Robust sparse nonnegative matrix factorization based on maximum correntropy criterion

Siyuan Peng, Wee Ser and Zhiping Lin
School of Electrical and Electronic Engineering
Nangyang Technological University, 639798, Singapore
Email: {PENG0074@e., EZPLin@, ewser@}ntu.edu.sg

Badong Chen
Institute of Artificial Intelligence and Robotics,
Xi'an Jiaotong University, Xian, ShanXi 710049, China
Email: chenbd@mail.xjtu.edu.cn

***Abstract*—Nonnegative matrix factorization (NMF) is a significant matrix decomposition technique for learning parts-based, linear representation of nonnegative data, which has been widely used in a broad range of practical applications such as document clustering, image clustering, face recognition and blind spectral unmixing. Traditional NMF methods, which mainly minimize the square of the Euclidean distance or the Kullback-Leibler (KL) divergence, seriously suffer the outliers and non-Gaussian noises. In this paper, we propose a robust sparse nonnegative matrix factorization algorithm, called l_1-norm nonnegative matrix factorization based on maximum correntropy criterion (l_1-CNMF). Specifically, l_1-CNMF is derived from the traditional NMF algorithm by incorporating the l_1 sparsity constraint and maximum correntropy criterion. Numerical experiments on the Yale database and the ORL database with and without apparent outliers show the effectiveness of the proposed algorithm for image clustering compared with other existing related methods.**

I. Introduction

In recent years, nonnegative matrix factorization (NMF) technique, as a fundamental tool for data representation, has received considerable attention due to its various applications in the fields of document clustering, image clustering, face recognition, blind spectral unmixing and so on [1], [2], [3]. The main ideal of NMF is to find an approximate decomposition of a nonnegative data matrix into two low-rank nonnegative matrix factors (basis matrix and coefficient matrix). Under the condition that the two matrix factors must be nonnegative, NMF can yield a parts-based representation of the original data [4], [5]. Generally speaking, most of the objective functions or optimization criteria for the matrix decomposition used in the traditional NMF algorithm and its extensions are based on the minimum the square of the Euclidean distance (ED) or the Kullback-Leibler (KL) divergence, due to their attractive advantages such as simplicity, nice mathematical properties and optimality under Gaussian noise. However, in many real world applications, the data usually contains different types of non-Gaussian noise or outliers. In this situation, the performance of the traditional NMF algorithm and its extensions will perform poorly.

In order to improve the robustness of the original NMF algorithm, some variants based on the robust objective functions have been successfully proposed in [6], [7], [8], which are insensitive to the outliers and non-Gaussian noises. Among these robust approaches, maximum correntropy criterion (M-CC) [9], [10], [11] based NMF approaches have demonstrated the superior performance in many applications of engineering. For example, Wang et al. developed a novel NMF algorithm based on MCC for the gene expression data-based cancer clustering problem [12]; Du et al. utilized the MCC into NMF, and proposed three new algorithms for face recognition [13]; Wang et al. presented a robust model for unsupervised hyperspectral unmixing with correntropy-based metric [14]. The main reason is that, as a nonlinear and local similarity measure, correntropy can directly measure the probability of how similar two random variables are in the joint space. Consequently, correntropy is robust to non-Gaussian noises and large outliers. However, up to now, sparse NMF based on MCC for image clustering has not been studied yet in the literature.

In this paper, a robust sparse nonnegative matrix factorization algorithm, called l_1-norm nonnegative matrix factorization based on maximum correntropy criterion (l_1-CNMF), has been proposed, which is derived by incorporating the l_1 sparsity constraint [15] and maximum correntropy criterion into the traditional NMF algorithm. l_1-CNMF is insensitive to outliers and non-Gaussian noises, and achieves better performance than the MCC-based NMF methods without sparsity constraint. Experimental results illustrate the effectiveness of the proposed l_1-CNMF algorithm on the Yale database and the ORL database with and without apparent outliers for image clustering in comparison with other existing related methods.

The rest of this paper is organized as follows. In Section II, after briefly introducing the MCC, we derive the l_1-CNMF algorithm. Experimental results are presented in Section III. Finally, Section IV draws the conclusion.

