Pattern Recognition Letters 98 (2017) 68–74

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Adaptive total-variation for non-negative matrix factorization on manifold

Chengcai Leng[a,b,c], Guorong Cai[c,d,*], Dongdong Yu[c], Zongyue Wang[d]

[a] School of Mathematics, Northwest University, Xi'an 710127, China
[b] School of Mathematics and Information Sciences, Nanchang Hangkong University, Nanchang 330063, China
[c] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[d] College of Computer Engineering, Jimei University, Xiamen 361021, China

ARTICLE INFO

Article history:
Received 20 September 2016
Available online 26 August 2017

JEL classification:
41A05
41A10
65D05
65D17

Keywords:
Adaptive total variation
Non-negative matrix factorization
Manifold learning

ABSTRACT

Non-negative matrix factorization (NMF) has been widely applied in information retrieval and computer vision. However, its performance has been restricted due to its limited tolerance to data noise, as well as its inflexibility in setting regularization parameters. In this paper, we propose a novel sparse matrix factorization method for data representation to solve these problems, termed Adaptive Total-Variation Constrained based Non-Negative Matrix Factorization on Manifold (ATV-NMF). The proposed ATV can adaptively choose the anisotropic smoothing scheme based on the gradient information of data to denoise or preserve feature details by incorporating adaptive total variation into the factorization process. Notably, the manifold graph regularization is also incorporated into NMF, which can discover intrinsic geometrical structure of data to enhance the discriminability. Experimental results demonstrate that the proposed method is very effective for data clustering in comparison to the state-of-the-art algorithms on several standard benchmarks.

1. Introduction

Matrix Factorization (MF) plays the fundamental role in various emerging applications ranging from information retrieval to data mining [1]. It typically adopts a sparse representation to obtain low-dimensional matrix, which can deal with many classical classification and clustering problems efficiently and robustly [2–4]. In order to avoid the curse of dimensionality, different forms of dimensionality reduction schemes like Principal Component Analysis (PCA), ISOMAP [5], Locally Linear Embedding (LLE) [6], Laplacian Eigenmap [7] and Isometric Projection [8]. NMF [9] incorporates the non-negativity constraint to achieve a parts-based representation.

NMF allows only additive, not subtractive, combination of the original data, and which is effective to capture the underlying structure of the data combining non-negative constraints in a parts-based low dimensional space. Usually, the rank of the NMF is generally chosen so that the matrix factorization can be regarded as a compressed form of the data [9,10]. NMF has been widely used for clustering [11,12], face recognition [13–15] and image or data analysis [2,16]. To overcome the difficulty in modeling the intrinsic geometrical structure, Manifold learning [4–6,17,18] has been introduced into NMF. For instance, Cai et al. [19] presented a graph regularized NMF (GNMF) by adding a graph manifold term to NMF, While promising, manifold-based NMF is typically sensitive to data noise.

Since NMF model does not consider noise signal, its performance has been restricted due to the fact that it is very hard to determine appropriate regularization parameters. In order to resolve these problems, we will add an adaptive total variation regularization item to NMF model. It is worth noting that Total variation (TV), first introduced by Rudin et al. [20], is effective for piecewise constant reconstruction, thus can preserve the boundary of large objects well. Since then TV regularization has been widely used for denoising tasks in image processing, computer vision and image reconstruction, such as data representation [21], face recognition [15,22]. To this end, total variation scheme has been proposed to handle data noise by combining TV term [23,24]. However, TV based NMF cannot well discover and reveal the intrinsic geometrical and structure information of data and it is difficult to fix the TV regularization parameter of TV term.

In this paper, we present a novel NMF scheme that correctly handles the data noise as well as modeling the intrinsic geometric structure of data, terms Adaptive Total-Variation Constrained based Non-negative Matrix Factorization on Manifold (ATV-NMF). First, in order to discover intrinsic geometrical structure, we

* Corresponding author at: College of Computer Engineering, Jimei University, Xiamen 361021, China.
E-mail address: guorongcai.jmu@gmail.com (G. Cai).

http://dx.doi.org/10.1016/j.patrec.2017.08.027
0167-8655/

incorporate the graph regularization to NMF. Second, the Adaptive Total-Variation (ATV) regularization is incorporated to choose adaptively the anisotropic smoothing scheme based on the data gradient to denoise or preserve adaptively the feature details. ATV also avoids choosing the regularization parameter to enhance the discrimination ability. Finally, we present a novel iterative update rule that achieve ATV-NMF. Experimental results show that the proposed method is better compared to the state-of-the-art schemes for data clustering.

The rest of this paper is organized as follows: In Section 2, we propose the ATV-NMF on manifold method. Section 3 presents experimental results and Section 4 gives conclusions and future work.

2. ATV-NMF On manifold

In this section, we first describe the basic idea of ATV method. In principle, ATV and graph regularization is introduced into NMF to preserve edge or details, as well as to discover and enhance the intrinsic geometrical data structure to improve the discriminability. As for data clustering, the database is regarded as an $m \times n$ matrix V, each column of which contains m non-negative values of one of the n images. Then the task of ATV-NMF is to construct approximate factorizations of the form $V = WH$, where W and H are respectively $m \times r$ and $r \times n$ matrix factors, and r denotes the rank of the factorization.

