ROBUST NONNEGATIVE MATRIX FACTORIZATION VIA $L_1$ NORM REGULARIZATION
BY MULTIPLICATIVE UPDATING RULES

Bin Shen\*  Bao-Di Liu\°  Qifan Wang\*  Rongrong Ji\†

\*Computer Science, Purdue University, West Lafayette, IN. 47907, U.S.
\° Information and Control Engineering, China University of Petroleum, Qingdao, 266580, China
\†Cognitive Science, Xiamen University, Xiamen, 361005, China

ABSTRACT

Nonnegative Matrix Factorization (NMF) is a widely used technique in many applications such as face recognition, motion segmentation, etc. It approximates the nonnegative data in an original high dimensional space with a linear representation in a low dimensional space by using the product of two nonnegative matrices. In many applications data are often partially corrupted with large additive noise. When the positions of noise are known, some existing variants of NMF can be applied by treating these corrupted entries as missing values. However, the positions are often unknown in many real world applications, which prevents the usage of traditional NMF or other existing variants of NMF. This paper proposes a Robust Nonnegative Matrix Factorization (RobustNMF) algorithm that explicitly models the partial corruption as large additive noise without requiring the information of positions of noise. In particular, the proposed method jointly approximates the clean data matrix with the product of two nonnegative matrices and estimates the positions and values of outliers/noise. An efficient iterative optimization algorithm with a solid theoretical justification has been proposed to learn the desired matrix factorization. Experimental results demonstrate the advantages of the proposed algorithm.

Index Terms— Nonnegative matrix factorization, sparse noise, 11 norm regularizer.

1. INTRODUCTION

Dictionary learning is important for many computer vision tasks [1, 2, 3, 4, 5, 6, 7]. Nonnegative Matrix Factorization (NMF) is one of the most popular approaches, and it has been widely applied in a lot of applications such as face recognition [8], motion segmentation [9], object tracking [10], etc. NMF has received substantial attention due to its theoretical interpretation and practical performance.

Several variants of NMF have been proposed recently to improve the performance. Sparseness constraints have been incorporated into NMF to obtain sparse solutions [11, 12]. NMF algorithms in [13, 14] are proposed to preserve the local structure on the low dimensional manifold(s). To be robust to outliers, [15] proposes RSNMF, which is based on an outlier resistant objective function. [16] maintains an outlier list in NMF for more robust performance.

In real applications, data samples are often partially corrupted (e.g., pepper and salt noise in images, occlusion on faces). Unfortunately, traditional methods based on least square estimation, such as NMF and PCA, are sensitive to this kind of noise [17], since the underlying assumption of Gaussian noise distribution is not valid. Some recent work [18, 19, 20] tries to deal with partial corruption. They usually assume the positions of the corruption are given ahead, and then ignore the corresponding data entries. However, it is unrealistic to assume that the positions of corruption are known in many real world applications. [21] proposes Robust PCA to recover the noise value and position.

This paper proposes a Robust Nonnegative Matrix Factorization(RobustNMF) approach, which is able to simultaneously learn the basis matrix, coefficient matrix and estimate the positions and values of noise. The underlying observation is that the clean data allow a nonnegative factorization and the noise is sparse. An efficient iterative multiplicative optimization algorithm with solid theoretical justification has been proposed to obtain the desired solution of the RobustNMF approach.

The rest part of this paper is organized as follows. Section 2 proposes the RobustNMF algorithm, followed by the iterative optimization method in section 3. Section 4 provides some theoretical justification of the optimization method, and the experimental results are shown in section 5. Finally, we conclude and discuss this work.

2. ROBUST NONNEGATIVE MATRIX FACTORIZATION

The proposed Robust Nonnegative Matrix Factorization(RobustNMF) algorithm explicitly models the partial corruption, which is treated as large additive noise. Let nonnegative matrix $X \in \mathbb{R}^{m \times n}$ denote the observed corrupted data, while each column of $X$ is a data sample. Let $\hat{X} \in \mathbb{R}^{m \times n}$ denote the clean data without pollution. We have $X = \hat{X} + E$, where $E \in \mathbb{R}^{m \times n}$ is the large additive noise. Note that the large additive noise $E$ is not Gaussian noise with zero mean, which is well handled by least square error minimization. Moreover, we are concerned with partial corruption, and partial means the noise distribution is sparse. In other words, only a small portion of entries of $E$ are nonzero. For example, in face recognition, the occlusion by glasses is an instance of this kind of noise, and it covers only a small portion of the entire face.