II. Robust Sparse Nonnegative Matrix Factorization

A. MCC Algorithm

Maximum correntropy criterion was proposed in information theoretic learning (ITL) [9], which has been widely used to process non-Gaussian noises and outliers. As a generalized similarity measure between two random variables A and B, correntropy is defined by [10], [11]

$$V(A,B) = E[\kappa(A,B)] = \int \kappa(a,b)dF_{AB}(a,b) \qquad (1)$$

where $\kappa(\cdot,\cdot)$ denotes a shift-invariant Mercer kernel, $E[\cdot]$ stands for the expectation operator, and $F_{AB}(a,b)$ is the joint distribution function of (a,b). In this paper, only the Gaussian kernel is used for correntropy with $\kappa_\sigma(a,b) = \exp(-\frac{(a-b)^2}{2\sigma^2})$, where σ denotes the kernel bandwidth.

In practical applications, the join distribution $F_{AB}(a,b)$ is usually unknown, and only a finite number of data $\{(a_s,b_s)\}_{s=1}^S$ are available. In this situation, we have the following sample estimator of correntropy:

$$\hat{V}_{S,\sigma} = \frac{1}{S}\sum_{s=1}^{S}\kappa_\sigma(a_s,b_s) \tag{2}$$

B. l_1-CNMF Algorithm

Assume that there is a nonnegative matrix $\mathbf{X} = [\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N] \in \mathbb{R}_+^{M\times N}$, where each column in $\mathbf{X}$ denotes a sample vector containing M elements. NMF aims to factorize $\mathbf{X}$ into two nonnegative matrices $\mathbf{W} \in \mathbb{R}_+^{M\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ (called basis matrix and coefficient matrix respectively), such that the produce of $\mathbf{W}$ and $\mathbf{H}$ closely approximates the original matrix $\mathbf{X}$:

$$\mathbf{X} \approx \mathbf{WH} \tag{3}$$

In order to to quantify the quality of the decomposition, two objective functions, namely, the square of the Euclidean distance and the Kullback-Leibler divergence, are frequently used in previous studies, which are respectively formulated by [5]

$$D_{ED} = \sum_{m=1}^{M}\sum_{n=1}^{N}\left(\mathbf{X}_{mn} - \sum_{k=1}^{K}\mathbf{W}_{mk}\mathbf{H}_{kn}\right)^2 \tag{4}$$

$$D_{KL} = \sum_{m=1}^{M}\sum_{n=1}^{N}\left(\mathbf{X}_{mn}\ln\frac{\mathbf{X}_{mn}}{(\mathbf{WH})_{mn}} - \mathbf{X}_{mn} + (\mathbf{WH})_{mn}\right) \tag{5}$$

where $\mathbf{X}_{mn}$, $\mathbf{W}_{mk}$, and $\mathbf{H}_{kn}$ denote the element in $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{H}$, respectively.

In this paper, instead of using the square of the Euclidean distance and the Kullback-Leibler divergence, we use the correntropy as the objective function, and maximize the correntropy between the original matrix $\mathbf{X}$ and the produce of $\mathbf{W}$ and $\mathbf{H}$. In addition, due to the fact that NMF does not always result in parts-based representations, sparsity constraint has been successfully applied in NMF to improve the found decomposition into parts [3], [16]. Therefore, we incorporate the l_1-norm sparsity constraint of the coefficient matrix $\mathbf{H}$ into MCC-based nonnegative matrix factorization technique, and propose a robust sparse nonnegative matrix factorization algorithm, called l_1-norm nonnegative matrix factorization based on maximum correntropy criterion. Accordingly, the objective function of l_1-CNMF is defined as follows:

$$\begin{aligned} D_{l_1-CNMF} = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\kappa_\sigma(\mathbf{X}_{mn},\sum_{k=1}^{K}\mathbf{W}_{mk}\mathbf{H}_{kn})\right) \\ - 2\lambda\sum_{n=1}^{N}\parallel \mathbf{H}_n \parallel_1 \\ s.t.\quad \mathbf{W}_{mk}\geq 0,\ \mathbf{H}_{kn}\geq 0\ \forall m,k,n \end{aligned} \tag{6}$$

where λ denotes a regularization parameter that controls the sparseness. Then we have the following optimization problem:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}} D_{l_1-CNMF} = \sum_{m=1}^{M}\sum_{n=1}^{N}\left(-\kappa_\sigma(\mathbf{X}_{mn},\sum_{k=1}^{K}\mathbf{W}_{mk}\mathbf{H}_{kn})\right) \\ + 2\lambda\sum_{n=1}^{N}\parallel \mathbf{H}_n \parallel_1 \\ s.t.\quad \mathbf{W}_{mk}\geq 0,\ \mathbf{H}_{kn}\geq 0,\ \forall m,k,n \end{aligned} \tag{7}$$