2.1. Adaptive total-variation

Our ATV-NMF model is inspired by the adaptive total variation regularization proposed in [25] so that the proposed model can adaptively choose the anisotropic smoothing scheme based on the gradient information of data to denoise or preserve feature details, which can be defined as:

$$E(H) = ||H||_{ATV} \tag{1}$$

where E is the energy function of H, $||H||_{ATV} = \int_\Omega \frac{1}{p(x,y)} |\nabla H|^{p(x,y)} dxdy$ denotes the adaptive TV regularization term, $p(x,y) = 1 + \frac{1}{1+|\nabla H|^2}$, $1 < p(x,\ y) < 2$, $(\nabla H)(i,j) = ((\partial_x H)(i,j), (\partial_y H)(i,j))$ is a discrete gradient form with $(\partial_x H)(i,j)$ and $(\partial_y H)(i,j)$, given as follows:

$$(\partial_x H)(i,j) = \begin{cases} H(i+1,j) - H(i,j) & if\ i < r \\ H(1,j) - H(r,j) & if\ i = r \end{cases}$$

$$(\partial_y H)(i,j) = \begin{cases} H(i,j+1) - H(i,j) & if\ j < n \\ H(i,1) - H(i,n) & if\ j = n \end{cases}$$

The adaptive TV regularization including a diffusion coefficient $\frac{1}{|\nabla H|^{2-p}}$ in Eq. (8), which is used to control the speed of the diffusion based on the gradient information. For edges, $|\nabla H|^{2-p}$ has big values, the $\frac{1}{|\nabla H|^{2-p}}$ is small and the diffusion is very weak along the edge directions, which helps preserve edges. In a smooth region, $|\nabla H|^{2-p}$ has small values, the $\frac{1}{|\nabla H|^{2-p}}$ is big and the diffusion is strong, which helps in denoising. In addition, the ATV model has some fundamental properties, which has numerical stability solution, can avoid the staircase effect, and is able to preserve or enhance finer scale data features, such as edges or textures, while denoising [25].

2.2. Multiplicative updating rules

Using the ATV as the regularization term, the refined ATV-NMF model is designed by solving the following objective function:

$$O_{ATV-NMF} = ||V - WH||_F^2 + \lambda Tr(HLH^T) + 2||H||_{ATV}. \quad s.t.\ W \geq 0, H \geq 0 \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\lambda \geq 0$ is a regularization parameter, $Tr(.)$ denotes the trace of a matrix, S is the weight matrix whose entry S_{ij} measures the similarity between each vertex pair (v_i, v_j), D is a diagonal matrix with column sums of S as its diagonal entries. i.e., $D_{ij} = \sum_{i=1}^{n} S_{ij}$, $L = D - S$ is called graph Laplacian matrix [26].

Since the objective function $O_{ATV-NMF}$ in Eq. (2) is not convex in W and H, we therefore resort to an iterative updating algorithm to obtain an approximate optimal solution of $O_{ATV-NMF}$. In order to obtain the solution of the objective function $O_{ATV-NMF}$ in Eq. (2), we need to find an iterative updating algorithm to achieve the minimization of $O_{ATV-NMF}$ by gradient descent algorithm [27]. The gradient of the objective function $O_{ATV-NMF}$ with respect to W and H are given as follows:

$$\frac{\partial O_{ATV-NMF}}{\partial W_{i,l}} = -2(VH^T - WHH^T)_{i,l} \tag{3}$$

$$\frac{\partial O_{ATV-NMF}}{\partial H_{l,j}} = -2\left(W^TV - W^TWH - \lambda HL + div\left(\frac{\nabla H}{|\nabla H|^{2-p}}\right)\right)_{l,j} \tag{4}$$

The additive update rules for problem (2) by Eqs. (3) and (4) can be obtained as follows:

$$W_{i,l} \leftarrow W_{i,l} + \xi_{i,l}(VH^T - WHH^T)_{i,l} \tag{5}$$

$$H_{l,j} \leftarrow H_{l,j} + \eta_{l,j}\left(W^TV - W^TWH - \lambda HL + div\left(\frac{\nabla H}{|\nabla H|^{2-p}}\right)\right)_{l,j} \tag{6}$$

where $\xi_{i,l} = \frac{W_{i,l}}{(WHH^T)_{i,l}}$ and $\eta_{l,j} = \frac{H_{l,j}}{(W^TWH + \lambda HD)_{l,j}}$ are the step sizes of the updates, and the multiplicative updating rules can be formulated as follows:

$$W_{i,l} \leftarrow W_{i,l} \frac{(VH^T)_{i,l}}{(WHH^T)_{i,l}} \tag{7}$$

$$H_{l,j} \leftarrow H_{l,j} \frac{(W^TV + \lambda HS + div(\frac{\nabla H}{|\nabla H|^{2-p}}))_{l,j}}{(W^TWH + \lambda HD)_{l,j}} \tag{8}$$

where div denotes the divergence, i.e., $div = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y})$, $\nabla H = (\partial_x H, \partial_y H)$ denotes the gradient, and $|\nabla H| = \sqrt{(\partial_x H)^2 + (\partial_y H)^2}$ is the norm of the gradient. The similar form of the Eq. (8) can be found in [22], and the discrete form of the $div(\frac{\nabla H}{|\nabla H|^{2-p}})$ can also be found based on the operator of the divergence and the gradient by using total variation principal [25]. The derivation of Eq. (8) is given as belows.

