Robust Constrained Concept Factorization

Wei Yan and Bob Zhang

Abstract Accurately representing data is a fundamental problem in many pattern recognition and computational intelligence applications. In this chapter, a robust constrained concept factorization (RCCF) method is proposed. RCCF allows the extraction of important information, while simultaneously utilizing prior information when it is available, and is noise invariant. To guarantee data samples share the identical cluster and obtain similar representation in the new laten space, the proposed method uses a constraint matrix that is embodied into the rudimentary concept factorization model. The $L_{2,1}$-norm is used for both the reconstruction function and the regularization, which allows the proposed model to be insensitive to outliers. Furthermore, the $L_{2,1}$-norm regularization assists in the selection of useful information with joint sparsity. An elegant and efficient iterative updating scheme is also introduced with convergence and correctness analysis. Experimental results on commonly used databases in pattern recognition and computational intelligence demonstrate the effectiveness of RCCF.

Keywords Concept factorization · Dimensionality reduction · Clustering

1 Introduction

Obtaining a suitable representation is a fundamental problem for many research areas. For example: machine learning [1], data mining [2, 3], signal processing [4–6], and in particular pattern recognition [7–9], and computational intelligence [10, 11]. Optimal data representation can boost the performance of a learning task by revealing the underlying structure with-in a high-dimensional space. Recently,

W. Yan · B. Zhang (✉)
Department of Computer and Information Science,
University of Macau, Macau, China
e-mail: bobzhang@umac.mo

W. Yan
e-mail: yb67410@umac.mo

W. Pedrycz and S.-M. Chen (eds.), *Computational Intelligence for Pattern Recognition*, Studies in Computational Intelligence 777,
https://doi.org/10.1007/978-3-319-89629-8_7

matrix factorization based methods, including Singular Value Decomposition (SVD) [12], Principal Component Analysis (PCA) [13], Vector Quantization (VQ) [14], Nonnegative Matrix Factorization (NMF) [15–18] and Concept Factorization (CF) [19–22], have been receiving considerable attention as useful techniques for learning meaningful representation.

Generally, the main goal of these methods is to represent the given matrix as a product of two or more matrices. Among them, NMF is superior to PCA, SVD, and VQ for providing meaningful factorization results. Moreover, NMF yields parts-based and sparse representation because the nonnegative constraints allow only additive combinations. Regards to the parts-based representation, there are physiological evidences [23, 24]. However, NMF performs only in original data space. It is an issue about how to successfully apply NMF in reproducing kernel Hilbert space (RKHS), e.g., the transformed data space [20]. Recently, Concept Factorization (CF), an important variation of NMF, which uses linear combination of input data to represent the bases, has been effectively employed in processing real data, such as text and image, due to the fact that CF inherits all the strengths from NMF. Besides this, CF can be employed effectively in the transformed space. When using the CF method in data clustering, each sample is reconstructed as a linear combination of the cluster centers, and each cluster center is expressed as a linear combination of the samples. Here, the task of data clustering can be regarded as finding two sets of coefficients. To further improve the clustering performance of CF, Locally Consistent Concept Factorization (LCCF) [20] preserves the intrinsic structure information of the data set by incorporating the manifold structure into the CF model.

Despite its impressive performances, there are three major drawbacks for basic CF: (1) It is prone to outliers since a few outliers or noisy features with large errors will play a dominative role in the least square error function. Indeed, in many applications, data are additionally corrupted and thus data always contains noisy features or outliers. A potential robust version of CF is needed to deal with these issues. (2) Basic CF does not always result in sparse representation since there is no constraints to manage the sparseness explicitly. That means the representation in the low-dimensional space may still contain redundant and useless information. Generally, adding sparsity regularization is one practical method to control the degree of sparseness in factorization results, but it was designed only for NMF [25, 26]. (3) CF obtains data representation in an unsupervised way. It may not effectively distinguish the constrained data from the unconstrained data. Especially when the prior information is collected and CF does not completely use this information. To bridge this gap, a constrained algorithm, named constrained concept factorization (CCF) [27] is proposed utilizing prior information as a constraint matrix.

However, there is no such a framework that addresses all these drawbacks simultaneously. In this chapter, we propose a robust constrained concept factorization (RCCF) method, which not only makes good use of the available label information, but also addresses noise and learns meaningful information at the same time. Specifically, we utilized the mixed norm $L_{2,1}$-norm instead of the F-norm that is used in basic CF as our loss function, thus improving the robustness of the model such that this new model can effectively deal with outliers and can be employed in

pattern recognition and computational intelligence. Then, a constraint matrix, which contains label information, is embedded into the original CF model to guarantee data belong to the same cluster obtain the identical representation in the new representation space. Hence, the learned representation achieves better distinguishing abilities. In addition, the $L_{2,1}$-norm regularization is added in RCCF to obtain sparse results that help select the most relevant information. For optimizing the new model, we derive efficient updating rules that are iterative. At the same time, we analyse the correctness and convergence of the updating rules. Experimental results on three different data sets have shown the effectiveness of RCCF.

The remainder of this chapter is organized as follows. Section 2 proposes the RCCF framework, followed with its updating rules. Section 3 presents the experimental results on three data sets. Finally, we summarize our work.

2 Robust Constrained Concept Factorization (RCCF)

2.1 NMF and CF

Given a matrix $\mathbf{V} = [v_1, v_2, \ldots, v_N] \in {\mathbb{R}_+}^{M \times N}$, N denotes the number of data points and M is the length of the vector. For a dimensionality number K, NMF tries to seek two nonnegative data matrices $\mathbf{W} \in {\mathbb{R}_+}^{M \times K}$ and $\mathbf{H} \in {\mathbb{R}_+}^{K \times N}$ whose product gives an approximation to the input data matrix. The objective function of NMF is:

$$\mathcal{O} = \|\mathbf{V} - \mathbf{WH}\|_F^2. \tag{1}$$

Since (1) is a nonconvex minimization problem, it is unrealistic to get the optimal solution. However it is convex in W only or H only. Based on this analysis, Lee [15] proposed the following updating rules:

$$\mathbf{W}_{ik} \leftarrow \mathbf{W}_{ik} \frac{(\mathbf{VH^T})_{ik}}{(\mathbf{WHH^T})_{ik}} \tag{2}$$

$$\mathbf{H}_{kj} \leftarrow \mathbf{H}_{kj} \frac{(\mathbf{W^T V})_{kj}}{(\mathbf{W}^T\mathbf{WH})_{kj}}. \tag{3}$$

In regards to the above solutions, if $\mathbf{W}$ and $\mathbf{H}$ are the solution to (1), $\mathbf{WQ}$ and $\mathbf{HQ}^{-1}$ can also be a solution for any matrix $\mathbf{Q}$, which is positive and diagonal. To ensure it unique, we normalize the solution. In practice, this can be obtained by:

$$\mathbf{W}_{ik} \leftarrow \frac{\mathbf{W}_{ik}}{\sqrt{\sum_i W_{ik}^2}} \tag{4}$$

$$\mathbf{H}_{kj} \leftarrow \mathbf{H}_{kj} \sqrt{\sum_i W_{ik}^2} \tag{5}$$

In [15], Lee gives the proof that the aforementioned updating rules could obtain a local solution of (1).

