Logdet Divergence Based Sparse Non-negative Matrix Factorization for Stable Representation

Qing Liao [1], Naiyang Guan [2], Qian Zhang [1]
[1] *Department of Computer Science and Engineering, The Hong Kong University of Science and Technology*
[2] *School of Computer Science, National University of Defense Technology*
qnature@ust.hk, ny_guan@nudt.edu.cn, qzhang@ust.hk

***Abstract*—Non-negative matrix factorization (NMF) decomposes any non-negative matrix into the product of two low dimensional non-negative matrices. Since NMF learns effective parts-based representation, it has been widely applied in computer vision and data mining. However, traditional NMF has the risk learning rank-deficient basis on high-dimensional dataset with few examples especially when some examples are heavily corrupted by outliers. In this paper, we propose a Logdet divergence based sparse NMF method (LDS-NMF) to deal with the rank-deficiency problem. In particular, LDS-NMF reduces the risk of rank deficiency by minimizing the Logdet divergence between the product of basis matrix with its transpose and the identity matrix, meanwhile penalizing the density of the coefficients. Since the objective function of LDS-NMF is non-convex, it is difficult to optimize. In this paper, we develop a multiplicative update rule to optimize LDS-NMF in the frame of block coordinate descent, and theoretically prove its convergence. Experimental results on popular datasets show that LDS-NMF can learn more stable representations than those learned by representative NMF methods.**

Keywords*—Non-negative matrix factorization, robust matrix decomposition, Logdet divergence.

I. Introduction

Many practical tasks confront the so-called "curse of dimensionality challenge" [1], and thus require effective methods to reduce the dimensionality of data at the preprocessing stage. Several dimension reduction methods aim to project high-dimensional data onto a lower-dimensional space, and boost subsequent processing. Among them, principle component analysis (PCA, [2]) and non-negative matrix factorization (NMF, [3] [4]) are two most popular unsupervised generative methods. PCA learns a group of orthogonal axes by maximizing variance of examples in the low-dimensional space, and NMF projects the high-dimensional data onto a positive low-dimensional orthant. Since many practical data, e.g., pixel values and video frames, are non-negative, it is natural to preserve such non-negativity property in the low-dimensional space.

In recent decades, NMF has been widely used in many data mining and computer vision tasks because its non-negativity constraint yields natural parts-based representation [5]. In particular, NMF [3][4] represents the examples as a combination of several non-negative bases, and allows only additive or non-subtractive coefficients. Since the non-negativity constraints over both basis and coefficients avoid cancellation of energies, NMF can learn parts-based representation. Since such parts-based representation is consistent with the psychological evidence in human brain [6][7][8], NMF has been widely applied in data representation. Figure 1 shows that our method can learn parts-based representation on the YaleB[9] dataset while PCA cannot. Recently, many NMF-based methods are developed by incorporating various regularization or constraints to enhance the representation capacity of the basis. For example, Hoyer [10] proposed NMF with sparseness constraint (NMFsc) to learn sparse representation of data. Cai *et al.* [11] proposed graph regularized NMF (GNMF) to enhance the image representation capacity by incorporating the data geometric structure regularization into NMF. Moreover, Chen *et al.* [12] proposed non-negative local coordinate factorization (NLCF) which adds a local coordinate into NMF to impose the basis to be close to the original data.

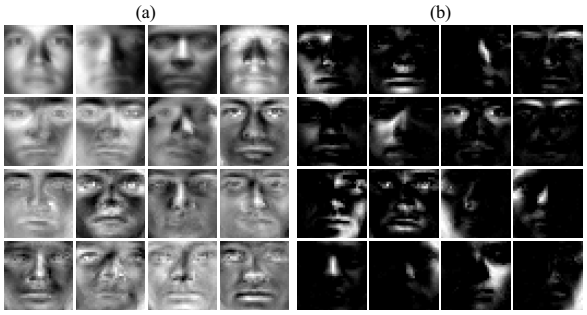

Fig. 1. The basis learned by (a) PCA and (b) our method on the YaleB dataset. PCA learns the holistic representation while our method induces the parts-based representation consistent with human intuition, i.e., the whole face consists of the components.

