Robust sparse nonnegative matrix factorization based on maximum correntropy criterion

Siyuan Peng, Wee Ser and Zhiping Lin
School of Electrical and Electronic Engineering
Nanyang 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 (MCC) [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]

$$V(A, B) = E[\kappa(A, B)] = \int \kappa(a, b)dF_{AB}(a, b)$$  (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 $\sigma$ denotes the kernel bandwidth.

In practical applications, the joint 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:

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

(2)

### B. $l_1$-CNMF Algorithm

Assume that there is a nonnegative matrix $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}_{+}^{M \times N}$, where each column in $X$ denotes a sample vector containing $M$ elements. NMF aims to factorize $X$ into two nonnegative matrices $W \in \mathbb{R}_{+}^{M \times K}$ and $H \in \mathbb{R}_{+}^{K \times N}$ (called basis matrix and coefficient matrix respectively), such that the produce of $W$ and $H$ closely approximates the original matrix $X$:

$$X \approx WH$$

(3)

In order 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]

$$D_{ED} = \sum_{m=1}^M \sum_{n=1}^N \left( x_{mn} - \sum_{k=1}^K W_{mk} H_{kn} \right)^2$$

(4)

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

(5)

where $x_{mn}$, $W_{mk}$, and $H_{kn}$ denote the element in $X$, $W$, and $H$, 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 $X$ and the produce of $W$ and $H$. 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 $H$ 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:

$$D_{l_1-CNMF} = \frac{1}{MN} \sum_{m=1}^M \sum_{n=1}^N \left( \kappa_\sigma(x_{mn}, \sum_{k=1}^K W_{mk} H_{kn}) \right)$$

$$- 2\lambda \sum_{n=1}^N \| H_n \|_1$$

s.t. $W_{mk} \geq 0, H_{kn} \geq 0 \forall m, k, n$

(6)

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

$$\min_{W, H} D_{l_1-CNMF} = \sum_{m=1}^M \sum_{n=1}^N \left( -\kappa_\sigma(x_{mn}, \sum_{k=1}^K W_{mk} H_{kn}) \right)$$

$$+ 2\lambda \sum_{n=1}^N \| H_n \|_1$$

s.t. $W_{mk} \geq 0, H_{kn} \geq 0, \forall m, k, n$

(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 $W$ and $H$ 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 $\varphi$ of $g(x)$ such that

$$g(x) = \max_u (ux - \varphi(u)),$$

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

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

$$\min_{W, H, U} \hat{D}_{l_1-CNMF}$$

s.t. $W_{mk} \geq 0, H_{kn} \geq 0, \forall m, k, n$

$$\hat{D}_{l_1-CNMF} = 2\lambda \sum_{n=1}^N \| H_n \|_1 +$$

$$\sum_{m=1}^M \sum_{n=1}^N \left( U_{mn} (x_{mn} - \sum_{k=1}^K W_{mk} H_{kn})^2 + \varphi(U_{mn}) \right)$$

(9)

where $U_{mn}$ denotes a element of the nonnegative matrix $U \in \mathbb{R}^{M \times N}$. 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 \( W \) and \( H \) are fixed, we derive
\[
U_{mn} = \kappa \sigma (X_{mn}, \sum_{k=1}^{K} W_{mk} H_{kn}) = \exp \left( -\frac{(X_{mn} - \sum_{k=1}^{K} W_{mk} H_{kn})^2}{2\sigma^2} \right)
\]
where the kernel bandwidth is computed by \([12], [13]\)
\[
\sigma^2 = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} (X_{mn} - \sum_{k=1}^{K} W_{mk} H_{kn})^2
\]
\[
(11)
\]