Xu and Gong [19] modeled the document clustering problem by using two data representations. In the first one, the intrinsic semantics (e.g. clusters) can be represented by related document samples that belong to similar semantics. That is, the entire samples can be used to construct the cluster, and this combination can is linearly. Let V_i denotes the term-frequency vector of sample i, where $i = 1, \ldots, n$, m is the dimensionality and R_c is the centroid of cluster c, where $c = 1, \ldots, k$. The first representation can be defined as:

$$R_c = \sum_i w_{ic} V_i \tag{6}$$

where w_{ic} is nonnegative coefficient that represents the coefficient of data point i relating to cluster c. In the second one, all the clusters can be used to reconstruct the samples. The corresponding weight denotes the coefficient of overlap between the related sample and the cluster. The above two representations can be formulated as:

$$V_i = \sum_c h_{ic} R_c \tag{7}$$

where h_{ic} is the coefficient value that gives the coefficient of overlap between the related sample V_i and the concept cluster R_c. We construct the document matrix $\mathbf{V} = [V_1, V_2, \ldots, V_n] \in {\mathbf{R}_+}^{m \times n}$ with the feature vector of sample i as the ith column. From (6) and (7) we have

$$\mathbf{V} \approx \mathbf{VWH} \tag{8}$$

where $\mathbf{W} = [w_{jk}] \in {\mathbf{R}_+}^{n \times k}$ and $\mathbf{H} = [h_{jk}] \in {\mathbf{R}_+}^{k \times n}$. From Eq. (8), we observe that it can be considered as a factorization process of input sample matrix $\mathbf{X}$ into $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{H}$. With the factorization results, we can find the cluster which are accomplished by constructed by $\mathbf{XW}$. The cluster coefficient of each sample is obtained by finding the $\mathbf{H}$. Thus, we term this process concept factorization (CF). As $k \ll m$ and $k \ll n$, concept factorization leads to low-dimensional representation of the input matrix. This means the object function is defined as:

$$\mathcal{O} = \|\mathbf{V} - \mathbf{VWH}\|_F^2 \tag{9}$$

where $\|\cdot\|_F^2$ is the Frobenius norm of a matrix. Using the formulation (9), the data clustering problem can be solved by finding $\mathbf{W}$ and $\mathbf{H}$ that minimizes the $\mathcal{O}$.

To minimize (9), the multiplicative updating rules are introduced as [19]:

$$w_{nk} \leftarrow w_{nk} \frac{(\mathbf{KH^T})_{nk}}{(\mathbf{KWHH}^T)_{nk}} \tag{10}$$

$$h_{kn} \leftarrow h_{kn} \frac{(\mathbf{W^T K})_{kn}}{(\mathbf{W}^T\mathbf{KWH})_{kn}}. \tag{11}$$

where $\mathbf{K} = \mathbf{V}^T\mathbf{V}$. It is natural to leverage kernel methods on CF. Therefore, CF can make use of kernel methods to improve its performance in real applications. More information can be found in [19].

Concept factorization is an effective tool of data clustering, which is a fundamental topic in data mining. Data mining is about extracting interesting information from raw data. Data clustering aims to efficiently separate a given data set into clusters, which is a kind of key information. Among various clustering methods, concept factorization is widely used since it can provide meaningful clustering results. From the definition of CF, the cluster is constructed by using the input samples. This combination is linearly. The cluster construction as well as the new data representation can be addressed by CF.

2.2 RCCF Model

According to recent semi-supervised algorithms [27–29] a few labeled samples could be used along with the unlabeled samples to improve learning accuracy of unlabeled data. Inspired by previous research CCF [27], we assume that the first l data samples are given label information with c clusters. Then we construct an constraint matrix $\mathbf{C}$, in which $c_{i,j} = 1$ if c_i belongs to the jth class; $C_{i,j} = 0$ otherwise. For example, given n data points, v_1, v_2 and v_3 come from class I, v_4 and v_5 belong to class II, v_6 is labeled with class III. Base on this illustration, the label indicator matrix C can be formulated as follows:

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{12}$$

Based on the indicator matrix C, a label constraint matrix $\mathbf{A}$ is defined as follows:

$$\mathbf{A} = \begin{pmatrix} \mathbf{C}_{l\times c} & 0 \\ 0 & \mathbf{I}_{N-l} \end{pmatrix}, \tag{13}$$

where $\mathbf{I}_{N-l}$ is an identity matrix. The obtained $\mathbf{H}$ of input data points in the new representation space is formulated as $\mathbf{H} = \mathbf{ZA}^T$. That means if samples v_i and v_j come from the identical category, the ith row and the jth row of $\mathbf{A}$ should be identical, that is $h_i = h_j$, which makes sure that document point with the identical class could obtain the identical low-dimensional representation. Thus, (9) could be reformulated:

$$\min_{\mathbf{W}\geq 0, \mathbf{Z}\geq 0} \|\mathbf{V} - \mathbf{VWZA}^T\|_F^2 \tag{14}$$