Although many NMF methods have succeeded in practical applications, they cannot perform well enough on some noisy datasets because their loss functions, i.e., Frobenius norm or Kullback-Leibler (KL) divergence, cannot handle outliers. To avoid this shortcoming, several works [13][14][15][16] incorporates additional knowledge into NMF. For example, Du *et al.* [13] proposed the correntropy induced metric based NMF method (CIM-NMF) which assumes that noise obeys non-Gaussian distribution. CIM-NMF significantly boosts NMF in terms of the subsequent classification or clustering performance. Yang *et al.* [14] integrated the Lasso regularization over data noises and Laplacian regularization over the coefficients into NMF, and proposed the robust NMF via joint sparse and graph regularization (RSGNMF) method. Kong *et al.* [15][16] proposed the $L_{2,1}$-NMF which utilizes $L_{2,1}$-norm to penalize the reconstruction loss meanwhile removing outliers from data. However, both traditional NMF and its variants cannot theoretically guarantee the rank of the learned basis to be equivalent

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.52

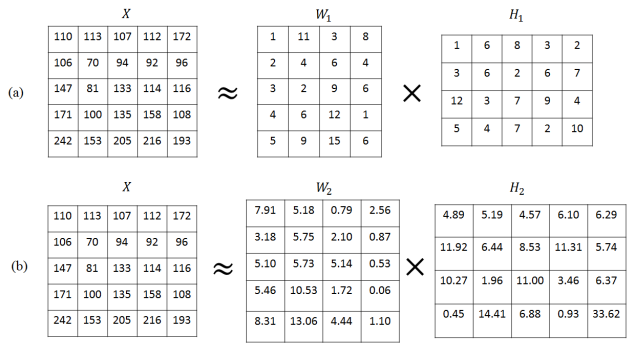

Fig. 2. Motivating example. (a) NMF obtains perfect factorization without any loss, but the rank-deficient basis may cause redundant representation, and (b) LDS-NMF obtains an approximate factorization with a small loss 8.28, but it can yield full column-rank basis.

to the predefined reduced dimensionality. This induces the so-called rank-deficiency problem, and prohibits them from stably representing limited high dimensional examples.

This paper proposes a Logdet divergence based sparse NMF method (LDS-NMF) to overcome the above drawbacks. Particularly, LDS-NMF reduces the risk of learning rank-deficient basis matrix by minimizing the Logdet divergence between the product of basis matrix with its transpose and the identity matrix. For example, to factorize the matrix given in Figure 2, the original NMF obtains a rank-deficient basis without any loss. Although LDS-NMF encourages a small loss, it can obtain more stable representation, i.e., the basis is full column-rank. To better represent a high-dimensional example in the learned low dimensional space, LDS-NMF simultaneously penalizes the density of its coefficient by minimizing its L_1-norm. Since sparse coefficients encourage to distribute energies over whole subspace, the incorporated Logdet divergence based regularization over the basis matrix and the L_1-norm regularization over the coefficients enhance each other. To take off the effect of outliers in the dataset, inspired by [15][16], LDS-NMF penalizes the reconstruction loss by $L_{2,1}$-norm due to its nice mathematical property. LDS-NMF is difficult to optimize because the objective function is jointly non-convex with respect to both basis matrix and coefficients. In this paper, we developed a multiplicative update rule to optimize LDS-NMF and proved its convergence. Experiments of both classification and clustering on popular image datasets suggest that LDS-NMF can learn more stable data representation than the representative NMF methods. The main contributions of this work are two-folds: 1) We incorporate the Logdet divergence regularization into NMF to reduce the risk of the rank-deficiency problem, and 2) We develop a multiplicative update rule (MUR) to optimize LDS-NMF and prove that the MUR converges theoretically.

II. LogDet Divergence Based Sparse NMF

This section, we proposed a Logdet divergence based sparse NMF (LDS-NMF) method to overcome the rank-deficiency problem of NMF as well as its variants based on the regularization theory.

A. The LDS-NMF Model

NMF [3][4] represents each example as a combination of several non-negative bases. It allows only additive or non-subtractive coefficients and avoids cancellation of energies, and thus can induce parts-based representation. Since such representation is in line with human intuition [6][7][8], NMF has been widely applied in data representation. Thus, the quality of the learned basis plays a key role for representing data. But rank deficient basis indicates one or several of bases to be linear combinations of the remaining bases so that the learned basis loses ability to represent some examples, and thus significantly undermines the representation ability [17][18].

The rank-deficiency problem of NMF implies the learned basis matrix to be not a full rank one, and might lead to a trivial solution in some situations. For example, in the extreme case, all the learned bases are identical when the given training set contains several copies of one example in scales. In the real world, the rank-deficiency problem of basis brings a big challenge in the clustering or classification tasks, because the centroid of multiple clusters are in the same position. It is hard to decide data samples belong to which clusters in this situation, because the centroids of clusters are identical.

