe data from different perspectives, they are not completely exclusive to each other mutually. We further tested MCNMF on a larger dataset COIL20. Again, the results are quite good and promising, with multiple clusters being obtained through different components (right rotation, pottery, etc). Figure 3 shows example results, due to page limitation.

Parameter analysis. We tested the effect of parameter α of MCNMF on the datasets. α varies from 0.01 to 0.05 with an increment of 0.01. Here we presented the accuracy of MCNMF with respect to α on Yale and ORL as examples. Seen from Figure 4, the accuracy varies slightly showing a relatively stable performance. Also, for all values of α, the performance of MCNMF is consistently better than NMF (Table 1). For example, for Yale, the worst result of MCNMF is about 0.4182, while NMF only gets 0.4048.

We also tested the effect of the number of components V. Here we fixed $\alpha = 0.01$ and varied V from 1 to 7 with an increment of 1. Seen from Figure 5, for both Yale and ORL, the accuracy with multiple components ($V \geq 2$) is always better than MCNMF with $V = 1$ (NMF). Specifically, the accuracy increases sharply when V is tuned from 1 to 3, which indicates the effectiveness of MCNMF by exploring multiple components. Then the accuracy fluctuates slightly when V increases from 3 to 7. The fluctuation could be due to a compromise between the amount of features for each representation and the diverse information among them. When V

[2]http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

[3]http://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php

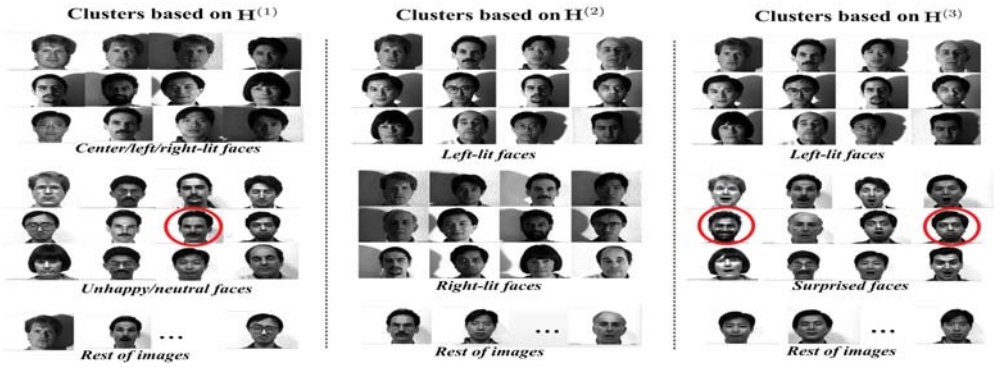

Figure 2: Sample clustering results of the Yale dataset based on each representation $\mathbf{H}^{(i)}$. Images circled in red are outliers.

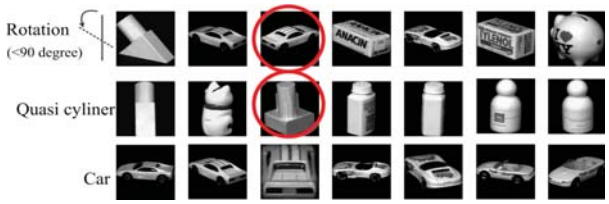

Figure 3: Sample clustering results of the COIL20 dataset. Each row, from top to bottom, represents a cluster based on the representations $\mathbf{H}^{(1)}$, $\mathbf{H}^{(2)}$ and $\mathbf{H}^{(3)}$, respectively. Images circled in red are outliers.

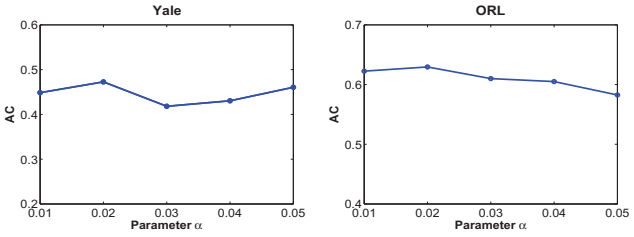

Figure 4: The effect of the parameter α.

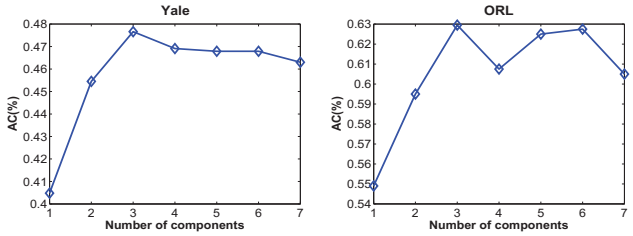

Figure 5: The effect of the number of components V.