However, F-norm is sensitive to outliers and could be unstable because the error for each sample in the objective function is expressed as squared. The objective function could be dominated by the large errors. To address this drawback, we utilize the mixed norm $L_{2,1}$-norm on the loss function to effectively remove outliers and noise. According to [30], the definition of $L_{2,1}$-norm is:

$$\|\mathbf{U}\|_{2,1} = \sum_{i=1}^{M} \sqrt{\sum_{j=1}^{N} \mathbf{U}_{ji}^2} = \sum_{i=1}^{M} \|u_i\|_2, \tag{15}$$

where u_i is the ith row of $\mathbf{U}$. We rewrite the error function:

$$\begin{aligned} \|\mathbf{V} - \mathbf{VWH}\|_{2,1} &= \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{M} (\mathbf{V} - \mathbf{VWH})_{ji}^2} \\ &= \sum_{i=1}^{N} \|v_i - \mathbf{VW}h_i\|. \end{aligned} \tag{16}$$

We can observe that the error for each sample in the new objective function (16) is not of the form x^2, so the large errors because of outliers do not impact the function in (16) dramatically. By employing the $L_{2,1}$-norm as measurement of the reconstruction error, the objective function can be reformulated as

$$\min_{\mathbf{W}\geq 0, \mathbf{Z}\geq 0} \|\mathbf{V} - \mathbf{VWZA}^T\|_{2,1}. \tag{17}$$

Furthermore, the real data usually contains meaningless features, i.e., not all the features are useful. Although basic CF can lead to sparse results that help extract meaningful features, it does not always result in such representation. Regarding this, we use the $L_{2,1}$-norm regularization term to control row sparsity on the new representation of data to extract informative features. Generally, the group sparsity imposing on the representation matrix $\mathbf{H}^T$ can be represented as follows,

$$\min_{\mathbf{H}\geq 0} \|\mathbf{H}^T\|_{2,1}. \tag{18}$$

As we make $\mathbf{H} = \mathbf{Z}\mathbf{A}^T$, our task is to get the minimum of matrix $\mathbf{A}\mathbf{Z}^T$. Since the constrained matrix $\mathbf{A}$ is given, the task in turn is to find matrix $\mathbf{Z}^T$.

By embedding the constrained matrix into basic CF, and imposing the $L_{2,1}$-norm on both the regularization and reconstruction function, a new model can be obtained as follows,

$$\min_{\mathbf{W}\geq 0,\mathbf{Z}\geq 0} \|\mathbf{V} - \mathbf{V}\mathbf{W}\mathbf{Z}\mathbf{A}^T\|_{2,1} + \alpha\|\mathbf{Z}^T\|_{2,1}, \tag{19}$$

where $\mathbf{V} \in {\mathbb{R}_+}^{M\times N}$, $\mathbf{W} \in {\mathbb{R}_+}^{N\times K}$, $\mathbf{Z} \in {\mathbb{R}_+}^{K\times(N-l+c)}$ and $\mathbf{A} \in {\mathbb{R}_+}^{N\times(N-l+c)}$. In this function, there is only one parameter, e.g., the parameter α. This item plays the role on controlling the sparse regularization.

2.3 Solutions of the RCCF Model

The solutions for the RCCF model via an iterative strategy is given as follows,

$$\mathbf{Z}_{ki} \leftarrow \mathbf{Z}_{ki}\frac{(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{D}_1\mathbf{A})_{ki}}{(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{Z}\mathbf{A}^T\mathbf{D}_1\mathbf{A} + \alpha\mathbf{D}_2\mathbf{Z})_{ki}}, \tag{20}$$

$$\mathbf{W}_{nk} \leftarrow \mathbf{W}_{nk}\frac{(\mathbf{V}^T\mathbf{V}\mathbf{D}_1\mathbf{A}\mathbf{Z}^T)_{nk}}{(\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{Z}\mathbf{A}^T\mathbf{D}_1\mathbf{A}\mathbf{Z}^T)_{nk}}, \tag{21}$$

The entries of $\mathbf{D}_1$ and $\mathbf{D}_2$ are defined as:

$$(\mathbf{D}_1)_{ii} = \frac{1}{\|\mathbf{V}_i - \mathbf{V}\mathbf{W}(\mathbf{Z}\mathbf{A}^T)_i\|}, i = 1, 2\ldots, N. \tag{22}$$

$$(\mathbf{D}_2)_{ii} = \frac{1}{\|(\mathbf{Z}^T)_i\|}, i = 1, 2\ldots, K. \tag{23}$$

2.4 RCCF Model Convergence

In this subsection, we give the analysis of the convergence of proposed updating rules with following two Theorems.

Theorem 1 *Obtaining* $\mathbf{Z}$ *utilizing the rule of* (20) *with* $\mathbf{W}$ *being fixed, the objective function of* (19) *is non-increasing,*

$$\begin{aligned}&\|\mathbf{V} - \mathbf{V}\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T\|_{2,1} + \alpha\|(\mathbf{Z}^{t+1})^T\|_{2,1}\\ &-\|\mathbf{V} - \mathbf{V}\mathbf{W}\mathbf{Z}^t\mathbf{A}^T\|_{2,1} - \alpha\|(\mathbf{Z}^t)^T\|_{2,1} \leq 0,\end{aligned} \tag{24}$$

where t *is the number of iteration.*

Theorem 2 *Obtaining* $\mathbf{W}$ *utilizing the solution proposed in* (21) *when* $\mathbf{Z}$ *is fixed, the objective function of* (19) *is nonincreasing,*

$$\|\mathbf{V}-\mathbf{V}\mathbf{W}^{t+1}\mathbf{Z}\mathbf{A}^T\|_{2,1}-\|\mathbf{V}-\mathbf{V}\mathbf{W}^t\mathbf{Z}\mathbf{A}^T\|_{2,1}\leq 0, \tag{25}$$

where t *represents the number of iteration.*

We use the following Lemma 1 to prove Theorem 1.