To overcome this deficiency, we incorporated a Logdet divergence regularization over the basis matrix and proposed a novel Logdet divergence regularized sparse NMF method. The Logdet divergence [19][20][21][22] is defined based on the Bregman vector divergence [23] to measure the discrepancy between two multivariate Gaussian distributions under certain constraints such as the linear constraint [20]. Given any two positive definite matrices, namely and with the same dimensionality, their Logdet divergence is defined by:

$$D_{\varphi}(A, A_0) = \varphi(A) - \varphi(A_0) - \langle \nabla\varphi(A), A - A_0 \rangle , \qquad (1)$$

where $\varphi(A) = -\log\det(A)$, and the operator $\langle B, C\rangle = tr(B^T C)$ denotes the inner product. By simple algebra, the formula (1) can be written as:

$$D_{ld}(A, A_0) = tr(AA_0^{-1}) - \log\det(AA_0^{-1}) - m, \qquad (2)$$

where m denotes the dimensionality of A . The Logdet divergence has various properties [22][24] including: 1) scale invariance, i.e., $D_{ld}(\alpha A, \alpha A_0) = D_{ld}(A, A_0)$ for any positive α; 2) translation invariance, i.e., $D_{ld}(SAS^T, SA_0S^T) = D_{ld}(A, A_0)$ for any invertible matrix S; and 3) range space preservation, i.e., $range(A) = range(A_0)$ if and only if $D_{ld}(A, A_0)$ is finite.

Based on the third property of Logdet divergence, it is easy to verify the following observation: for any matrix $W \in R^{m\times r}$, $range(W^TW) = r$ if and only if $D_{ld}(W^TW, I_r)$ is finite, where I_r signifies the r-dimensional identity matrix. According to [22], we note that there is a feasible W with a finite $D_{ld}(W^TW, I_r)$ when minimizing $D_{ld}(W^TW, I_r)$. Therefore, we can preserve the range space of W by minimizing $D_{ld}(W^TW, I_r)$ according to the following *Lemma* 1.

Lemma 1: Given two positives m and r which satisfy $r \le m$, if the matrix $W' \in R^{m\times r}$ minimizes the following objective:

$$W' = \arg\min_{W} D_{ld}(W^TW, I_r), \qquad (3)$$

then $range(W') = r$.

Algorithm 1 The MUR algorithm for LDS-NMF

Input: Examples $X \in R^{m \times n}$, positive constant $\gamma > 1$, penalty parameter $\lambda > 0$
Output: W and H

1: Initialize: W and H.
2: **repeat**
3: Update W as follows:

$$W \leftarrow W \otimes \frac{XDH^T + \lambda W \left\lfloor (W^T W)^{-1} \right\rfloor_+}{WHDH^T + \lambda W + \lambda W \left\lfloor (W^T W)^{-1} \right\rfloor_-}.$$

4: Update H as follows: $H \leftarrow H \otimes \frac{W^T XD}{W^T WHD + \gamma}$.
5: Update D as follows: $D_{ii} \leftarrow 1/\sqrt{\sum_{j=1}^{m} (X - WH)_{ji}^2}$.
6: **until** {Converged.}

We leave the proof in **Appendix** A.

According to *Lemma* 1, the Logdet divergence based can reduce the risk of NMF to suffer from the rank-deficiency problem, i.e., minimizing $D_{ld}(W^T W, I_r)$ enforces that there exist at least r independent bases, in NMF. Based on the regularization theory, we have the following objective:

$$\min_{W \geq 0, H \geq 0} D(X|WH) + \frac{\lambda}{2} D_{ld}(W^T W, I_r), \tag{4}$$

where $D(\cdot|\cdot)$ signifies the loss function of NMF, e.g., Frobenius norm and KL-divergence, H denotes the coefficients, and λ signifies a positive tradeoff parameter. Although (4) preserves the range space of W, it still cannot prohibit W from a trivial solution, e.g., the resultant W contains a zero column. We therefore expected the energy to be distributed over all the low dimensions to overcome this deficiency. To this end, we incorporated sparsity regularization over coefficients H by minimizing the L_1-norm of each column of H. We have the objective of Logdet divergence regularized sparse NMF, i.e.,

$$\min_{W \geq 0, H \geq 0} D(X|WH) + \frac{\lambda}{2} D_{ld}(W^T W, I_r) + \gamma \sum_{j=1}^{n} \|H_{\cdot j}\|_1, \tag{5}$$

where $H_{\cdot j}$ represents the j-th column of H, and γ denotes the penalty parameter.