The clean data $\hat{X}$ is approximated by $UV (U \in \mathbb{R}^{m \times k}, V \in \mathbb{R}^{k \times n})$ as in traditional NMF, thus we have

$$X \approx UV + E \quad (1)$$

Considering the above model and the sparseness of the large additive noise $E$, the objective function of RobustNMF is defined as follows:

$$O_{RobustNMF} = ||X - UV - E||_F^2 + \lambda \sum_j ||E_j||_0^2 \quad (2)$$

where $\lambda$ is a parameter that controls the degree of sparsity.
The first term is to approximate the clean data; the second term is obtained from the sparseness constraint of E. The parameter λ controls the tradeoff between the two terms, thus it is dependent on how large portion of entries are corrupted.

However, the L0 norm in the second term makes this objective function difficult to optimize, so L1 norm is employed to approximate it, which has been a popular strategy in prior research [22]. Substituting the L1 norm into the objective function, we have

\[
O_{\text{RobustNMF}} = \|X - UV - E\|_F^2 + \lambda \sum_j \left(\|E_j\|_1 + ||E_j||_1\right)^2
\]

where E = E^p - E^n, E^p = \frac{|E| + E}{2}, E^n = \frac{|E| - E}{2}, and E^p \geq 0, E^n \geq 0. Now we have squared L1 norm penalty for sparseness, which has been proved to be effective and computationally convenient [12, 23, 14]. Note that E is the sparse large additive noise, which could be either negative or nonnegative. We need to decompose E into two nonnegative matrices E^p and E^n described above to gain the nonnegativity which results in the convenience in optimization. We also set constraint that X - E \geq 0, since the clean data should be nonnegative. Finally, the objective function should be minimized with respect to U, V, E^p, and E^n subject to the constraints that U \geq 0, V \geq 0, E^p \geq 0, E^n \geq 0, and X - E \geq 0.

3. MULTIPLICATIVE OPTIMIZATION

Since O_{\text{RobustNMF}} is not convex with U, V, E^p, and E^n jointly, it is difficult to find the global minimum for O_{\text{RobustNMF}}. Instead, we aim to find a local minimum by iteratively updating U, V, E^p and E^n in a similar way with the work [24] for NMF. Note, we are still able to optimize w.r.t. V, E^p and E^n jointly.

3.1. Update U

Given V, E^p, and E^n, we update U to decrease the value of objective function.

\[
U = \arg \min_{U \geq 0} \|X - [U, I, -I] \left( V_{(E^n)} \right)\|_F^2 + \lambda \sum_j \left(\||E|^j\|_1 + ||E_n|^j||_1\right)^2
\]

\[
= \arg \min_{U \geq 0} \|X - UV - E\|_F^2
\]

The updating rule for U to reduce the objective function is as follows, which can be proven in a similar way as in [24].

\[
U_{ij} = U_{ij} \left( \hat{X}V^T \right)_{ij} / (UVV^T)_{ij}
\]

where \( \hat{X} = X - E \). Note that at this step E is given, and it satisfies the constraint that X - E \geq 0.

3.2. Update V, E^p, and E^n Simultaneously

Now we decrease the objective function with respect to V, E^p and E^n given U. Let \( \bar{V} = \left( V_{(E^p)} \right). \)

The updating rule for \( \bar{V} \) is:

\[
\bar{V}_{ij} = \max(0, \bar{V}_{ij} - \frac{\bar{V}_{ij}(U^T \bar{V} \bar{V})_{ij}}{(SV)_{ij}} + \frac{\bar{V}_{ij}(U^T \tilde{X})_{ij}}{(SV)_{ij}})
\]

where \( \tilde{X} = (X_{(01x^n)}) \), \( \bar{V} = (U_{01x^n} \sqrt{X_{1x1m} \sqrt{X_{1x1m}}}) \), and S is defined as

\[
S_{ij} = \|U^T \tilde{V} \|_F^2
\]