Note that Eq. (6) is the additive update rule, where $\eta_{l,j} = \frac{H_{l,j}}{(W^TWH + \lambda HD)_{l,j}}$. Let $L = D - S$ be the graph Laplacian matrix [26], thus we have:

$$H_{l,j} \leftarrow H_{l,j} + \eta_{l,j}\left(W^TV - W^TWH - \lambda HL + div\left(\frac{\nabla H}{|\nabla H|^{2-p}}\right)\right)_{l,j}$$

$$H_{l,j} \leftarrow H_{l,j} + \eta_{l,j}\left(W^TV - W^TWH - \lambda H(D-S) + div\left(\frac{\nabla H}{|\nabla H|^{2-p}}\right)\right)_{l,j}$$

$$H_{l,j} \leftarrow H_{l,j} + \eta_{l,j}\left(-W^TWH - \lambda HD + W^TV + \lambda HS + div\left(\frac{\nabla H}{|\nabla H|^{2-p}}\right)\right)_{l,j}$$

$$H_{l,j} \leftarrow H_{l,j} + \eta_{l,j}(-W^TWH - \lambda HD)_{l,j} + \eta_{l,j}\left(W^TV + \lambda HS + div\left(\frac{\nabla H}{|\nabla H|^{2-p}}\right)\right)_{l,j}$$

$$H_{l,j} \leftarrow \eta_{l,j}\left(W^TV + \lambda HS + div\left(\frac{\nabla H}{|\nabla H|^{2-p}}\right)\right)_{l,j}$$

As a consequent, we have $H_{l,j} \leftarrow H_{l,j} \frac{(W^T V+\lambda HS+div(\frac{\nabla H}{|\nabla H|^{2-p}}))_{l,j}}{(W^T WH+\lambda HD)_{l,j}}$. The detailed multiplicative updating procedure is summarized in Algorithm 1.

Algorithm 1 ATV-NMF algorithm.

Input: $V \in R^{m\times n}$, D , S and $1 \le r \le \min\{m, n\}$.
Initialization: W_0 , H_0 , λ and $k = 0$.
For $k = 0, 1, \ldots$ until convergence or maximum iteration.
Update H^{k+1} according to
$H^{k+1} = H^k \frac{(W^T V+\lambda HS+div(\frac{\nabla H}{|\nabla H|^{2-p}}))^k}{(W^T WH+\lambda HD)^k}$
Update W^{k+1} according to
$W^{k+1} = W^k \frac{(VH^T)^k}{(WHH^T)^k}$
$k = k + 1$
Output: $W \in R^{m\times r}$, $H \in R^{r\times n}$.

2.3. Convergence analysis

In the following subsection, the convergence is analyzed according to the multiplicative updating rules in Eqs. (7) and (8).

Definition: $G(x, x')$ is an auxiliary function of $F(x)$ if the conditions $G(x, x') \ge F(x)$ and $G(x, x) = F(x)$ are satisfied.

Lemma 1. *If G is an auxiliary function of F , then F is non-increasing under the update rule:*

$$x^{t+1} = \arg\min_x G\left(x, x^t\right) \tag{9}$$

Proof. $F(x^{t+1}) \le G(x^{t+1}, x^t) \le G(x^t, x^t) = F(x^t)$. □

Considering an element w_{ab} in W, we use $F_{w_{ab}}$ to denote the part of the objective $O_{ATV-NMF}$ which is only relevant to w_{ab}. From which one can see that:

$$\begin{aligned} F'_{w_{ab}} &= \left(\frac{\partial O_{ATV-NMF}}{\partial W}\right)_{ab} \\ &= (-2VH^T + 2WHH^T)_{ab} \end{aligned} \tag{10}$$

and

$$F''_{w_{ab}} = \left(\frac{\partial^2 O_{ATV-NMF}}{\partial W^2}\right)_{ab} = (2HH^T)_{bb}. \tag{11}$$

Lemma 2. *If $G(w, w^t_{ab})$ satisfies*

$$\begin{aligned} G(w, w^t_{ab}) = F_{w_{ab}}(w^t_{ab}) + F'_{w_{ab}}(w^t_{ab})(w - w^t_{ab}) \\ + \frac{(WHH^T)_{ab}}{w^t_{ab}}(w - w^t_{ab})^2, \end{aligned} \tag{12}$$

then $G(w, w^t_{ab})$ is an auxiliary function of $F_{w_{ab}}$.