As analyzed in Section I, NMF often suffers from the rank-deficiency problem on the corrupted dataset, i.e., some coordinates of data matrix are contaminated by outliers. Therefore, to filter out the effect of outliers in LDS-NMF, we choose the $L_{2,1}$-norm to measure the reconstruction error in (5) due to its nice mathematical property [15][16], i.e., $D(X|WH) = \|X - WH\|_{2,1}$.

B. Optimization Algorithm

Although both the second and the third terms of (5) are convex, the objective function of LDS-NMF is jointly non-convex because the loss function in the first term are jointly non-convex with respect to both W and H. Here we developed a block coordinate descent based algorithm to optimize LDS-NMF by alternating updating W and H with multiplicative rules. Based on the auxiliary function technique, we can establish the following *Theorem* 1, where we denote the inverse operator of a matrix by $(\cdot)^{-1}$, and denote the positive and negative components of a matrix by $\lfloor \cdot \rceil_+$ and $\lfloor \cdot \rceil_-$, respectively.

Theorem 1: The objective function (5) is non-increasing under updating W, H, and D with steps 3, 4, and 5 in **Algorithm** 1.

We leave the proof in **Appendix** B.

Based on *Theorem* 1, we can summarize the optimization procedure of LDS-NMF in **Algorithm** 1. The major computation of **Algorithm** 1 is cost by steps 3, 4 and 5. They correspond to the updating rules for W, H and D, respectively. Step 3 requires computing the matrix inverse of a $r \times r$ matrix, which takes time of $O(r^3)$, and meanwhile involves a few matrix multiplications and lement-wise product and division operations that cost $O(mr^2 + n^2 r + nmr + mr + r^2 n)$. Thus, the total time complexity of Step 3 is $O(r^3 + mr^2 + n^2 r + nmr + mr + r^2 n)$. Step 4 executes the same matrix operations with ordinary MUR including a few matrix multiplications, and then the total running time is $O(rmn + rn^2 + mr^2 + rn)$. For Step 5, the matrix multiplication and subtraction costs time of $O(mnr)$ and $O(mn)$, respectively. Therefore, the total time complexity for **Algorithm** 1 equals to $O(r^3 + mr^2 + n^2 r + nmr + mr + r^2 n + rn)$.

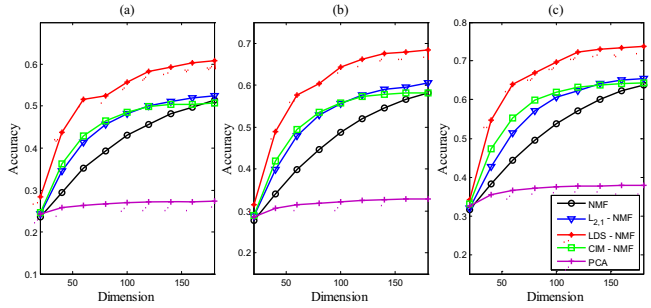

Fig. 3. Average accuracy versus reduced dimensionality when (a) 5, (b) 7, and (d) 9 images of each subject were randomly selected for training on the AR dataset.

TABLE I. THE HIGHEST FACE-RECOGNITION ACCURACY AND CORRESPONDING REDUCED DIMENSIONALITY (IN BRACKET) ON THE AR DATASET.

Method	P5(NN/SVM)	P7(NN/SVM)	P9(NN/SVM)
LDS-NMF	**0.6085(0.8377)**	**0.6836(0.8983)**	**0.7384(0.9283)**
$L_{2,1}$-NMF [15]	0.5251(0.8346)	0.6052(0.8838)	0.6646(0.9139)
CIM-NMF [13]	0.5064(0.8128)	0.5815(0.8680)	0.6361(0.9034)
NMF [4]	0.5137(0.7797)	0.5817(0.8369)	0.6383(0.8717)
PCA [2]	0.2733(0.7684)	0.3290(0.8391)	0.3781(0.8815)

III. EXPERIMENTS

A. Face Recognition

We first evaluates the effectiveness and efficiency of LDS-NMF compared to the representative NMF [4], $L_{2,1}$-NMF [15], CIM-NMF [13] and PCA [2] on two face image datasets including YaleB [9] and AR [25]. All algorithms are implemented in Matlab 7.10.0 on a workstation which contains four 3.4GHz Intel (R) Core (TM) processor and an 8 GB RAM.

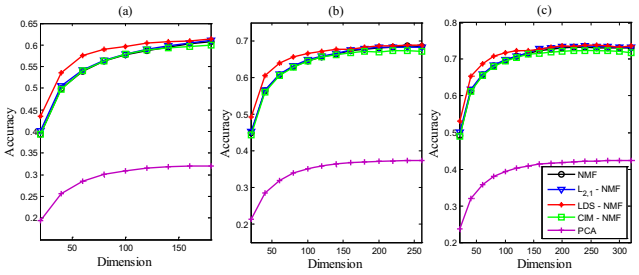

Fig. 4. Average accuracy versus reduced dimensionality when (a) 5, (b) 7, and (d) 9 images of each subject were randomly selected for training on the YaleB dataset.