Lemma 1 *With the solution in* (20)*, we have the following inequation:*

$$\begin{aligned}&Tr((\mathbf{V}-\mathbf{V}\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{V}\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T)^T)\\&+\alpha Tr((\mathbf{Z}^{t+1})^T\mathbf{D}_2\mathbf{Z}^{t+1})\\&\leq Tr((\mathbf{V}-\mathbf{V}\mathbf{W}\mathbf{Z}^t\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{V}\mathbf{W}\mathbf{Z}^t\mathbf{A}^T)^T)\\&+\alpha Tr((\mathbf{Z}^t)^T\mathbf{D}_2\mathbf{Z}^t).\end{aligned} \tag{26}$$

Proof Following [31], we introduce an auxiliary function approach to help prove Lemma 1. Firstly, we have

$$\begin{aligned}J(\mathbf{Z})=&Tr((\mathbf{V}-\mathbf{V}\mathbf{W}\mathbf{Z}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{V}\mathbf{W}\mathbf{Z}\mathbf{A}^T)^T)\\&+\alpha Tr(\mathbf{Z}^T\mathbf{D}_2\mathbf{Z}).\end{aligned} \tag{27}$$

Next we re-express (26) as

$$J(\mathbf{Z}^{t+1})\leq J(\mathbf{Z}^t). \tag{28}$$

Base on (27), the following equation can be obtained

$$\begin{aligned}J(\mathbf{Z})&=Tr(\mathbf{V}\mathbf{D}_1\mathbf{V}^T-2\mathbf{V}\mathbf{D}_1\mathbf{A}\mathbf{Z}^T\mathbf{W}^T\mathbf{V}^T)\\&+Tr(\mathbf{V}\mathbf{W}\mathbf{Z}\mathbf{A}^T\mathbf{D}_1\mathbf{A}\mathbf{Z}^T\mathbf{W}^T\mathbf{V}^T)+\alpha Tr(\mathbf{Z}^T\mathbf{D}_2\mathbf{Z})\\&\leq Tr(\mathbf{V}\mathbf{D}_1\mathbf{V}^T-2\mathbf{V}\mathbf{D}_1\mathbf{A}\mathbf{Z}^T\mathbf{W}^T\mathbf{V}^T)\\&+\sum_{k=1}^{K}\sum_{i=1}^{(N-l+c)}\frac{(\mathbf{S}_1\mathbf{H}'\mathbf{B}_1)_{ki}(\mathbf{H}^2)_{ki}}{\mathbf{H}'_{ki}}\\&+\sum_{k=1}^{K}\sum_{i=1}^{(N-l+c)}\frac{(\mathbf{S}_2\mathbf{H}'\mathbf{B}_2)_{ki}(\mathbf{H}^2)_{ki}}{\mathbf{H}'_{ki}}\\&=Tr(\mathbf{V}\mathbf{D}_1\mathbf{V}^T-2\mathbf{V}\mathbf{D}_1\mathbf{A}\mathbf{Z}^T\mathbf{W}^T\mathbf{V}^T)\\&+\sum_{k=1}^{K}\sum_{i=1}^{(N-l+c)}\frac{(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{Z}'\mathbf{A}^T\mathbf{D}_1\mathbf{A}+\alpha\mathbf{D}_2\mathbf{Z}')_{ki}(\mathbf{Z}^2)_{ki}}{\mathbf{Z}'_{ki}}\\&=F(\mathbf{Z},\mathbf{Z}'),\end{aligned} \tag{29}$$

where $\mathbf{S}_1 = \mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}$, $\mathbf{B}_1 = \mathbf{A}^T\mathbf{D}_1\mathbf{A}$, $\mathbf{H} = \mathbf{Z}$, $\mathbf{H}' = \mathbf{Z}'$, $\mathbf{B}_2 = \mathbf{I}$, and $\mathbf{S}_2 = \alpha\mathbf{D}_2$. The equality holds in case of $\mathbf{Z} = \mathbf{Z}'$. The auxiliary function of $J(\mathbf{Z})$ is $F(\mathbf{Z}, \mathbf{Z}')$.

Let

$$\mathbf{Z}^{t+1} = \arg\min_{\mathbf{Z}} F(\mathbf{Z}, \mathbf{Z}^t), \tag{30}$$

we can get

$$J(\mathbf{Z}^{t+1}) = F(\mathbf{Z}^{t+1}, \mathbf{Z}^{t+1}) \leq F(\mathbf{Z}^{t+1}, \mathbf{Z}^t) \leq J(\mathbf{Z}^t), \tag{31}$$

From (31), we can have the provement that $J(\mathbf{Z}^t)$ is non-increasing.

Let $f(\mathbf{Z}) = F(\mathbf{Z}, \mathbf{Z}')$, the gradient of $f(\mathbf{Z})$ is

$$\begin{aligned}\frac{\partial f(\mathbf{Z})}{\partial \mathbf{Z}_{ki}} &= -2(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{D}_1\mathbf{A})_{ki} \\ &\quad + 2\frac{(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{Z}'\mathbf{A}^T\mathbf{D}_1\mathbf{A} + \alpha\mathbf{D}_2\mathbf{Z}')_{ki}(\mathbf{Z})_{ki}}{\mathbf{Z}'_{ki}}.\end{aligned} \tag{32}$$

The Hessian matrix of $f(\mathbf{Z})$ is

$$\frac{\partial^2 f(\mathbf{Z})}{(\partial \mathbf{Z}_{ki})(\partial \mathbf{Z}_{lj})} = 2\frac{(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{Z}'\mathbf{A}^T\mathbf{D}_1\mathbf{A} + \alpha\mathbf{D}_2\mathbf{Z}')_{ki}}{\mathbf{Z}'_{ki}}\delta_{ij}\delta_{kl}. \tag{33}$$

Since $f(\mathbf{Z})$ is convex and the second-order derivatives is semi-positive definite, we can obtain the solution for $f(\mathbf{Z})$. By letting (32) be zero, we can obtain the update rule of $\mathbf{Z}$ as:

$$\mathbf{Z}_{ki} \leftarrow \mathbf{Z}'_{ki}\frac{(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{D}_1\mathbf{A})_{ki}}{(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{Z}'\mathbf{A}^T\mathbf{D}_1\mathbf{A} + \alpha\mathbf{D}_2\mathbf{Z}')_{ki}}, \tag{34}$$

Let $\mathbf{Z}^t \leftarrow \mathbf{Z}'$, $\mathbf{Z}^{t+1} \leftarrow \mathbf{Z}$, (34), we can obtain the iterative solution of (20). When we use this strategy to update $\mathbf{Z}$, the objective function of (27) is non-increasing.

Until now, Lemma 1 is proved.