Proof. Obviously, $G(w, w) = F_{w_{ab}}(w)$. According to the definition of auxiliary function, we need to prove $G(w, w^t_{ab}) \ge F_{w_{ab}}(w)$. Therefore, we expand the Taylor series of $F_{w_{ab}}(w)$ as follows:

$$\begin{aligned} F_{w_{ab}}(w) = F_{w_{ab}}(w^t_{ab}) + F'_{w_{ab}}(w^t_{ab})(w - w^t_{ab}) \\ + [(HH^T)_{bb}](w - w^t_{ab})^2. \end{aligned} \tag{13}$$

By combing Eqs. (12) and (13), one can see that $G(w, w^t_{ab}) \ge F_{w_{ab}}(w)$ is equivalent to:

$$\frac{(WHH^T)_{ab}}{w^t_{ab}} \ge (HH^T)_{bb}, \tag{14}$$

Table 1
Data information of the four data sets.

data sets	size	dimensionality	classes
COIL20	1440	1024	20
ORL	400	1024	40
PIE	2856	1024	68
Yale	165	4096	15

Therefore, we have:

$$(WHH^T)_{ab} = \sum_{l=1}^{r} w^t_{al}(HH^T)_{lb} \ge w^t_{ab}(HH^T)_{bb} \tag{15}$$

Thus Eq. (14) is derived accordingly. □

Since Eq. (12) is auxiliary function for $F_{w_{ab}}$, $F_{w_{ab}}$ is non-increasing under the updating rule in Eq. (7). The updating rules of Eq. (8) has the similar form as the reference [22], therefore, we can similarly construct the auxiliary function $G(h, h^t_{ab})$ for $F_{h_{ab}}$, and the detailed proofs for convergence under the updating rule for H in Eq. (8) can be also followed by Yin and Liu [22].

Theorem 1. *The objective function $O_{ATV-NMF}$ in Eq. (2) is non-increasing under the multiplicative updating rules of Eqs. (7) and (8).*

Proof. According to Lemmas 1 and 2, $G(w, w^t_{ab})$ is an auxiliary function of $F_{w_{ab}}$. As a consequent, $F_{w_{ab}}$ is non-increasing under the update equation $w^{t+1}_{ab} = \arg\min_w G(w, w^t_{ab})$.

Therefore, we have:

$$\frac{\partial G(w, w^t_{ab})}{\partial w} = F'_{w_{ab}}(w^t_{ab}) + \frac{2(WHH^T)_{ab}}{w^t_{ab}}(w - w^t_{ab}) = 0.$$

That is:

$$(-2VH^T + 2WHH^T)_{ab} + 2\frac{(WHH^T)_{ab}}{w^t_{ab}}(w - w^t_{ab}) = 0.$$

Consequently,

$$w = w^t_{ab}\frac{(VH^T)_{ab}}{(WHH^T)_{ab}}.$$

By substituting w into Eq. (9), we then have:

$$w^{t+1}_{ab} = \arg\min_w G(w, w^t_{ab}) = w^t_{ab}\frac{(VH^T)_{ab}}{(WHH^T)_{ab}}.$$

Similarly, we have:

$$\begin{aligned} h^{t+1}_{ab} &= \arg\min_h G(h, h^t_{ab}) \\ &= h^t_{ab}\frac{(W^T V + \lambda HS + div(\frac{\nabla H}{|\nabla H|^{2-p}}))_{ab}}{(W^T WH + \lambda HD)_{ab}}. \end{aligned}$$

□

Theorem 1 guarantees that the objective function $O_{ATV-NMF}$ in Eq. (2) converges to a local optimum under the multiplicative updating rules in Eqs. (7) and (8).

In addition, the related theory can also be found in [28].

3. Experimental results

In this section, we introduce some experimental evaluation on the task of data clustering to demonstrate the efficiency and effectiveness of the proposed ATV-NMF algorithm. The results have been compared with state-of-the-art methods, including NMF [9] and Graph-regularized NMF (GNMF) [19].

Table 2
Clustering results on COIL 20 and ORL dataset.

Data	k	Clustering Accuracy (%)			Normalized MI (%)		
		NMF	GNMF	ATV-NMF	NMF	GNMF	ATV-NMF
COIL20	4	50.417	72.292	**84.306**	67.657	86.312	**90.998**
	7	62.431	77.569	**79.514**	73.962	**89.828**	88.806
	10	60.764	75.764	**79.236**	71.334	87.889	**89.069**
	13	62.778	72.361	**73.958**	71.822	87.017	**87.261**
	16	64.167	**80.417**	76.111	74.747	**90.025**	86.968
	19	64.236	74.167	**82.847**	71.947	86.271	**90.684**
	20	66.736	79.306	**79.722**	74.361	88.515	**89.743**
	Avg.	61.647	75.982	**79.385**	72.261	87.980	**89.076**
ORL	2	47.250	48.750	**53.250**	67.609	69.300	**70.347**
	3	45.250	48.500	**51.500**	66.246	67.765	**69.802**
	4	43.500	46.750	**49.250**	67.320	66.512	**69.026**
	5	45.750	46.500	**47.750**	67.214	68.262	**69.159**
	6	46.000	46.250	**47.500**	66.241	67.816	**69.167**
	7	47.500	45.750	**49.750**	68.173	67.292	**68.822**
	8	48.500	47.500	**53.250**	68.429	69.012	**70.550**
	9	49.000	47.000	**53.500**	69.818	66.643	**70.564**
	10	46.750	43.500	**47.500**	**68.742**	65.160	68.133
	Avg.	46.611	46.722	**50.361**	67.755	67.529	**69.508**

Table 3
Approximation reconstruction error on the COIL20 for the different cluster numbers.