Obviously, it is difficult to solve the above optimization problem directly, since the objective function in (7) is non-quadratic and non-convex with respect to $\mathbf{W}$ and $\mathbf{H}$ together. However, in recent years, the half-quadratic technique has been successfully employed to solve nonlinear Information theoretic learning optimization problem [17]. In this paper, we adopt the half-quadratic technique to solve (7). Convex conjugate function is a commonly used transformation for transforming a non-convex optimization problems into its dual problem which is easier to solve. Based on the property of the convex conjugated function, we have the following proposition:

Proposition 1: There exists a convex conjugated function φ of $g(x)$ such that

$$g(x) = \max_u\left(ux - \varphi(u)\right), \tag{8}$$

and for a fixed x, the maximum is reached at $u = -g(x)$.

Combining (7) and (8), we derive the following augmented objective function

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{H},\mathbf{U}} \hat{D}_{l_1-CNMF} \\ &s.t.\quad \mathbf{W}_{mk}\geq 0,\ \mathbf{H}_{kn}\geq 0,\ \forall m,k,n \\ &\hat{D}_{l_1-CNMF} = 2\lambda\sum_{n=1}^{N}\parallel \mathbf{H}_n \parallel_1 + \\ &\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\mathbf{U}_{mn}(\mathbf{X}_{mn} - \sum_{k=1}^{K}\mathbf{W}_{mk}\mathbf{H}_{kn})^2 + \varphi(\mathbf{U}_{mn})\right) \end{aligned} \tag{9}$$

where $\mathbf{U}_{mn}$ denotes a element of the nonnegative matrix $\mathbf{U} \in \mathbb{R}_+^{M\times N}$. Therefore, minimizing (7) is equivalent to minimizing (9). By the following alternate minimization, the augmented objective function $\hat{D}_{l_1-CNMF}$ can be optimized.

*1) Computation of **U**:* When $\mathbf{W}$ and $\mathbf{H}$ are fixed, we derive

$$\begin{aligned}\mathbf{U}_{mn} &= \kappa_\sigma(\mathbf{X}_{mn}, \sum_{k=1}^{K} \mathbf{W}_{mk}\mathbf{H}_{kn}) \\ &= \exp\left(- \frac{\left(\mathbf{X}_{mn} - \sum\limits_{k=1}^{K} \mathbf{W}_{mk}\mathbf{H}_{kn}\right)^2}{2\sigma^2} \right) \end{aligned} \tag{10}$$

where the kernel bandwidth is computed by [12], [13]

$$\sigma^2 = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(\mathbf{X}_{mn} - \sum_{k=1}^{K} \mathbf{W}_{mk}\mathbf{H}_{kn} \right)^2 \tag{11}$$

*2) Computation of **W**:* When $\mathbf{U}$ is fixed, minimizing $\hat{D}_{l_1-CNMF}$ with respect to $\mathbf{W}$ is equivalent to minimizing

$$\begin{aligned} L(\mathbf{W}) = &\sum_{m=1}^{M} (\mathbf{X}_{m*} - \mathbf{W}_{m*}\mathbf{H})\, diag(\mathbf{U}_{m*})\, (\mathbf{X}_{m*} - \mathbf{W}_{m*}\mathbf{H})^T \\ &+ 2\lambda \sum_{n=1}^{N} \parallel \mathbf{H}_n \parallel_1 + Tr(\Phi\mathbf{W}) \end{aligned} \tag{12}$$

where $\mathbf{X}_{m*}$ and $\mathbf{W}_{m*}$ denote the mth row in $\mathbf{X}$ and $\mathbf{W}$ respectively, $diag(\cdot)$ is an operator that converts the vector to a diagonal matrix, T stands for transpose operator, $Tr(\cdot)$ denotes the trace of a matrix, and $\Phi = [\Phi_{mk}] \in \mathbb{R}^{M\times K}$ is the Lagrange multiplier for the nonnegative constraint $\mathbf{W}_{mk} \geq 0$. Similar to [13], using the partial derivative of $L(\mathbf{W})$ respect to $\mathbf{W}$ and the Karush-Kuhn-Tucker condition (i.e., $\Phi_{mk}\mathbf{W}_{mk} = 0$) [18], [19], we derive the following update rule

$$\mathbf{W}_{mk} = \mathbf{W}_{mk} \frac{(\mathbf{U} \otimes \mathbf{X}\mathbf{H}^T)_{mk}}{(\mathbf{U} \otimes (\mathbf{W}\mathbf{H})\mathbf{H}^T)_{mk}} \tag{13}$$

where $\otimes$ denotes the Hadamard product.