4. CORRECTNESS OF UPDATING RULES

To decrease the objective function with respect to V, E^p and E^n. We have:

\[
(V, E^p, E^n) = \arg \min_{V, E^p, E^n \geq 0} O_{\text{RobustNMF}}
\]

\[
= \arg \min_{V, E^p, E^n \geq 0} \|X - [U, I, -I] \left( V_{(E^n)} \right)\|_F^2 + \lambda \sum_j \left(\||E|^j\|_1 + ||E_n|^j||_1\right)^2
\]

\[
= \arg \min_{V, E^p, E^n \geq 0} \|X - UV - E\|_F^2
\]

\[
= \arg \min_{V, E^p, E^n \geq 0} \|X - UV - E\|_F^2
\]

where \( \tilde{X} = (X_{(01x^n)}) \), \( \bar{V} = (U_{01x^n} \sqrt{X_{1x1m} \sqrt{X_{1x1m}}}) \), \( \tilde{V} = \left( V_{(E^p)} \right). \)

Updating V, E^p and E^n is more involved than updating U, since \( \tilde{U} \) contains some negative values. Now we prove the correctness of the updating rules for V, E^p and E^n proposed in section 4.

4.1. Decrease Objective Function

Definition 1 [24] Z(\( \tilde{V}, \tilde{V}^n \)) is an auxiliary function for F(\( \tilde{V} \)), if it satisfies the following conditions

Z(\( \tilde{V}, \tilde{V}^n \)) \geq F(\( \tilde{V} \)), Z(\( \tilde{V}, \tilde{V}^n \)) = F(\( \tilde{V} \))

Lemma 1 [24] If Z is an auxiliary function, then F is nonincreasing under the update

\( \tilde{V}^{n+1} = \arg \min_{\tilde{V}} Z(\( \tilde{V}, \tilde{V}^n \)) \)

Now we generalize the Lemma 1 to Lemma 2.

Lemma 2 If Z is an auxiliary function, then F is nonincreasing as long as \( \tilde{V}^{n+1} \) satisfies the following condition:

Z(\( \tilde{V}^{n+1}, \tilde{V} \)) \leq Z(\( \tilde{V}, \tilde{V} \))

Proof:

F(\( \tilde{V}^{n+1} \)) \leq Z(\( \tilde{V}^{n+1}, \tilde{V} \)) \leq Z(\( \tilde{V}, \tilde{V} \)) \leq F(\( \tilde{V} \)) \square

This generalization from Lemma 1 to Lemma 2 is similar to the generalization from EM to Generalized EM.
In our problem, $\bar{U}$ contains some negative value. Thus the updating rules in [24] do not hold. So we begin to seek new updating rules.

**Lemma 3** If $K(\tilde{v}^t)$ is the diagonal matrix that

$$K_{ab}(\tilde{v}^t) = \delta_{ab}(S\tilde{v}^t)_a/\tilde{v}^t_a$$

then

$$Z(\tilde{v}, \tilde{v}^t) = F(\tilde{v}^t) + (\tilde{v} - \tilde{v}^t)\nabla F(\tilde{v}^t) + \frac{1}{2}(\tilde{v} - \tilde{v}^t)^T K(\tilde{v}^t)(\tilde{v} - \tilde{v}^t)$$

is an auxiliary function for

$$F(\tilde{v}) = \frac{1}{2} \sum_i (\tilde{x}_i - \sum_a \tilde{U}_{ia}\tilde{v}_a)^2$$

**Proof:**

$Z(\tilde{v}, \tilde{v}^t) = F(\tilde{v})$, obviously. Now we prove that $Z(\tilde{v}, \tilde{v}^t) \geq F(\tilde{v})$.