Methods	Approximation reconstruction error			
	k = 5	k = 10	k = 15	k = 20
GNMF	157.253	156.126	156.060	156.303
ATV-NMF	155.451	153.595	153.532	153.363

3.1. Evaluation metrics

In order to evaluate the effectiveness parts-based representation for the methods mentioned above. The following two popular evaluation metrics are used to evaluate the clustering performance. The first performance measure is the Clustering Accuracy (ACC), which is defined as [11,29,30]

$$ACC = \frac{\sum_{i=1}^{n} \delta(s_i, map(r_i))}{n}$$

where s_i is the true class label and r_i is the obtained cluster label of x_i, n is the total number of documents, $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and $map(\cdot)$ is the mapping function that maps each label r_i to the equivalent label from the data corpus. A larger ACC indicates a better clustering performance [31].

The second evaluation metric is the Normalized Mutual Information (NMI), which is defined as [29,31]

$$NMI(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))}$$

where C is a set of the true labels, and C' is a set of clusters obtained from the clustering algorithms. $H(C)$ and $H(C')$ are the entropies of C and C', respectively, and $MI(C, C')$ is the mutual information between two sets of clusters C and C', which is defined as [30]

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)}$$

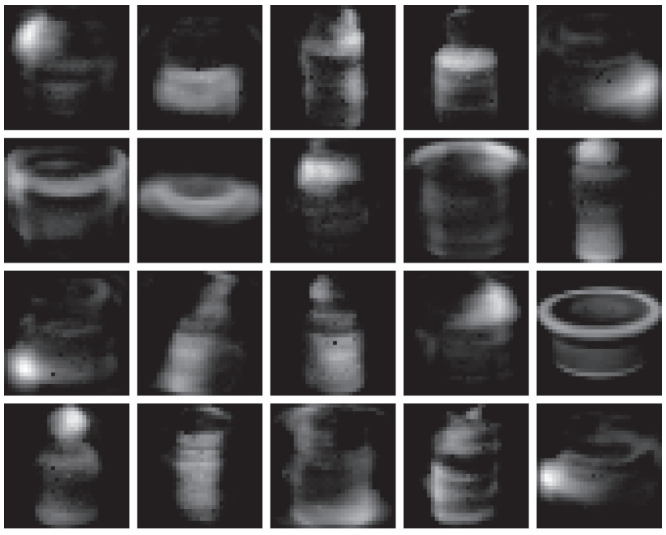
where $p(c_i)$ and $p(c'_j)$ are the probabilities of a document belonging to the clusters c_i and c'_j, respectively. $p(c_i, c'_j)$ denotes the joint probability that this arbitrarily selected document belongs to the clusters c_i as well as c'_j at the same time. The higher the NMI score, the better the clustering quality [31].

To show the data clustering performance and the experiments have been conducted on four widely used benchmarks, namely COIL20, ORL, PIE and Yale. Our main task is to generate the cluster label of each sample via the represent process, and then the results are compared with the ground truth.

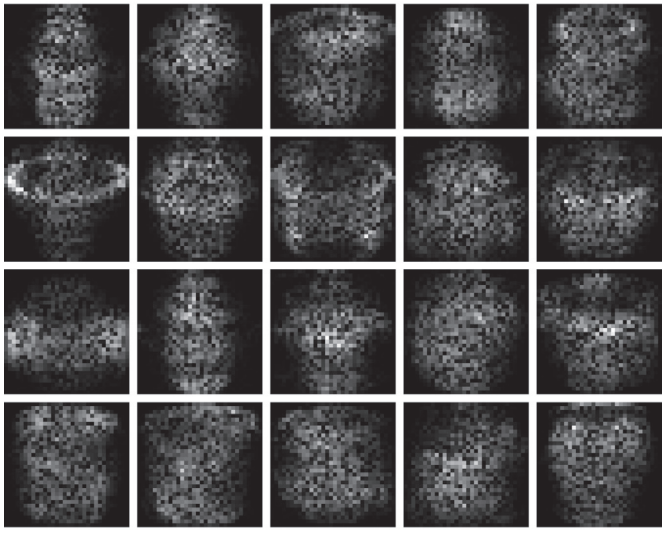
As for the clustering process, we first initialize the parameters, including W_0 , H_0 and λ, randomly. Second, W^k and H^k are updated via the multiplicative iterate rules. As a consequent, W

Table 4
Clustering results on PIE and Yale dataset.