Lemma 2 *In order to finish the proof of this, we refer to the matrix inequality in [32]. If matrices* $\mathbf{S} \geq \mathbf{0}$, $\mathbf{B} \geq \mathbf{0}$, $\mathbf{H} \geq \mathbf{0}$, *the sizes are suitable and* $\mathbf{B} = \mathbf{B}^T$, $\mathbf{S} = \mathbf{S}^T$, *we obtain the matrix inequality:*

$$Tr(\mathbf{H}^T\mathbf{S}\mathbf{H}\mathbf{B}) \leq \sum_{ik}(\mathbf{S}\mathbf{H}'\mathbf{B})\frac{\mathbf{H}_{ik}^2}{\mathbf{H}'_{ik}} \tag{35}$$

Lemma 3 *According to the solution in* (20), *the following in-equation holds*

$$
\begin{aligned}
&\|\mathbf{V}-\mathbf{VWZ}^{t+1}\mathbf{A}^T\|_{2,1}+\alpha\|(\mathbf{Z}^{t+1})^T\|_{2,1}\\
&-\|\mathbf{V}-\mathbf{VWZ}^{t}\mathbf{A}^T\|_{2,1}-\alpha\|(\mathbf{Z}^{t})^T\|_{2,1}\\
&\leq\frac{1}{2}[Tr((\mathbf{V}-\mathbf{VWZ}^{t+1}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{VWZ}^{t+1}\mathbf{A}^T)^T)\\
&+\alpha Tr((\mathbf{Z}^{t+1})^T\mathbf{D}_2\mathbf{Z}^{t+1})\\
&-Tr((\mathbf{V}-\mathbf{VWZ}^{t}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{VWZ}^{t}\mathbf{A}^T)^T)\\
&-\alpha Tr((\mathbf{Z}^{t})^T\mathbf{D}_2\mathbf{Z}^{t})].
\end{aligned}
\tag{36}
$$

Proof Lemma 3 can be proved with the same method of [33]. Then, we can derive (36).

With the characteristic of $\mathbf{D}_1$ and $\mathbf{D}_2$, we have

$$
\begin{aligned}
&Tr((\mathbf{V}-\mathbf{VWZ}^{t+1}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{VWZ}^{t+1}\mathbf{A}^T)^T)\\
&+\alpha Tr((\mathbf{Z}^{t+1})^T\mathbf{D}_2\mathbf{Z}^{t+1})\\
&=\sum_{i=1}^{N}\|\mathbf{V}_i-\mathbf{VW}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii}\\
&+\alpha\sum_{i=1}^{K}\|(\mathbf{Z}_i^{t+1})^T\|^2(\mathbf{D}_2)_{ii},
\end{aligned}
\tag{37}
$$

$$
\begin{aligned}
&Tr((\mathbf{V}-\mathbf{VWZ}^{t}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{VWZ}^{t}\mathbf{A}^T)^T)\\
&+\alpha Tr((\mathbf{Z}^{t})^T\mathbf{D}_2\mathbf{Z}^{t})\\
&=\sum_{i=1}^{N}\|\mathbf{V}_i-\mathbf{VW}(\mathbf{Z}^{t}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii}\\
&+\alpha\sum_{i=1}^{K}\|(\mathbf{Z}_i^{t})^T\|^2(\mathbf{D}_2)_{ii}.
\end{aligned}
\tag{38}
$$

The right-hand side (RHS) of (36) becomes

$$
\begin{aligned}
RHS=&\frac{1}{2}\sum_{i=1}^{N}(\|\mathbf{V}_i-\mathbf{VW}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii}\\
&-\|\mathbf{V}_i-\mathbf{VW}(\mathbf{Z}^{t}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii})\\
&+\frac{1}{2}\alpha\sum_{i=1}^{K}(\|(\mathbf{Z}^{t+1})_i^T\|^2(\mathbf{D}_2)_{ii}-\|(\mathbf{Z}^{t})_i^T\|^2(\mathbf{D}_2)_{ii}).
\end{aligned}
\tag{39}
$$

Combining (22) and (23),

$$\begin{aligned} RHS = & \frac{1}{2}\sum_{i=1}^{N}\left(\|\mathbf{V}_i - \mathbf{V}\mathbf{W}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii} - \frac{1}{(\mathbf{D}_1)_{ii}}\right) \\ & + \frac{1}{2}\alpha\sum_{i=1}^{K}(\|(\mathbf{Z}^{t+1})_i^T\|^2(\mathbf{D}_2)_{ii} - \frac{1}{(\mathbf{D}_2)_{ii}}). \end{aligned} \tag{40}$$

The left-hand side (LHS) of (36) becomes

$$\begin{aligned} LHS = & \|\mathbf{V} - \mathbf{V}\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T\|_{2,1} + \alpha\|(\mathbf{Z}^{t+1})^T\|_{2,1} \\ & - \|\mathbf{V} - \mathbf{V}\mathbf{W}\mathbf{Z}^{t}\mathbf{A}^T\|_{2,1} - \alpha\|(\mathbf{Z}^{t})^T\|_{2,1} \\ = & \sum_{i=1}^{N}(\|\mathbf{V}_i - \mathbf{V}\mathbf{W}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\| - \frac{1}{(\mathbf{D}_1)_{ii}}) \\ & + \alpha\sum_{i=1}^{K}(\|(\mathbf{Z}^{t+1})_i^T\| - \frac{1}{(\mathbf{D}_2)_{ii}}). \end{aligned} \tag{41}$$

Therefore, we have

$$\begin{aligned} & LHS - RHS \\ = & \sum_{i=1}^{N}\frac{-(\mathbf{D}_1)_{ii}}{2}(\|\mathbf{V}_i - \mathbf{V}\mathbf{W}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\| - \frac{1}{(\mathbf{D}_1)_{ii}})^2 \\ & + \sum_{i=1}^{K}\frac{-(\mathbf{D}_2)_{ii}}{2}(\|(\mathbf{Z}^{t+1})_i^T\| - \frac{1}{(\mathbf{D}_2)_{ii}})^2 \leq 0. \end{aligned} \tag{42}$$

Until now, the proof of Lemma 3 is accomplished.

With the usage of Lemmas 1–3, the proof of Theorem 1 can be obtained. It means the objective function of (19) is non-increasing under the solution in (20).

We can take the same strategy to prove the Theorem 2, we do not provide details here.

2.5 *Correctness of the RCCF Analysis*

In the following, we will prove that the proposed algorithms is guaranteed to converge to the Karush-Kuhn-Tucker (KKT) points.