2) **Computation of W**: When \( U \) is fixed, minimizing \( \hat{D}_{l_1-CNMF} \) with respect to \( W \) is equivalent to minimizing
\[
L(W) = \sum_{m=1}^{M} (X_{m*} - W_{m*}H) \text{diag}(U_{m*}) (X_{m*} - W_{m*}H)^T + 2\lambda \sum_{n=1}^{N} \| H_n \|_1 + Tr(\Phi W)
\]
\[
(12)
\]
where \( X_{m*} \) and \( W_{m*} \) denote the \( m \)th row in \( X \) and \( W \) respectively, \( \text{diag}(\cdot) \) 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 \( \Phi = \{ \varphi_{mk} \} \in \mathbb{R}^{MK \times K} \) is the Lagrange multiplier for the nonnegative constraint \( W_{mk} \geq 0 \). Similar to [13], using the partial derivative of \( L(W) \) respect to \( W \) and the Karush-Kuhn-Tucker condition (i.e., \( \varphi_{mk} W_{mk} = 0 \)) [18], [19], we derive the following update rule
\[
W_{mk} = W_{mk} \frac{(U \odot XH^T)_{mk}}{(U \odot (WH)H^T)_{mk}}
\]
\[
(13)
\]
where \( \odot \) denotes the Hadamard product.

3) **Computation of H**: When \( U \) is fixed, minimizing \( \hat{D}_{l_1-CNMF} \) with respect to \( H \) is equivalent to
\[
L(H) = \sum_{n=1}^{N} (X_{n*} - WH_{n*})^T \text{diag}(U_{n*}) (X_{n*} - WH_{n*}) + 2\lambda \sum_{n=1}^{N} \| H_n \|_1 + Tr(\Psi H)
\]
\[
(14)
\]
where \( X_{n*} \) and \( H_{n*} \) denote the \( n \)th column in \( X \) and \( H \) respectively, and \( \Psi = \{ \psi_{kn} \} \in \mathbb{R}^{KN \times N} \) is the Lagrange multiplier for the nonnegative constraint \( H_{kn} \geq 0 \). Then, the partial derivative of \( L(H) \) with respect to \( H \) is
\[
\frac{\partial L(H)}{\partial H_{kn}} = -2(W^T \text{diag}(U_{n*})X_{n*})_k + 2(W^T WH_{n*})_k + 2\lambda(\psi_{kn}) + \Psi_{kn}
\]
\[
(15)
\]
Similarly, using the Karush-Kuhn-Tucker condition (i.e., \( \psi_{kn} H_{kn} = 0 \)), we obtain
\[
(-2(W^T \text{diag}(U_{n*})X_{n*})_k + 2(W^T WH_{n*})_k + 2\lambda(\psi_{kn}) + \Psi_{kn}) H_{kn} = 0
\]
\[
(16)
\]
After some straightforward manipulations, we have the following update rule
\[
H_{kn} = H_{kn} \frac{(W^T(U \odot X))_{kn}}{(W^T(U \odot (WH)))_{kn} + \lambda}
\]
\[
(17)
\]
It is worth noting that, if \( W \) and \( H \) are the solution for the CNMF-\( l^1 \) algorithm, then \( WY \) and \( YH \) are also a solution for any positive diagonal matrix \( Y \) due to \( WH = (W^T)^{-1}(YH) \), which will lead \( H \) to be zero. To eliminate this problem in many practical areas, one strategy is frequently used to normalize each column of \( W \) to be 1 [14], [20]. Then we have
\[
W_{mk} \leftarrow \frac{W_{mk}}{\sqrt{\sum_{m=1}^{M} W_{mk}^2}}, H_{kn} \leftarrow H_{kn} \sqrt{\sum_{m=1}^{M} W_{mk}^2}
\]
\[
(18)
\]
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 \( X \in \mathbb{R}^{M \times N} \), the initial matrices \( W \in \mathbb{R}^{M \times K} \), \( H \in \mathbb{R}^{K \times N} \), and the regularization parameter \( \lambda \).

**Output**: \( W, H \)

**Computation**: repeat

1) Update \( U \) by using (10);
2) Update \( W \) by using (13);
3) Update \( H \) 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 (MCC-NMF) [12], and correntropy 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 \( X \) as a whole, while \( l_1 \)-CENMF considers entire row in \( X \) as a whole. In addition, for all NMF based methods, the matrices \( W \) and \( H \) 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.
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