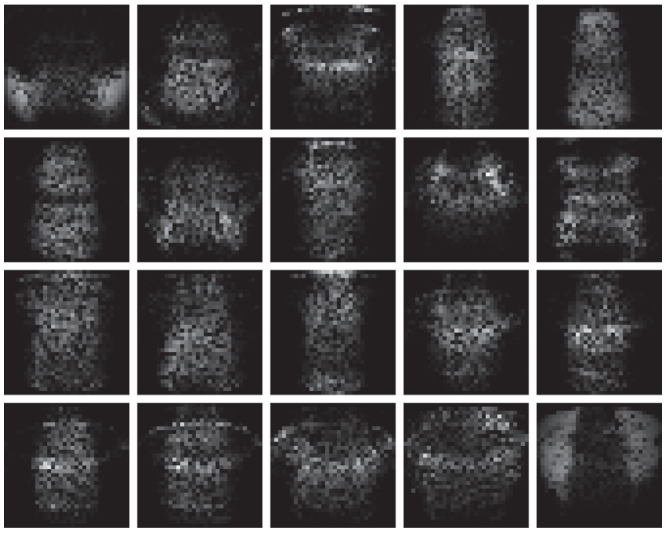
Data	k	Clustering Accuracy (%)				Normalized Mutual Information (%)			
		NMF	GNMF	DNMF	ATV-NMF	NMF	GNMF	DNMF	ATV-NMF
PIE	30	31.092	33.718	35.714	**37.651**	58.261	59.375	60.217	**61.917**
	40	31.653	32.703	36.224	**38.861**	58.507	58.590	60.661	**61.854**
	50	33.718	34.594	35.014	**37.006**	59.084	59.193	59.907	**62.667**
	60	33.333	34.944	36.520	**37.916**	59.291	59.435	60.656	**62.247**
	68	32.143	33.193	34.209	**36.111**	57.788	59.177	59.207	**61.327**
	Avg.	32.388	33.830	35.536	**37.509**	58.586	59.154	60.130	**62.002**
Yale	3	48.485	50.303	51.515	**52.121**	51.957	53.735	**54.847**	54.674
	6	46.061	49.697	52.121	**53.545**	50.665	52.991	53.192	**54.203**
	9	46.878	47.879	48.485	**52.333**	50.189	51.371	53.156	**54.236**
	12	48.485	49.091	50.303	**55.182**	52.621	52.871	54.541	**56.548**
	15	46.061	47.879	49.091	**51.515**	51.659	52.006	53.809	**55.042**
	Avg.	47.194	48.970	50.303	**52.939**	51.418	52.595	53.909	**54.941**



(a) Basis vectors learned by NMF



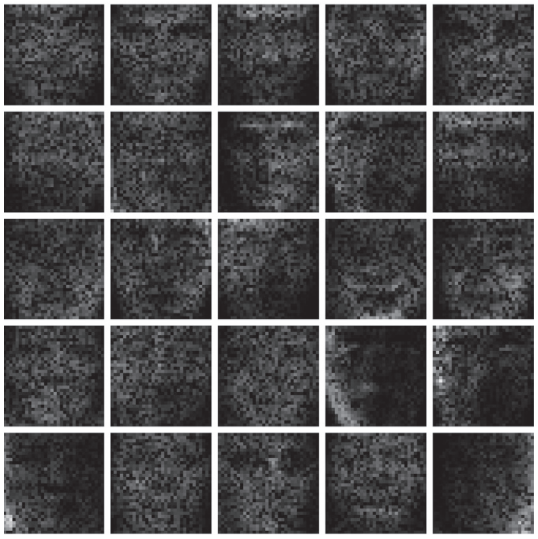
(b) Basis vectors learned by GNMF



(c) Basis vectors learned by ATV-NMF

Fig. 1. Basis vectors (column vectors of W) learned from the COIL20 dataset.

(a) Basis vectors learned by NMF



(b) Basis vectors learned by GNMF

(c) Basis vectors learned by ATV-NMF

Fig. 2. Basis vectors (column vectors of W) learned from the ORL dataset.

and weight coefficient H can be obtained via the iteration process. Finally, the clustering results are derived from W and H.

3.2. Data sets

As for the datasets, the first one is the COIL20 image data set and the second one is the ORL image data set. The third data set is the CMU PIE face database and the fourth Yale database. Table 1 summarizes the statistics for both data sets, and more details are introduced as below:

- *COIL20*. This data set contains 32×32 gray scale images of 20 objects viewed at varying angles, with each object having 72 images.
- *ORL Database*. The ORL database is consisted of 40 distinct subjects each with 10 images, which are taken at different times, varying the lighting, with different facial expression and facial details (with glasses or no glasses). Each image is 32×32 pixels with 256 gray levels per pixel [30].
- *CMU PIE face database*. This database contains 32×32 gray scale images of 68 people. Each person has 42 facial images under different light and illumination conditions [19].
- *Yale database*. Yale database contains 165 grayscale images with the size of 64×64 in GIF format of 15 subjects each with 11 images, with different facial expression or configuration: center-light, glasses, happy, left-light, no glasses, normal, right-light, sad, sleepy, surprised, and wink.

3.3. Clustering comparisons

We first compared our method with the other related methods such as NMF and GNMF on two well known datasets. In all experimental results, the F-norm formulation has been used to measure the quality of the approximation accuracy [32]. As for Graph-regularized NMF (GNMF) and ATV-NMF, we use the 0-1 weighting scheme for constructing the k -nearest neighbor graph, where $k = 5$ and the parameter λ is set to be 100, which were recommended in [19].