TABLE II. THE HIGHEST RECOGNITION ACCURACY AND CORRESPONDING REDUCED DIMENSIONALITY (IN BRACKET) ON THE YALEB DATASET.

Method	P5	P7	P9
LDS-NMF	**0.6140(0.6909)**	**0.6894(0.7641)**	**0.7382(0.8264)**
$L_{2,1}$-NMF [15]	0.6113(0.6868)	0.6837(0.7484)	0.7367(0.8193)
CIM-NMF [13]	0.6002(0.6497)	0.6733(0.7237)	0.7235(0.7868)
NMF [4]	0.6109(0.6551)	0.6847(0.7268)	0.7348(0.7801)
PCA [2]	0.3089(0.6382)	0.3737(0.7107)	0.4242(0.7695)

For each dataset, we aligned all frontal face images according to the eye position and randomly selected different numbers of images from each individual as the training set, and the remainder is divided into the validation set and the testing set. All compared methods adopt the raw pixel values of images as features to learn a non-negative basis on the same training set, select the parameter on the same validation set, and are subsequently evaluated on the identical testing set.

To quantify the performance of all compared methods, we choose two well-known classifiers: the nearest neighbor (NN) and SVM to calculate the percentage of correctly classified testing images as the accuracy of face recognition. For fair comparison, each experiment was independently conducted 50 times to remove the influence of randomness. For NMF, $L_{2,1}$-NMF and CIM-NMF, they do not involve any parameter tuning. For LDS-NMF, the parameter selection will be introduced in the next subsection.

1) AR Dataset: The AR Database [25] consists of 2600 face images totally of 100 subjects. Each subject has 26 images with more varying facial expressions, different illumination conditions and more outliers, such as glasses, sunglasses and mufflers. In the experiment, we randomly select five, seven and nine images for per subject as three train datasets termed P5, P7 and P9, respectively. The rest of images are divided equally into validation dataset and test dataset. Figure 3 shows that LDS-NMF significantly outperforms other three methods under different configurations based on nearest neighbor classifier. More important, LDS-NMF achieves 10.03%, 9.80% and 9.45% relative improvement against the second best method in P5, P7 and P9, respectively. Figure 3 indicates that LDS-NMF achieves more improvements when the reduced dimensionality is higher. Moreover, Table I reports the highest recognition accuracy of the compared methods by two classifiers including nearest neighbor (NN) and support vector machine(SVM). It also implies the effectiveness of LDS-NMF in term of face recognition accuracy.

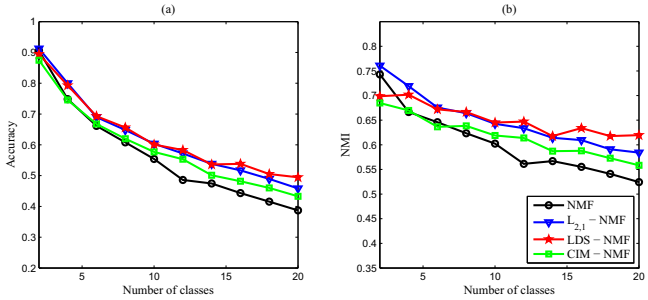

Fig. 5. Average accuracy (a) and averaged NMI (b) of LDS-NMF, $L_{2,1}$-NMF, CIM-NMF and NMF on the COIL20 dataset.

2) YaleB Dataset: The YaleB database [9] is an extension of the Yale face Database. There are totally 2424 face images of 38 individuals. Each individual has 59 images at least and 64 images at most under different illumination conditions. In our experiment, we randomly select five, seven and nine images for per individual as three train datasets termed P5, P7 and P9, respectively. The remainder is divided equally into the validation and test dataset. Figure 4 shows that LDS-NMF is superior to the compared methods under varying dimensionalities in terms of average recognition accuracy. Table II also shows that LDS-NMF method also achieves the highest recognition accuracy compared with the representative methods in both classifiers.

B. Image Clustering

To evaluate the clustering performance of LDS-NMF, we compare LDS-NMF with NMF, $L_{2,1}$-NMF and CIM-NMF on COIL [26] and JAFFE [27] datasets with clustering accuracy and normalized mutual information (NMI) [28]. In clustering tasks, we adopt the raw pixels to learn the coefficients of examples and then utilize the K-means method to cluster the learned coefficients. Note that the number of the clusters as a prior knowledge is set to the number of selected objects. For NMF, $L_{2,1}$-NMF and CIM-NMF, they do not involve any parameter tuning. For LDS-NMF, the parameter selection will be introduced in the next subsection.