Theorem 3 *Using the updating rule in* (20)*, the obtained solution of* $\mathbf{Z}$ *satisfies the Karush-Kuhn-Tucker condition.*

Proof. The Karush-Kuhn-Tucker condition for $\mathbf{Z}$ with the constrains $(\mathbf{Z})_{ki} \geq 0$, $k = 1, 2, \ldots, K$; $i = 1, 2, \ldots, (N - l + c)$, is

$$\frac{\partial J(\mathbf{Z})}{\partial (\mathbf{Z})_{ki}} (\mathbf{Z})_{ki} = 0, \forall k, i. \tag{43}$$

The derivative is

$$\frac{\partial J(\mathbf{Z})}{\partial (\mathbf{Z})_{ki}} = -2((\mathbf{W}^T\mathbf{V}^T\mathbf{V}(1 - \mathbf{W}\mathbf{Z}\mathbf{A}^T)\mathbf{D}_1\mathbf{A})_{ki} + \alpha(\mathbf{D}_2\mathbf{Z})_{ki}). \tag{44}$$

Then, the Karush-Kuhn-Tucker condition for $\mathbf{Z}$ is

$$\begin{aligned} &[-(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{D}_1\mathbf{A})_{ki} + (\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{Z}\mathbf{A}^T\mathbf{D}_1\mathbf{A})_{ki} \\ &+ \alpha(\mathbf{D}_2\mathbf{Z})_{ki}](\mathbf{Z})_{ki} \\ &= 0, \forall k, i. \end{aligned} \tag{45}$$

If the $\mathbf{Z}$ converges under updating rule of (20), the obtained solution $\mathbf{Z}^*$ satisfies

$$\mathbf{Z}^*_{ki} \leftarrow \mathbf{Z}^*_{ki} \frac{(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{D}_1\mathbf{A})_{ki}}{(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{Z}^*\mathbf{A}^T\mathbf{D}_1\mathbf{A} + \alpha\mathbf{D}_2\mathbf{Z}^*)_{ki}}, \tag{46}$$

which can be reformulated as

$$\begin{aligned} &[-(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{D}_1\mathbf{A})_{ki} + (\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{Z}^*\mathbf{A}^T\mathbf{D}_1\mathbf{A})_{ki} \\ &+ \alpha(\mathbf{D}_2\mathbf{Z}^*)_{ki}](\mathbf{Z}^*)_{ki} \\ &= 0, \forall k, i. \end{aligned} \tag{47}$$

We observe that (47) is the same as (45). This means the learned solution for $\mathbf{Z}^*$ satisfies the Karush-Kuhn-Tucker condition. Until now, we finish the proof.

Theorem 4 *With the solution* $\mathbf{W}$ *under the updating rule of* (21)*, the proposed algorithm converges to the Karush-Kuhn-Tucker points.*

In regards to proving Theorem 4, we can take the same strategy to finish it.

3 Experimentation

3.1 Description of the Data

We used three data sets in our experiments. There are two face data sets and one handwritten digits images database. The proposed RCCF method is evaluated on data clustering. Table 1 show details of the selected data sets.

Table 1 Details of the datasets

Datasets	Size	Dimentions	Classes
Yale	165	1024	15
ORL	400	1024	40
MNIST	1000	784	10

Fig. 1 Yale Faces database

Yale Database.[1] The Yale database contains 165 images in gray scale. These images belong to 15 different people. For each person, there are 11 facial images of size 32×32. Each picture is with a different facial expression or configuration. Similar to [34], all samples are normalized in orientation and scale to ensure that two eyes are aligned at the same position. Figure 1 shows some face images from this database.

ORL Database.[2] This database contains 400 gray scale face images of 40 individuals. For each individual, images are in different facial expressions or configurations. All these pictures are collected at different time, varying the lighting. We use the same way as the Yale data set to preprocess this data set. Figure 2 shows some examples from this database.

MNIST Database.[3] The MNIST database contains 10000 images of handwritten digits from 0 to 9 in gray scale. For each subject, there are 1000 images. We resize

[1] http://www.face-rec.org/databases/.

[2] http://www.face-rec.org/databases/.

[3] http://yann.lecun.com/exdb/mnist/.

Fig. 2 ORL Faces database

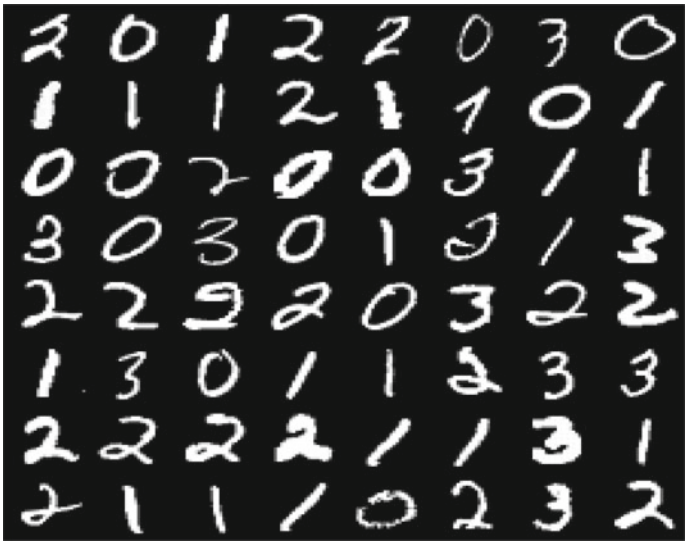

Fig. 3 MNIST database

each image to 16×16, thus the dimensionality of feature vector is 256. Figure 3 provides several example pictures from this data set.

3.2 Evaluation Metrics

Following [34, 35], the normalized mutual information metric (*NMI*) and the accuracy (*AC*) are employed to evaluate the clustering performance. Accuracy reflects the percentage of correctly predicted cluster number. Given a database with n samples, r_i is cluster information provided by the database, for each sample, let l_i be cluster information that we obtain by using different methods. The definition of AC is:

$$AC = \frac{\sum_{i=1}^{n} \delta(r_i, map(l_i))}{n} \tag{48}$$

where $\delta(x, y)$ be set 1 if $x = y$ and $\delta(x, y)$ be set 0 if $x \neq y$, and $map(l_i)$ denotes the permutation mapping function that maps each cluster label l_i to the corresponding label from the data set. We utilize the KM algorithm [36] to obtain the best map.