Table 2 shows the clustering results of the three algorithms on the COIL20 and the ORL data sets, which have been normalized and measured by the *ACC* and the *NMI*. The COIL20 data set is tested for the different cluster numbers for iteration 100 times, and the iteration times are also set to 100 on the second ORL data set with different cluster numbers. The NMF model is the worst of the three clustering algorithms for the first COIL20 data set. The ATV-NMF model is better than GNMF and NMF, this is because the adaptive TV regularization can choose adaptively the anisotropic smoothing scheme based on the gradient information of data, or to denoise or preserve adaptively the feature details of data. The GNMF algorithm achieves better performance than NMF which shows that the geometric structure of the data in learning is preserved by incorporating manifold graph regularization.

As for the ORL data set, our method outperforms GNMF and NMF significantly. Overall, the performance is degenerated with the increasing number of clusters. Under such a circumstance, the GNMF algorithm does not work well as shown in Table 2. The best results are bold faced, from which we have found that the clustering accuracy and normalized mutual information is bigger than ours for the underlined cases and certain classes, as demonstrated in Table 2. Overall, our method has better performance and outperforms the NMF and GNMF by incorporating the adaptive TV regularization into ATV-NMF, which can significantly improve the discriminability of the data.

Below, we'll use one specific example to investigate the *Approximation Reconstruction Error* (*ARR*) learned in two methods GNMF and ATV-NMF that achieve compared to the reconstruction of the original matrix V based on the W and H obtained via the iteration process under the same iteration times set to be 100 on the COIL20 for the different cluster numbers. From the reconstruction error results shown in Table 3, we can see that the approximation reconstruction error of the ATV-NMF is smaller than that of the GNMF, which suggests that the proposed method can achieve better approximation reconstruction. The approximation reconstruction error is defined as

$$ARR = ||V - WH||_F^2$$

where V is the original matrix (data) with the size of 1024×1440, W and H is obtained via the updating rules Eqs. (7) and (8).

In order to further test clustering performance, we also compared ATV-NMF with other related NMF methods such as Graph dual regularization NMF (DNMF) [33], NMF and GNMF on the other two datasets. There are two regularization parameters which are set the same value in DNMF [33] and their parameter setting of DNMF are set to be the same as that of GNMF, i.e., the all regularization parameters are set to be 100, and the iteration times is also set to be 100 for the different cluster numbers on the PIE and Yale. Meanwhile, we also use the 0-1 weighting scheme for constructing the k -nearest neighbor graph, where $k = 5$. The clustering results of the four clustering algorithms on the PIE and Yale data sets are shown in Table 4, in which the best results are bold faced. From Table 4, we can see that GNMF, DNMF and ATV-NMF consider the geometrical structure of the data by adding graph regularization to NMF, which have better clustering performance than the NMF. Moreover, the proposed method outperforms the other three methods according to *ACC* and *NMI*. This indicates that the proposed method can learn better parts-based for data representation.

3.4. Sparseness study

NMF is a special parts-based representation learning method by non-negative constraints, and NMF only allows additive, not subtractive, combinations. In this subsection, we investigate the sparseness of the basis vectors learned by NMF, GNMF and ATV-NMF by using two specific examples. Figs. 1 and 2 show the basis vectors learned on the COIL 20 and ORL data sets, respectively. We plot these learned basis vectors as gray scale images. It is clear to see that the basis vectors learned by ATV-NMF are sparser than those learned by NMF and GNMF from the results as shown Figs. 1 and 2. In addition, some noises are removed and some feature or details are enhanced as shown in Fig. 2. This sparseness study reveals that the proposed ATV-NMF can learn better parts-based representation of data than both NMF and GNMF.

4. Conclusions

In this paper, we presented a sparse parts-based representation method for matrix factorization, called Adaptive Total-Variation Constrained based Non-negative Matrix Factorization on Manifold (ATV-NMF). The purpose is to reduce the influence of noise data during the process of data representation. By using the adaptive TV model, the proposed method can choose adaptively the anisotropic smoothing scheme to denoise or preserve adaptively the feature details of data based on the gradient information of data. We also exploit the graph regularization into ATV-NMF, which can discover intrinsic geometrical structure information of the data to conduct NMF over the data manifold. Experimental results on four widely used image data sets show that our method outperform state-of-the-art for data clustering.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant Nos. 61363049, 61702251 and 41201462,

the Scientific Research Fund of Jiangxi Provincial Education Department under Grant No. GJJ150751, Fujian Provincial Key Projects of Technology under Grant Nos. 2014H0034 and 2017H6015, the Natural Science Foundation of Jiangxi Province under Grant No. 20161BAB212033, the Natural Science Foundation of Fujian Province under Grant Nos. 2016J01310 and 2016J01309. The authors would like to thank Prof. D. Cai, in the College of Computer Science at Zhejiang University, China, for providing his code of GNMF.