1) COIL20 Dataset: The COIL20 dataset contains 1,440 images with the uniform black background for 20 objects. Each object has 72 images under different angles. Each image is cropped to 32×32 pixels. We randomly select different numbers of objects to evaluate the clustering performance, and meanwhile each trail is independently conducted 50 times to remove the influence of randomness for fair comparison. Figure 5 shows that LDS-NMF is superior to the compared methods under different number of classes in terms of average accuracy and NMI.

2) JAFFE Dataset: The JAFFE dataset contains 213 face images belong on 10 Japanese females. Each person has around 20-23 images with varying facial expressions. Each image is cropped to 256×256 pixels. We randomly select different numbers of individuals as the training subsets. Figure 6 shows that LDS-NMF outperforms the compared methods under varying number of classed in terms of average accuracy and NMI.

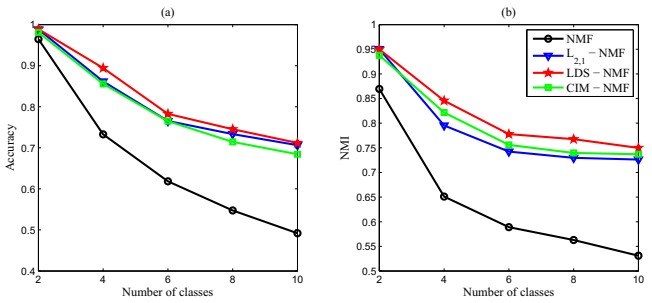

Fig. 6. Average accuracy (a) and averaged NMI (b) of LDS-NMF, $L_{2,1}$-NMF, CIM-NMF and NMF on the JAFFE dataset.

C. Parameter Settings

Firstly, we show how to select proper parameters λ and γ for LDS-NMF in face recognition. It is time-consuming to select the value of parameters based on the grid search. Another alternative to this problem is to select a feasible range for them, because the proper range has little effect on the performance of LDS-NMF. Due to the symmetric property of both parameters, we merely introduce the procedure to select the parameter λ with γ fixed as follows: a) solve the basis and the coefficients with setting both parameters to zero on the training set, b) calculate both parameters using $\lambda_0 = 2\|X - WH\|_{2,1}/D_{ld}(W^TW, I_r)$ and $\gamma_0 = \|X - WH\|_{2,1}/\sum_{j=1}^{n}\|H_{.j\|_1}$, and c) collect all the average accuracy on the validation set based on the leaned basis when we fix $\gamma = \gamma_0$ and adapt varying values of λ from 10^{d_1} to 10^{d_2}, wherein the integral d_1 gradually increases with one step until no larger than d_2. To this end, we keep the parameter value corresponding to the highest average accuracy as the resultant parameter value. In the empirical studies, we set the range of λ from 10^{-5} to 10, and range of γ from 10^{-3} to 10^3 on YaleB and AR datasets. For parameter selection in image clustering, we adopt the identical strategy used in image classification for this purpose.

IV. Conclusion

This paper proposes a Logdet divergence based sparse NMF method to solve the rank-deficiency problem of the learned lower dimensional basis. Particularly, LDS-NMF incorporates the Logdet divergence based regularization over the basis matrix into NMF together with an L_1-norm based regularization over the coefficients. Since these two regularizations positively effect each other, LDS-NMF reduces the risk of rank-deficiency problem and enhance the representation ability of NMF. Experimental results of face recognition on popular face image datasets verify that LDS-NMF outperforms NMF, $L_{2,1}$-NMF and CIM-NMF in quantities.Based on the effective Logdet divergence based regularization, our future works will further integrate it into NMF variants for other vision tasks.

Acknowledgement

The research was supported in part by grants from 973 project 2013CB329006, China NSFC under Grant 61173156, RGC under the contracts CERG 622613, 16212714, HKUST6/CRF/12R, and M-HKUST609/13, as well as the grant from Huawei-HKUST joint lab.