Comparing

$$F(\tilde{v}) = F(\tilde{v}^t) + (\tilde{v} - \tilde{v}^t)\nabla F(\tilde{v}^t) + \frac{1}{2}(\tilde{v} - \tilde{v}^t)^T (U^T \bar{U})(\tilde{v} - \tilde{v}^t)$$

with $Z(\tilde{v}, \tilde{v}^t)$, we find that we only need to show

$$(\tilde{v} - \tilde{v}^t)^T [K(\tilde{v}^t) - U^T \bar{U}](\tilde{v} - \tilde{v}^t) \geq 0$$

To prove the positive semidefiniteness, consider the matrix $M(\tilde{v}^t)$ with $M_{ab}(\tilde{v}^t) = \tilde{v}^t_a [K(\tilde{v}^t) - U^T \bar{U}]_{ab} \tilde{v}^t_b$. $M$ is a rescaling of $K(\tilde{v}^t) - U^T \bar{U}$. It is semipositive definite if and only if $K(\tilde{v}^t) - U^T \bar{U}$ is.

$$\mu^T M \mu = \sum_{ab} \mu_a M_{ab} \mu_b$$

$$= \sum_{ab} \tilde{v}^t_a S_{ab} \tilde{v}^t_b \mu_a^2 - \mu_a \tilde{v}^t_a (U^T \bar{U})_{ab} \tilde{v}^t_b \mu_b$$

$$= \sum_{ab} \frac{1}{2} \mu_a^2 + \frac{1}{2} \mu_b^2 - \text{sgn}((U^T \bar{U})_{ab}) \mu_a \mu_b$$

$$= \sum_{ab} \frac{1}{2} \mu_a^2 - \text{sgn}((U^T \bar{U})_{ab}) \mu_a \mu_b$$

$$\geq 0$$

where $\text{sgn}(x) = \begin{cases} -1: & x < 0 \\ 0: & x = 0 \\ 1: & x > 0 \end{cases}$

Note in our setting, $\bar{U}$ contains some negative values, but $\tilde{V}$ is nonnegative. □

Substitute Lemma 3 into Lemma 1, the updating rule is:

$$\tilde{v}^{t+1} = \tilde{v}^t - K(\tilde{v}^t)^{-1}\nabla F(\tilde{v}^t)$$

$$= \tilde{v}^t - K(\tilde{v}^t)^{-1}(U^T \bar{U})\tilde{v}^t + K(\tilde{v}^t)^{-1}U^T \bar{x}$$

Writing the components explicitly, we get:

$$\tilde{v}_{a}^{t+1} = \tilde{v}_{a}^{t} - \tilde{v}_{a}^{t}(U^T \bar{U})_{aa} / (S\tilde{v}^t)_a + \tilde{v}_{a}^{t}(U^T \bar{x})_a / (S\tilde{v}^t)_a$$

The proposed updating rules can deal with negative values by explicitly considering the negative part of large additive noise in the $\bar{U}$. If $\bar{U} \geq 0$, the first two terms in the above updating rule would cancel each other, resulting in the same rule as in [24].

The $\tilde{V}$ gained by equation 16 is made up of three parts: $V, E^p$, and $E^o$. All of them should be nonnegative. Unfortunately, the value $\tilde{v}^{t+1}$ gained by equation 16 does not guarantee the nonnegativity. Now we discuss how to keep it nonnegative while updating the values.

In the auxiliary function $Z$, we see that the $K(\tilde{v}^t)$ is a diagonal matrix. Thus the second order terms only involve the form $\tilde{v}^a$. This results in a very important property in Lemma 4.

**Lemma 4** If $\tilde{v}^{t+1} = \arg \min_{\tilde{v}} Z(\tilde{v}, \tilde{v}^t)$ and $\tilde{v}^t \geq \tilde{v}^{t+1} \geq 0$,

then $Z(\tilde{v}^{t+1}, \tilde{v}^t) \leq Z(\tilde{v}, \tilde{v}^t)$.

According to Lemma 2 and Lemma 4, we have $F(\tilde{v}^{t+1}) \leq F(\tilde{v}^t)$.

So, to ensure the nonnegativity, we can simply threshold $\tilde{v}^{t+1}$ by 0. This operation will introduce the nonnegativity, while keeping the value of $F$ nonincrease from $\tilde{v}^t$ to the thresholded $\tilde{v}^{t+1}$. Thus, we have the updating rule described in equation 6.

### 4.2. Convergence Analysis

Since the objective function has a lower bound, e.g., 0, and the updating rules for $U, V$, and $E$ will all cause the objective function nonincrease, the algorithm always converges.