*3) Computation of **H**:* When $\mathbf{U}$ is fixed, minimizing $\hat{D}_{l_1-CNMF}$ with respect to $\mathbf{H}$ is equivalent to

$$\begin{aligned} L(\mathbf{H}) = &\sum_{n=1}^{N} (\mathbf{X}_{*n} - \mathbf{W}\mathbf{H}_{*n})^T\, diag(\mathbf{U}_{*n})\, (\mathbf{X}_{*n} - \mathbf{W}\mathbf{H}_{*n}) \\ &+ 2\lambda \sum_{n=1}^{N} \parallel \mathbf{H}_n \parallel_1 + Tr(\Psi\mathbf{H}) \end{aligned} \tag{14}$$

where $\mathbf{X}_{*n}$ and $\mathbf{H}_{*n}$ denote the nth column in $\mathbf{X}$ and $\mathbf{H}$ respectively, and $\Psi = [\Psi_{kn}] \in \mathbb{R}^{K\times N}$ is the Lagrange multiplier for the nonnegative constraint $\mathbf{H}_{kn} \geq 0$. Then, the partial derivative of $L(\mathbf{H})$ with respect to $\mathbf{H}$ is

$$\begin{aligned} \frac{\partial L(\mathbf{H})}{\partial \mathbf{H}_{kn}} = &- 2(\mathbf{W}^T diag(\mathbf{U}_{*n})\mathbf{X}_{*n})_k + \\ &2(\mathbf{W}^T\mathbf{W}\mathbf{H}_{*n})_k + 2\lambda(\mathbf{H}_{*n})_k + \Psi_{kn} \end{aligned} \tag{15}$$

Similarly, using the Karush-Kuhn-Tucker condition (i.e., $\Psi_{kn}\mathbf{H}_{kn} = 0$), we obtain

$$\begin{aligned} \big(-2(\mathbf{W}^T diag(\mathbf{U}_{*n})\mathbf{X}_{*n})_k + 2(\mathbf{W}^T\mathbf{W}\mathbf{H}_{*n})_k + \\ 2\lambda(\mathbf{H}_{*n})_k\big)\, \mathbf{H}_{kn} = 0 \end{aligned} \tag{16}$$

After some straightforward manipulations, we have the following update rule

$$\mathbf{H}_{kn} = \mathbf{H}_{kn} \frac{(\mathbf{W}^T(\mathbf{U} \otimes \mathbf{X}))_{kn}}{(\mathbf{W}^T(\mathbf{U} \otimes (\mathbf{W}\mathbf{H})))_{kn} + \lambda} \tag{17}$$

It is worth noting that, if $\mathbf{W}$ and $\mathbf{H}$ are the solution for the CNMF-l^1 algorithm, then $\mathbf{W}\mathbf{Y}^{-1}$ and $\mathbf{Y}\mathbf{H}$ are also a solution for any positive diagonal matrix $\mathbf{Y}$ due to $\mathbf{W}\mathbf{H} = (\mathbf{W}\mathbf{Y}^{-1})(\mathbf{Y}\mathbf{H})$, which will lead $\mathbf{H}$ to be zero. To eliminate this problem in many practical areas, one strategy is frequently used to normalize each column of $\mathbf{W}$ to be 1 [14], [20]. Then we have

$$\mathbf{W}_{mk} \leftarrow \frac{\mathbf{W}_{mk}}{\sqrt{\sum\limits_{m=1}^{M} \mathbf{W}_{mk}^2}},\ \mathbf{H}_{kn} \leftarrow \mathbf{H}_{kn} \sqrt{\sum_{m=1}^{M} \mathbf{W}_{mk}^2} \tag{18}$$