Appendix

A. *Proof of Lemma 1*

Given $W' \in R^{m\times r}$, it is obvious that the objective (3) results from (2) when $A = W'^TW'$ and $A_0 = I$. By simple algebra, there holds $range(W'^TW') \le range(W')$. According to [22], we obtain $range(W'^TW') = range(I_r) = r$. Thus, we have:

$$r = range(W'^TW') \le range(W') \le \min\{m, r\} = r. \quad (6)$$

This completes the proof. ■

B. *Proof of Theorem 1*

We prove by utilizing the majorization minimization technique. According to [15][16], we have

$$\begin{aligned}&\left\|X - WH^{t+1}\right\|_{2,1} - \left\|X - WH^t\right\|_{2,1}\\ &\le \frac{1}{2}\left[tr(X - WH^{t+1})D(X - WH^t)^T\right], \end{aligned} \quad (7)$$

where $D_{ii} = 1/\|x_i - Wh_i^t\|$, t is the iteration counter.

Since $tr(Z^TAZB) \le \sum_{ik}(AZ'B)_{ik}\frac{Z_{ik}^2}{Z'_{ik}}$, let $A = I$, $Z = W$, and $B = HDH^T$, we have:

$$tr(WHDH^TW^T) \le \sum_{ik}\frac{(W'HDH^T)_{ik}W_{ik}^2}{W'_{ik}}. \quad (8)$$

Thus, auxiliary function of $tr(X - WH)D(X - WH)^T$ is:

$$tr(XDX^T - 2H^TW^TXD) + \sum_{ik}\frac{(W'HDH^T)_{ik}W_{ik}^2}{W'_{ik}}. \quad (9)$$

Combining (7)-(9), we have the following auxiliary function of $\|X - WH\|_{2,1}$:

$$\frac{1}{2}\left(tr(XDX^T - 2H^TW^TXD) + \sum_{ik}\frac{(W'HDH^T)_{ik}W_{ik}^2}{W'_{ik}}\right). \quad (10)$$

Likewise, we use the matrix inequality $tr(Z^TAZB) \le \sum_{ik}(AZ'B)_{ik}\frac{Z_{ik}^2}{Z'_{ik}}$, and let the left term $A = I$, $Z = W$, and $B = I$, and thus there holds:

$$\frac{\lambda}{2}tr(W^TW) \le \frac{\lambda}{2}\sum_{ik}\frac{W'_{ik}W_{ik}^2}{W'_{ik}}. \quad (11)$$

Besides, the auxiliary function of $\frac{\lambda}{2}\log\det(W^TW)$ is:

$$\begin{aligned}&\frac{\lambda}{2}\log\det(W'^TW')\\ &+ \lambda\sum_{ik}(W'(W'^TW')_+^{-1})_{ik}W'_{ik}(1 + \log\frac{W_{ik}}{W'_{ik}})\\ &- \lambda\sum_{ik}(W'(W'^TW')_-^{-1})_{ik}\frac{W_{ik}'^2 + W_{ik}^2}{2W'_{ik}} + 2r. \end{aligned} \quad (12)$$

Based on the first-order Taylor series expansion, we can expand $\log\det(W^TW)$ as: $\log\det(W^TW) \approx \log\det(W'^TW') + 2tr(W'(W'^TW')^{-1}(W - W')^T) + 2r$.

Then we can obtain the upper bound of $\log\det(W^TW)$ as follows:

$$
\begin{aligned}
&-\log\det(W'^TW') - 2tr(W^TW'(W'^TW')_+^{-1}) \\
&+ 2tr(W^TW'(W'^TW')_-^{-1}) - 2r \leq -\log\det(W'^TW') \\
&- 2\sum_{ik}(W'(W'^TW')_+^{-1})_{ik}W'_{ik}(1+\log\frac{W_{ik}}{W'_{ik}}) \\
&+ 2\sum_{ik}(W'(W'^TW')_-^{-1})_{ik}\frac{W'^2_{ik}+W^2_{ik}}{2W'_{ik}} - 2r. \qquad (13)
\end{aligned}
$$

According to (10)-(13), we can construct the auxiliary function $G(W,W')$ of $J(W,H)$ as follows:

$$
\begin{aligned}
&\frac{1}{2}\left(tr(XDX^T - 2H^TW^TXD) + \sum_{ik}\frac{(W'HDH^T)_{ik}W^2_{ik}}{W'_{ik}}\right) \\
&+ \frac{\lambda}{2}\sum_{ik}\frac{W'_{ik}W^2_{ik}}{W'_{ik}} - \frac{\lambda}{2}\log\det(W'^TW') \\
&- \lambda\sum_{ik}(W'(W'^TW')_+^{-1})_{ik}W'_{ik}(1+\log\frac{W_{ik}}{W'_{ik}}) \\
&+ \lambda\sum_{ik}(W'(W'^TW')_-^{-1})_{ik}\frac{W'^2_{ik}+W^2_{ik}}{2W'_{ik}} - 2r. \qquad (14)
\end{aligned}
$$