The normalized mutual information matrix is employed to measure the similarity of two clusters. Let C be the information of clusters achieved from the ground truth and C' obtained from the proposed algorithm. Its mutual information matrix is measured as:

$$MI(C, C') = \sum_{c_i \in C, c_j' \in C'} p(c_i, c_j') \cdot log \frac{p(c_i, c_j')}{p(c_i) \cdot p(c_j')}, \tag{49}$$

where $p(c_i, c_j')$ denotes the joint probability that the chosen sample comes from the cluster c_i and $p(c_j')$ simultaneously. $p(c_i)$ and $p(c_j')$ are the probabilities that a randomly chosen data comes from the clusters c_i and $p(c_j')$, respectively. In these experiments, we used the $NMI(C, C')$, which gets scores ranging from 0 to 1.

$$NMI(C, C') = \frac{MI(C, C')}{max(H(C), H(C'))}, \tag{50}$$

where $H(C)$ and $H(C')$ denote the entropies of C and C'. *NMI* equals to 1 when two selected samples are the same, and it is 0 when these two samples come from two different clusters.

3.3 Experimental Results

Testing was carried out on the proposed RCCF method in terms of clustering performance on three public datasets. At the same time we also make comparisons with related methods as follow:

1. Traditional KMeans clustering method (KMeans for short).
2. Concept-Factorization-based clustering (CF for short) [19].
3. Constrained concept factorization (CCF for short) [27].

The former two methods are unconstrained and the last one is a constrained algorithm.

Experiments are conducted with different cluster numbers K. K takes the value between 2 and 10. For each data set, K categories are selected randomly. Similar to [28], 30% of the data points are extracted to construct the training dataset and the rest for constructing the test dataset. Then, matrix factorization methods are used to obtain low-dimensional representations. The reduced dimensionality is set to be equal to the cluster number K. Once the new representation is obtained, we utilize KMeans by choosing cosine distance to the new representation for data clustering. KMeans is repeated 20 times with various initiations. The result with minimum cost function is recorded to measure accuracy and mutual information. There is an important tunable parameter for the proposed method. To make the experimental results persuasive, we perform grid search in the parameter space for our method and the best results are recorded. In particular, the search of α from 0.1 to 90 was carried out and we set $\alpha = 20, 10, 20$ for the YALE, ORL and MNIST respectively.

The Yale data set clustering results are shown in Table 2. Average AC and NMI versus K can be found in the last row. We can see that RCCF outperforms others most of the time, especially in terms of AC, while comparing to the second best results, i.e., average results in terms of AC and NMI for CCF, our algorithm RCCF achieves 3.1 percent and 5.3 percent improvements respectively.

Table 3 provides the clustering results for the data set named ORL. In this table it can be observed that RCCF achieves the best results for most cases. RCCF obtains the highest results 8/9 times in AC and the highest results 7/9 times in NMI. RCCF achieves a 3.7 percent improvement in AC and a 3.4 percent improvement in NMI on average, compared to the next best method (i.e. CCF)

Using MNIST, the details of the clustering results are given in Table 4. It can be observed from the table that the superiority of our method is obvious when K is

Table 2 Clustering methods' performance on Yale Database

K	Accuracy(%)				Normalized mutual information(%)			
	KMeans	CF	CCF	RCCF	KMeans	CF	CCF	RCCF
2	73.6	85.0	83.2	88.4	25.9	50.1	46.7	59.3
3	70.3	73.0	80.9	82.1	44.1	48.9	61.1	57.2
4	51.6	62.0	69.8	71.0	32.1	38.0	49.6	55.7
5	46.4	57.3	61.5	67.5	38.2	41.3	47.6	56.6
6	49.09	49.5	61.4	62.3	36.3	37.1	50.2	56.3
7	45.8	48.8	55.5	62.9	39.4	39.5	49.4	56.2
8	44.3	48.3	55.1	57.5	41.0	43.1	49.5	54.3
9	43.8	48.4	55.5	56.3	43.4	45.2	53.1	54.9
10	40.6	43.9	52.2	54.7	40.9	41.8	52.0	55.9
Avg.	52.0	57.4	63.9	67.0	37.9	42.8	51.0	56.3

Table 3 Clustering methods' performance on ORL Database

K	Accuracy(%)				Normalized mutual information(%)			
	KMeans	CF	CCF	RCCF	KMeans	CF	CCF	RCCF
2	95.0	87.1	90.7	93.6	76.1	56.7	64.3	75.6
3	60.0	70.4	80.9	83.8	43.2	49.5	62.9	68.7
4	57.5	61.4	65.4	73.9	47.4	50.0	55.8	63.0
5	63.0	57.7	61.7	63.7	52.7	50.3	55.6	55.9
6	56.7	60.5	61.2	63.1	55.2	59.7	59.4	59.1
7	54.5	57.4	62.9	65.3	58.0	58.4	63.8	64.0
8	53.6	58.5	63.8	67.3	59.3	61.2	65.8	67.4
9	59.9	62.2	63.2	68.3	67.0	65.2	65.9	69.6
10	58.3	55.4	62.0	65.9	65.3	63.3	68.1	68.7
Avg.	62.0	63.4	67.9	71.6	58.3	57.2	62.4	65.8

Table 4 Clustering methods' performance on MNIST Database

K	Accuracy(%)				Normalized mutual information(%)			
	KMeans	CF	CCF	RCCF	KMeans	CF	CCF	RCCF
2	90.1	90.1	94.6	98.6	59.9	60.1	73.3	90.6
3	82.2	81.2	82.8	87.0	56.5	54.7	60.0	65.2
4	74.4	70.8	80.9	84.0	53.7	50.9	63.1	69.8
5	67.9	63.2	80.0	86.8	51.5	46.0	62.5	73.7
6	65.9	68.2	77.7	74.6	53.9	52.2	63.1	62.0
7	63.2	61.6	75.1	74.3	53.5	50.2	63.2	61.3
8	57.1	61.2	67.3	68.4	50.9	50.0	57.6	58.0
9	53.9	57.0	67.6	67.6	51.1	48.5	58.5	57.7
10	54.1	53.8	68.9	58.6	50.9	45.8	59.2	50.1
Avg.	67.6	67.4	77.2	77.8	53.5	51.0	62.3	65.4

small. On average, RCCF and CCF have similar performance, however, RCCF still achieves the best results. When matched with the algorithm that performed second best (CCF), RCCF obtains a 3.1 percent improvement in *NMI*,

4 Conclusion

A robust constrained concept factorization (RCCF) method is proposed in this chapter. This new model learns discriminative results since it fully utilizes the labeled information with a constraint matrix. In addition, $L_{2,1}$-norm is applied on both the reconstruction function and the regularization. The $L_{2,1}$-norm based reconstruction

function improves the robustness of RCCF, and the $L_{2,1}$-norm regularization is used to select useful information. In order to solve the new model, we have derived an efficient iterative updating algorithm, along with proofs of convergence. Evaluating the proposed method on three data sets showed the superiority of the algorithm as a generalized method in pattern recognition and computational intelligence applications.