Furthermore, based on the multiplicative updates, the computational complexity of the proposed algorithm is $O(MKN)$, which is the same as the traditional NMF approach. The detail steps for l_1-CNMF are shown in algorithm 1.

```
Algorithm 1 $l_1$-CNMF Algorithm
Input: The data matrix $\mathbf{X} =\in \mathbb{R}_+^{M\times N}$, the initial matrices
  $\mathbf{W} \in \mathbb{R}_+^{M\times K}$, $\mathbf{H} \in \mathbb{R}_+^{K\times N}$, and the regularization parameter $\lambda$.
Output: $\mathbf{W}$, $\mathbf{H}$
Computation:
repeat
    1) Update $\mathbf{U}$ by using (10);
    2) Update $\mathbf{W}$ by using (13);
    3) Update $\mathbf{H}$ by using (17);
    4) Update $\sigma^2$ by using (11);
until convergence
```

III. Experiments

In this section, we investigate the performance of the proposed l_1-CNMF algorithm for image clustering with/without apparent outliers, compared with six existing related methods including the standard Kmeans algorithm (Kmeans)[21], NMF [5], l_1-norm based NMF (l_1-NMF) [16], NMF method based on the correntropy induced metric (CIM-NMF) [13], MCC-based NMF (NMF-MCC) [12], and corrrentropy based NMF with an l_1 sparse penalty term (l_1-CENMF) [14]. Note that the main difference between l_1-CNMF and l_1-CENMF is that, l_1-CNMF considers every element in $\mathbf{X}$ as a whole, while l_1-CENMF considers entire row in $\mathbf{X}$ as a whole. In addition, for all NMF based methods, the matrices $\mathbf{W}$ and $\mathbf{H}$ have the same initial values, which are selected randomly, and the parameter K is set to be the same as the number of clusters. Experiment results are averaged over 30 independent Monte Carlo runs with different initial values.

A. Data Sets and Evaluation Metric

Two image data sets are used in the experiments. The first data set is the Yale data set[1], which includes 165 gray scale

[1] http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

TABLE I
CLUSTERING PERFORMANCE ON YALE DATABASE

K	2	3	4	5	6	7	8	9	10	12	15
Kmeans	69.77	59.55	54.55	52.63	48.48	46.10	42.15	38.93	40.03	36.89	36.78
NMF	67.96	66.06	56.02	52.90	51.67	48.44	45.34	43.03	43.59	41.97	38.84
l_1-NMF	66.59	64.09	56.93	54.09	52.05	48.89	47.55	42.67	43.18	41.51	36.97
NMF-MCC	64.09	66.82	57.61	54.54	51.89	48.57	44.94	46.21	45.55	43.14	40.48
CIM-NMF	83.86	73.78	59.20	52.45	52.05	51.16	48.75	46.41	48.50	44.28	42.36
l_1-CENMF	66.14	64.39	61.13	53.54	52.65	50.58	47.89	45.45	44.63	40.11	40.39
l_1-CNMF	**87.05**	**77.12**	**62.71**	**58.63**	**55.68**	**54.35**	**51.42**	**48.18**	**49.51**	**45.03**	**43.67**

TABLE II
CLUSTERING PERFORMANCE ON ORL DATABASE

K	2	3	4	5	6	7	8	9	10	15	20
Kmeans	75.50	70.33	59.43	53.03	49.83	50.29	48.06	49.61	49.30	46.83	40.55
NMF	60.20	64.00	56.83	53.33	48.33	52.13	48.06	52.00	47.32	45.43	43.47
l_1-NMF	60.40	64.33	55.27	54.67	51.83	50.93	47.75	53.06	47.21	46.63	43.22
NMF-MCC	89.50	77.67	78.40	71.83	67.67	69.29	68.56	69.53	69.18	64.80	60.05
CIM-NMF	92.50	73.67	74.83	71.49	68.17	72.57	66.06	68.78	70.80	64.63	61.62
l_1-CENMF	93.50	76.00	76.03	72.12	66.83	71.93	69.13	68.06	67.45	64.90	59.12
l_1-CNMF	**96.00**	**82.33**	**80.00**	**77.34**	**73.00**	**73.36**	**72.54**	**70.28**	**70.95**	**67.43**	**61.92**

images of 15 objects viewed under different angles. Each object have 11 different images, and each image contains a size of 32×32 pixels. The second data set is the ORL data set[2], which contains 400 gray scale images of 40 objects. Each object have 10 facial images under different light and illumination conditions, and each image also contains a size of 32×32 pixels.

We adopt the accuracy (ACC) as the evaluation metric to evaluate the performance of the proposed algorithm quantitatively, which is defined by

$$\text{ACC} = \frac{\sum\limits_{n=1}^{N} \delta(r_n, map(c_n))}{N} \tag{19}$$

where N denotes the total number of a data set, r_n and c_n are respectively the label provided by the data set and the cluster label, $\delta(h, z)$ stands for the delta function that equals 1 if $h = z$ and equals 0 otherwise, and $map(\cdot)$ is the mapping function that map each cluster label to the equivalent label from the data set by using the Kuhn-Munkres algorithm [22].