Then we can obtain the derivative of $G(W,W')$ as follows:

$$
\begin{aligned}
&\frac{\partial}{\partial W}G(W,W') = -(XDH^T)_{ik} + \frac{(W'HDH^T)_{ik}W_{ik}}{W_{ik}} \\
&+ \lambda\frac{W'_{ik}W_{ik}}{W'_{ik}} - \lambda(W'(W'^TW')_+^{-1})_{ik}\frac{W'_{ik}}{W_{ik}} \\
&- \lambda(W'(W'^TW')_+^{-1})_{ik}\frac{W_{ik}}{W'_{ik}}. \qquad (15)
\end{aligned}
$$

By (15), we derive the update rule for W (cf. Step 3 in **Algorithm** 1). Likewise, we can yield the update rule for H (cf. Step 4 in **Algorithm** 1). The auxiliary function can help the objective function to get the following inequality:

$$J(W^{t+1},H^{t+1}) \leq J(W^t,H^{t+1}) \leq J(W^t,H^t), \qquad (16)$$

where $J(W,H)$ is the objective function. Since the objective (5) has the lower bound, we know that the sequences generated by **Algorithm** 1 converge to a local saddle point, according to (16). Moreover, the derivatives equal to zero when the equality of (16) holds. Thus, such saddle point is still a local solution as well. This completes the proof. ■

REFERENCES

[1] R. Bellman and R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 1961, vol. 4.

[2] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1, pp. 37–52, 1987.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[4] ——, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[5] P. O. Hoyer, "Non-negative sparse coding," in *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 557–565.

[6] S. E. Palmer, "Hierarchical structure in perceptual representation," *Cognitive psychology*, vol. 9, no. 4, pp. 441–474, 1977.

[7] N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," *Annual review of neuroscience*, vol. 19, no. 1, pp. 577–621, 1996.

[8] E. Wachsmuth, M. Oram, and D. Perrett, "Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque," *Cerebral Cortex*, vol. 4, no. 5, pp. 509–522, 1994.

[9] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[10] P. O. Hoyer, "Non-negative matrix factorization with sparsensss constraints," *Journal of machine learning research*, vol. 5, pp. 1457–1469, 2004.

[11] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[12] Y. Chen, J. Zhang, D. Cai, W. Liu, and X. He, "Nonnegative local coordinate factorization for image representation," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 969–979, 2013.

[13] L. Du, X. Li, and Y.-D. Shen, "Robust nonnegative matrix factorization via half-quadratic minimization." in *ICDM*, 2012, pp. 201–210.

[14] S. Yang, C. Hou, C. Zhang, Y. Wu, and S. Weng, "Robust non-negative matrix factorization via joint sparse and graph regularization," in *The International Joint Conference on Neural Networks*, 2013, pp. 1–5.

[15] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l21-norm," in *The 20th ACM international conference on Information and knowledge management*, 2011, pp. 673–682.

[16] C. Ding, D. Zhou, X. He, and H. Zha, "R1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 281–288.

[17] R. M. OBrien, "Visualizing rank deficient models: a row equation geometry of rank deficient matrices and constrained-regression," *PloS one*, vol. 7, no. 6, p. e38923, 2012.

[18] P. Deuflhard and W. Sautter, "On rank-deficient pseudoinverses," *Linear Algebra and its Applications*, vol. 29, pp. 91–111, 1980.

[19] S. Sra and I. S. Dhillon, "Generalized nonnegative matrix approximations with bregman divergences," in *Advances in neural information processing systems*, 2005, pp. 283–290.

[20] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 209–216.

[21] I. S. Dhillon and J. A. Tropp, "Matrix nearness problems with bregman divergences," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 4, pp. 1120–1146, 2007.

[22] B. Kulis, M. A. Sustik, and I. S. Dhillon, "Low-rank kernel learning with bregman matrix divergences," *The Journal of Machine Learning Research*, vol. 10, pp. 341–376, 2009.

[23] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR computational mathematics and mathematical physics*, vol. 7, no. 3, pp. 200–217, 1967.

[24] B. Kulis, "Metric learning: A survey," *Foundations and Trends in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2012.

[25] A. M. Martinez, "The ar face database," *CVC Technical Report*, 1998.

[26] S. A. Nene, S. K. Nayar, H. Murase *et al.*, "Columbia object image library (coil-20)," Technical Report CUCS-005-96, Tech. Rep., 1996.

[27] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200–205.

[28] C. Studholme, D. L. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern recognition*, vol. 32, no. 1, pp. 71–86, 1999.