Acknowledgements This work is supported by the Science and Technology Development Fund (FDCT) of Macao SAR (124/2014/A3) and the National Natural Science Foundation of China (61602540).

References

1. Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
2. M. Tepper, G. Sapiro, Compressed nonnegative matrix factorization is fast and accurate. IEEE Trans. Signal Process. **64**(9), 2269–2283 (2016)
3. Z. Ma, A.E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, J. Guo, Variational bayesian matrix factorization for bounded support data. IEEE Trans. Pattern Anal. Mach. Intell. **37**(4), 876–889 (2015)
4. Z. Yang, Y. Xiang, Y. Rong, K. Xie, A convex geometry-based blind source separation method for separating nonnegative sources. IEEE Trans. Neural Netw. Learn. Syst. **26**(8), 1635–1644 (2015)
5. I. Domanov, L.D. Lathauwer, Generic uniqueness of a structured matrix factorization and applications in blind source separation. IEEE J. Sel. Top. Signal Process. **10**(4), 701–711 (2016)
6. X. Fu, W.K. Ma, K. Huang, N.D. Sidiropoulos, Blind separation of quasi-stationary sources: exploiting convex geometry in covariance domain. IEEE Trans. Signal Process. **63**(9), 2306–2320 (2015)
7. Y. Xu, B. Zhang, Z. Zhong, Multiple representations and sparse representation for image classification. Pattern Recognit. Lett. **68**, 9–14 (2015)
8. H. Zhang, J. Yang, J. Xie, J. Qian, B. Zhang, Weighted sparse coding regularized nonconvex matrix regression for robust face recognition. Inf. Sci. **394**, 1–17 (2017)
9. W. Jia, B. Zhang, J. Lu, Y. Zhu, Y. Zhao, W. Zuo, H. Ling, Palmprint recognition based on complete direction representation. IEEE Trans. Image Process. (2017)
10. Y. Xiao, Z. Zhu, Y. Zhao, Y. Wei, S. Wei, X. Li, Topographic nmf for data representation. IEEE Trans. Cybern. **44**(10), 1762–1771 (2014)
11. L. Luo, L. Chen, J. Yang, J. Qian, B. Zhang, Tree-structured nuclear norm approximation with applications to robust face recognition. IEEE Trans. Image Process. **25**(12), 5757–5767 (2016)
12. R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification* (Wiley, New York, 2012)
13. I. Jolliffe, *Principal Component Analysis* (Wiley Online Library, Hoboken, 2002)
14. R. Gray, Vector quantization. IEEE Assp Mag. **1**(2), 4–29 (1984)
15. D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
16. Z. Yang, Y. Zhang, W. Yan, Y. Xiang, S. Xie, A fast non-smooth nonnegative matrix factorization for learning sparse representation. IEEE Access **4**, 5161–5168 (2016)
17. X. Zhang, L. Zong, X. Liu, J. Luo, Constrained clustering with nonnegative matrix factorization. IEEE Trans. Neural Netw. Learn. Syst. **27**(7), 1514–1526 (2016)
18. Z. Yang, Y. Xiang, K. Xie, Y. Lai, Adaptive method for nonsmooth nonnegative matrix factorization. IEEE Trans. Neural Netw. Learn. Syst. **28**(4), 948–960 (2017)

19. W. Xu, Y. Gong, Document clustering by concept factorization, in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, 2004), pp. 202–209
20. D. Cai, X. He, J. Han, Locally consistent concept factorization for document clustering. IEEE Trans. Knowl. Data Eng. **23**(6), 902–913 (2011)
21. H. Liu, Z. Yang, J. Yang, Z. Wu, X. Li, Local coordinate concept factorization for image representation. IEEE Trans. Neural Netw. Learn. Syst. **25**(6), 1071–1082 (2014)
22. W. Yan, B. Zhang, S. Ma, Z. Yang, A novel regularized concept factorization for document clustering. Knowl.-Based Syst. (2017)
23. S.E. Palmer, Hierarchical structure in perceptual representation. Cogn. Psychol. **9**(4), 441–474 (1977)
24. N.K. Logothetis, D.L. Sheinberg, Visual object recognition. Annu. Rev. Neurosci. **19**(1), 577–621 (1996)
25. P.O. Hoyer, Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**, 1457–1469 (2004)
26. R. Peharz, F. Pernkopf, Sparse nonnegative matrix factorization with 0-constraints. Neurocomputing **80**, 38–46 (2012)
27. H. Liu, G. Yang, Z. Wu, D. Cai, Constrained concept factorization for image representation. IEEE Trans. Cybern. **44**(7), 1214–1224 (2014)
28. H. Liu, Z. Wu, D. Cai, T.S. Huang, Constrained nonnegative matrix factorization for image representation. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1299–1311 (2012)
29. D. Wang, X. Gao, X. Wang, Semi-supervised nonnegative matrix factorization via constraint propagation. IEEE Trans. Cybern. **46**(1), 233–244 (2016)
30. C. Ding, D. Zhou, X. He, H. Zha, R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization, in *Proceedings of the 23rd International Conference on Machine Learning* (ACM, 2006), pp. 281–288
31. D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in *Advances in Neural Information Processing Systems* (2001), pp. 556–562
32. C.H.Q. Ding, T. Li, M.I. Jordan, Convex and semi-nonnegative matrix factorizations. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 45–55 (2010)
33. D. Kong, C. Ding, H. Huang, Robust nonnegative matrix factorization using L21-norm, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (ACM, 2011), pp. 673–682
34. D. Cai, X. He, J. Han, T.S. Huang, Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1548–1560 (2011)
35. D. Cai, X. He, J. Han, Document clustering using locality preserving indexing. IEEE Trans. Knowl. Data Eng. **17**(12), 1624–1637 (2005)
36. L. Lovász, M.D. Plummer, *Matching Theory*, vol. 367 (American Mathematical Society, Providence, 2009)