B. Clustering Results

Two parameters are used in l_1-CNMF algorithm: the kernel bandwidth σ and the regularization parameter λ. Throughout the experiments, the kernel bandwidth σ is adopted by using (11), and the regularization parameter λ is empirically set to 0.1 for all sparse methods. In fact, good clustering results for the proposed algorithm can be obtained when choosing λ value in the range of $[0.05, 0.3]$ on all data sets.

Table I illustrates the detailed clustering results measured by accuracy using the Yale database without apparent outliers. The clustering results are conducted with the cluster number $K = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15\}$. Table II shows the detailed clustering results measured by Accuracy using the ORL database with apparent outliers. In this experiment, salt and pepper noise (the noise density is 0.05) is used to simulate outliers, and 5 percents of images are randomly selected to add this noise. The clustering results are conducted with the cluster number $K = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20\}$. From experiment results, we can observe that, compared with other related methods, the proposed l_1-CNMF algorithm can achieve the best accuracy performance for all the cases.

IV. CONCLUSION

In this paper, a robust sparse nonnegative matrix factorization algorithm, called l_1-norm nonnegative matrix factorization based on maximum correntropy criterion (l_1-CNMF), has been proposed. The l_1-CNMF algorithm incorporates the l_1 sparsity constraint and maximum correntropy criterion into the traditional NMF technique to improve the clustering performance. Compared with other existing related algorithms, experiment results on two image databases have demonstrated the effectiveness of the proposed l_1-CNMF algorithm.

ACKNOWLEDGMENT

This work was partially supported by 973 Program (No. 2015CB351703) and National Natural Science Foundation of China (No. 61372152, 91648208).

REFERENCES

[1] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, John Wiley & Sons, 2009.

[2] http://www.uk.research.att.com/facedatabase.html.

[2] C. Deng, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 6, pp. 902-913, 2011.

[3] R. Zhi, M. Flierl, Q. Ruan, and W. Bastiaan Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38-52, 2011.

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

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

[6] R. Sandler, M. Lindenbaum, "Nonnegative matrix factorization with earth mover's distance metric for image analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1590-1602, 2011.

[7] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l_{21}-norm," *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 673-682, 2011.

[8] Y. Wang, S. Wu, B. Mao, X. Zhang, and Z. Luo, "Correntropy induced metric based graph regularized non-negative matrix factorization," *Neurocomputing*, vol. 204, no. 2, pp. 172-182, 2016.

[9] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, Springer: New York, NY, USA, 2010.

[10] B. Chen, Y. Zhu, J. Hu, and J. C. Principe, *System Parameter Identification: Information Criteria and Algorithms*, Newnes, 2013.

[11] S. Peng, B. Chen, L. Sun, W. Ser, and Z. Lin, "Constrained maximum correntropy adaptive filtering," *Signal Processing*, vol. 140, pp. 116-126, 2017.

[12] J. Wang, X. Wang, and X. Gao, "Non-negative matrix factorization by maximizing correntropy for cancer clustering," *BMC bioinformatics*, vol. 14, no. 1, pp. 107-118, 2013.

[13] L. Du, X. Li, and Y. Shen, "Robust nonnegative matrix factorization via half-quadratic minimization," *Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE*, pp. 201-210, 2012.

[14] Y. Wang, C. Pan, S. Xiang, and F. Zhu, "Robust hyperspectral unmixing with correntropy-based metric," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4027-4040, 2015.

[15] S. Minaee, Y. Wang, "Subspace learning in the presence of sparse structured outliers and noise," *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*, pp. 1-4, 2017.

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

[17] R. He, B. Hu, X. Yuan, and L. Wang, *Robust recognition via information theoretic learning*, Springer, 2014.

[18] D. P. Bertsekas, *Nonlinear programming*, Belmont: Athena scientific, 1999.

[19] M. Kaneko, "KKT-condition inspired solution of DVFS with limited number of voltage levels," *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*, pp. 1-4, 2017.

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

[21] D. J. MacKay, *Information theory, inference and learning algorithms*, Cambridge university press, 2003.

[22] L. Lovasz, M. Plummer, *Matching Theory*, North Holland, 1986.