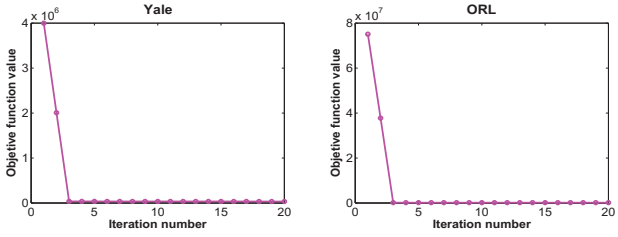

Figure 6: Convergence curves.

increases, more diverse information can be utilized. However, given a fixed $k = \sum_{i=1}^{3} k^{(i)}$, the increase of V will result in reduction of the feature dimension $k^{(i)}$ for each representation.

Convergence analysis. Having proven the convergence of our update rules of MCNMF in previous sections, here we experimentally demonstrate its convergence in Figure 6, where the horizontal axis is the number of iterations and the vertical axis is the value of objective function. It can be seen that the objective function values are non-increasing and drop sharply within 5 iterations on both datasets.

5 Conclusion

In this paper, we have proposed a Multi-Component Nonnegative Matrix Factorization (MCNMF) approach to find multi-representation of data by exploring embedded latent components. Different from existing NMF-based approaches that seek for a single representation matrix, MCNMF learns multiple representations simultaneously. Utilizing Hilbert Schmidt Independence Criterion (HSIC) as a penalty term, MCNMF explicitly enforces the diversity of different data representations. Extensive experiments have demonstrated that MCNMF can not only obtain multiple representations with each one reflecting one property of data, but also improves the accuracy by aggregating multiple representations. For future work, we will extend MCNMF to semi-supervised MCNMF by utilizing the label information of data to obtain a much clearer correspondence with the real component of data.

Acknowledgements

This work was supported by Royal Society Newton Mobility grant, EU H2020 project (No. 691215) and National Natural Science Foundation of China (No. 61503281, 61502334).

References

[Cai *et al.*, 2011] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1548–1560, 2011.

[Cao *et al.*, 2015] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–594, 2015.

[Changpinyo *et al.*, 2013] Soravit Changpinyo, Kuan Liu, and Fei Sha. Similarity component analysis. In *Advances in Neural Information Processing Systems*, pages 1511–1519, 2013.

[Ding *et al.*, 2006] 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. ACM, 2006.

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

[Gretton *et al.*, 2005] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[Guan *et al.*, 2012] Naiyang Guan, Dacheng Tao, Zhigang Luo, and John Shawe-Taylor. Mahnmf: Manhattan non-negative matrix factorization. *arXiv preprint arXiv:1207.3438*, 2012.

[Kong *et al.*, 2011] Deguang Kong, Chris Ding, and Heng Huang. Robust nonnegative matrix factorization using l21-norm. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 673–682. ACM, 2011.

[Lee and Seung, 1999] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[Lee and Seung, 2001] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[Liu *et al.*, 2012] Haifeng Liu, Zhaohui Wu, Xuelong Li, Deng Cai, and Thomas S Huang. Constrained nonnegative matrix factorization for image representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1299–1311, 2012.

[Liu *et al.*, 2013] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proc. of SDM*, volume 13, pages 252–260. SIAM, 2013.

[Liu *et al.*, 2016] Tongliang Liu, Mingming Gong, and Dacheng Tao. Large-cone nonnegative matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[Liutkus *et al.*, 2015] Antoine Liutkus, Derry Fitzgerald, and Roland Badeau. Cauchy nonnegative matrix factorization. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*, pages 1–5. IEEE, 2015.

[Niu *et al.*, 2010] Donglin Niu, Jennifer G Dy, and Michael I Jordan. Multiple non-redundant spectral clustering views. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 831–838, 2010.

[Ou *et al.*, 2015] Mingdong Ou, Peng Cui, Fei Wang, Jun Wang, and Wenwu Zhu. Non-transitive hashing with latent similarity components. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 895–904. ACM, 2015.

[Qian *et al.*, 2016] Wei Qian, Bin Hong, Deng Cai, Xiaofei He, Xuelong Li, et al. Non-negative matrix factorization with sinkhorn distance. International Joint Conferences on Artificial Intelligence, 2016.

[Song *et al.*, 2007] Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 823–830. ACM, 2007.

[Tao *et al.*, 2009] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. Geometric mean for subspace selection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):260–274, 2009.

[Wang *et al.*, 2016a] Jing Wang, Xiao Wang, Feng Tian, Chang Hong Liu, Hongchuan Yu, and Yanbei Liu. Adaptive multi-view semi-supervised nonnegative matrix factorization. In *International Conference on Neural Information Processing*, pages 435–444. Springer, 2016.

[Wang *et al.*, 2016b] Yuanyuan Wang, Shuyi Wu, Bin Mao, Xiang Zhang, and Zhigang Luo. Correntropy induced metric based graph regularized non-negative matrix factorization. *Neurocomputing*, 204:172–182, 2016.

[Zhang and Zhou, 2010] Yin Zhang and Zhi-Hua Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):14, 2010.