### 5. EXPERIMENTAL RESULTS

This section presents experimental results on two different applications of face reconstruction and image denoising.

#### 5.1. Reconstruction of Faces

The experiments are based on the ORL face dataset. Each face image is of size 32×32, thus is represented by a 1024 dimensional vector. For each face in a randomly selected subset, 50 pixels are randomly selected and replaced with the values of 255 to simulate the large additive noise. The polluted faces make up the data matrix $X$, each column of which corresponds to a polluted face image, while the original data samples form the matrix $\tilde{X}$. Mean Squared Error(MSRE) is used to measure the reconstruction performance.

For NMF, matrices $U$ and $V$ are learned based on $X$, and then are used in reconstruction. Compared with the original noise free matrix $\tilde{X}$, the MSRE is calculated as $\frac{1}{N}||\tilde{X} - UV||^2_F$, where $N$ is the number of samples. For RobustNMF, the MSRE definition is the same as for NMF. For RobustNMF+WNMF, based on $X$, RobustNMF learns $U$, $V$, and $E$. Since $E$ is an indicator for whether a pixel is polluted or not. Taking $E$ as a mask, WNMF learns the new matrix $U$, $\tilde{V}$. The MSRE is defined as $\frac{1}{N}||\tilde{X} - U\tilde{V}||^2_F$.

Experiments are conducted with varying number of pixels polluted and faces. In the first set of experiments, we fix the number of faces to be 50 or 100, and then vary the number of pixels polluted, from 10 to 100 with a step of 10. This means that there are about 1 percent to 10 percent pixels corrupted in each face. The results are...
shown in top row of figure 2. In the second set of experiments, the number of pixels polluted is fixed at 50 or 100, which means about 5 or 10 percent of pixels on each face are corrupted. Experiments are conducted with various numbers of faces, from 10 to 100 with a step of 10. The results are shown in bottom row of figure 2.

It can be seen from these experiments that both RobustNMF and RobustNMF+WNMF consistently outperform the traditional NMF with varying number of data samples. With the increasing amount of noise, the advantages of proposed algorithms become even larger. This is because RobustNMF is able to detect the positions of the large value noises, i.e. the partial corruption, which enables the application of WNMF. Considering the approximation of $L_0$ norm by $L_1$ norm, the large noise is underestimated, and that is why we prefer Robust+WNMF to pure RobustNMF, even though both methods outperform the traditional NMF.

5.2. Image Denoising(Reconstruction of Patches)

This subsection presents some experiment results on image denoising by using RobustNMF. Pepper and salt noise is added to natural images. The noise density is set to 5%, which means about 5% of pixels are affected. The noisy image is converted into a set of patches, to which RobustNMF is applied. $\lambda$ is set to 0.04 and $k$ is set to 10. $UV$ is used to reconstruct the original image. Some denoising results are shown in figure 1. The first row shows the generated polluted images, the second row shows the denoised results by traditional NMF, and the third row is the results by RobustNMF. It can be seen that RobustNMF outperforms traditional NMF.

6. CONCLUSION

Data in many real world applications are often partially corrupted without the explicit information of positions of noise, which prevents the usage of NMF and other existing variants. This paper proposes a RobustNMF algorithm, which is able to simultaneously locate and estimate the corruption and learn the basis matrix $U$ and coefficient matrix $V$ in the framework of NMF. This proposed algorithm also paves the way to apply other variants of NMF(e.g. WNMF) to data with missing values by estimating the positions of noise. An efficient multiplicative optimization algorithm with a solid theoretical justification is designed for RobustNMF, and experimental results demonstrate the advantages of our algorithm in applications of image denoising and face reconstruction.

Acknowledgements Bao-Di Liu is supported by the Fundamental Research Funds for the Central Universities (No.R1405012A). Rongrong Ji is supported by the Nature Science Foundation of China (No. 61373076), the Fundamental Research Funds for the Central Universities (No.2013121026), the 985 Project of Xiamen University.
7. REFERENCES

[1] Liangyue Li, Sheng Li, and Yun Fu, “Discriminative dictionary learning with low-rank regularization for face recognition,” in Automatic Face and Gesture Recognition, 10th IEEE International Conference and Workshops on, April 2013, pp. 1–6.