References

[1] Y.X. Wang, Y.J. Zhang, Nonnegative matrix factorization: a comprehensive review, IEEE Trans. Knowl. Data. Eng. 25 (6) (2013) 1336–1353.
[2] F.G. Germain, G.J. Mysore, Stopping criteria for non-negative matrix factorization based supervised and semi-supervised source separation, IEEE Signal Process. Lett. 21 (10) (2014) 1284–1288.
[3] Y.Y. Liu, L.C. Jiao, F.H. Shang, An efficient matrix factorization based low-rank representation for subspace clustering, Pattern Recognit. 46 (1) (2013) 284–292.
[4] W.Y. Ren, G.H. Li, D. Tu, L. Jia, Nonnegative matrix factorization with regularizations, IEEE J. Emerg. Sel. Top.Circuits Syst. 4 (1) (2014) 153–164.
[5] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5000) (2000) 2319–2323.
[6] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[7] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. (2001) 585–591.
[8] D. Cai, X.F. He, J.W. Han, Isometric projection, in: Proceeding of the National Conference on Artificial Intelligence, 2007, pp. 528–533.
[9] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature. 401 (6755) (1999) 788–791.
[10] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Adv. Neural Inf. Process. Syst. (2001) 556–562.
[11] W. Xu, X. Liu, Y.H. Gong, Document clustering based on non-negative matrix factorization, in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 267–273.
[12] F. Shahnaz, M.W. Berry, V. Pauca, R.J. Plemmons, Document clustering using nonnegative matrix factorization, Inform. Process. Manag. 42 (2) (2006) 373–386.
[13] S.Z. Li, X.W. Hou, H.J. Zhang, Q.S. Cheng, Learning spatially localized, parts-based representation, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2001, pp. 207–212.
[14] S. Zafeiriou, A. Tefas, I. Buciu, I. Pitas, Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification, IEEE Trans. Neural Netw. 17 (3) (2006) 683–695.
[15] T.P. Zhang, B. Fang, Y.Y. Tang, G.H. He, J. Wen, Topology preserving non-negativematrix factorization for face recognition, IEEE Trans. Image Process. 17 (4) (2008) 574–584.
[16] R. Sandler, M. Lindenbaum, Nonnegative matrix factorization with earth mover's distance metric for image analysis, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1590–1602.
[17] N.Y. Guan, D.C. Tao, Z.G. Luo, B. Yuan, Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent, IEEE Trans. Image Process. 20 (7) (2011) 2030–2048.
[18] R. Peharz, F. Pernkopf, Sparse nonnegative matrix factorization with l^0-constraints, Neurocomputing. 80 (1) (2012) 38–46.
[19] D. Cai, X.F. He, J.W. Han, T.S. Huang, Graph regularized nonnegative matrix factorization for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1548–1560.
[20] L. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Phys. D. 60 (1–4) (1992) 259–268.
[21] T.P. Zhang, B. Fang, W.N. Liu, Y.Y. Tang, G.H. He, J. Wen, Total variation norm-based nonnegative matrix factorization for identifying discriminant representation of image patterns, Neurocomputing. 71 (10–12) (2008) 1824–1831.
[22] H.Q. Yin, H.W. Liu, Nonnegative matrix factorization with bounded total variational regularization for face recognition, Pattern Recognit. Lett. 31 (16) (2010) 2468–2473.
[23] H. Gao, H.K. Zhao, Multilevel bioluminescence tomography based on radiative transfer equation part 2: total variation and l1 data fidelity, Opt. Express. 18 (3) (2010) 2894–2912.
[24] J.C. Feng, C.H. Qin, K.B. Jia, S.P. Zhu, K. Liu, D. Han, X. Yang, Q.S. Gao, J. Tian, Total variation regularization for bioluminescence tomography with the split bregman method, Appl. Optics. 51 (19) (2012) 4501–4512.
[25] S. Levine, J. Stanich, Y.M. Chen, Image restoration via nonstandard diffusion, Technical Report 4(1), 2004.
[26] F.R.K. Chung, Spectral Graph Theory, Providence RI: American Mathematical Society, 1997.
[27] P. Sajda, S.Y. Du, L. Parra, Recovery of constituent spectra using non-negative matrix factorization, Proc. SPIE (2003) 321–331.
[28] M.Y. Hong, M. Razaviyayn, Z.Q. Luo, J.S. Pang, A unified algorithmic framework for block-structured optimization involving big data: with applications in machine learning and signal processing, IEEE Signal Process. Mag. 33 (1) (2016) 57–77.
[29] D. Cai, X.F. He, J.W. Han, Document clustering using locality preserving indexing, IEEE Trans. Knowl. Data. Eng. 17 (12) (2005) 1624–1637.
[30] H.F. Liu, Z.H. Wu, X.L. Li, D. Cai, T.S. Huang, Constrained nonnegative matrix factorization for image representation, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2012) 1299–1311.
[31] F.H. Shang, L.C. Jiao, J.R. Shi, F. Wang, M.G. Gong, Fast affinity propagation clustering: a multilevel approach, Pattern Recognit. 45 (1) (2012) 474–486.
[32] P. Paatero, U. Tapper, Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, Environmetrics 5 (2) (1994) 111–126.
[33] F.H. Shang, L.C. Jiao, F. wang, Graph dual regularization non-negative matrix factorization for co-clustering, Pattern Recognit. 45 (6) (2012) 2237–2250.