Robust Non-negative Matrix Factorization via Joint Sparse and Graph Regularization

Shizhun Yang, Chenping Hou, Member, IEEE, Changshui Zhang, Member, IEEE, Yi Wu, Shifeng Weng

Abstract—In real world applications, we often have to deal with some high-dimensional, sparse and noisy data. In this paper, we aim to handle this kind of complex data by a Robust Non-negative Matrix Factorization via joint Sparse and Graph regularization model (RSGNMF). We provide a novel efficient and elegant iterative updating algorithm with rigorous convergence analysis for RSGNMF model. Experimental results on image data sets demonstrate that our RSGNMF model outperforms existing start-of-art methods.

I. INTRODUCTION

In real world applications, data have some new challenges, such as high-dimensionality, sparsity, containing noises and outliers, etc. These challenges motivate us to develop new technology to deal with this kind of complex data.

The high dimensionality of data poses challenges such as the curse of dimensionality [1]. A common way to solve this problem is dimension reduction. Non-negative matrix factorization technique (NMF), which originally proposed as a method for finding matrix factors with parts-of-whole interpretations [2][3], has been successfully applied to dimension reduction areas in the past decades [4][5][6].

When data contain noises or outliers, the error for both features and samples are squared, thus a few noisy features or a few outliers with large errors will dominate objective function. Robust NMF (RNMF) version was proposed to handle high-dimensional and noisy data [7][8]. In order to alleviate the impact of noises or outliers, we also employ RNMF technique to handle our concerned problems. Sometimes, data are sparse, and then sparse regularization technique is popular due to some important benefits, such as simple and robust; can avoid over-fitting problems; can discover the most relevant samples or features, etc. Motivated by recent interesting works of sparse NMF (SNMF), we also introduce sparse regularization in our models to learn meaningful features or samples.

II. RELATION TO PRIOR WORK

The most related work is Non-negative Matrix Factorization (NMF), which has gained great success in real applications. Cai et al. [4][5] proposed graph regularized non-negative matrix factorization (GNMF) model to preserve geometrical information by constructing nearest neighbor graph. Afterwards, there are some variations of GNMF, such as Constrained NMF [5], Dual Regularized Co-Clustering [6]. When real world data contain noises or outliers, some robust methods were proposed to handle high-dimensional and noisy data. Zhang et al. [7] proposed a robust non-negative matrix factorization model (RNMF) with $L_1$-norm regularization. Kong et al. [8] proposed a robust formulation of NMF using $L_{2,1}$-norm loss function (Robust NMF).

GNMF models only considered local structures of data, which can’t handle noisy or sparse data. RNMF model with $L_1$-norm regularization does not guarantee all data are sparse in the same features. Neither RNMF model nor Robust NMF model takes local structures of data into consideration.

In our paper, we take all these factors into consideration. We not only employ graph regularization to preserve local structures of data, but also employ a more suitable regularization term to make sure our model can handle noisy and sparse data simultaneously.
III. ROBUST NMF VIA JOINT SPARSE AND
GRAPH REGULARIZATION MODEL (RSGNMF)

A. RSGNMF model

To handle high-dimensional, sparse and noisy data simultaneously, we propose a novel Robust NMF via joint Sparse and Graph regularization model (RSGNMF).

We employ NMF technique to handle high-dimensional data. Given a data matrix $X$, each column of $X$ is a sample vector. NMF aims to find two non-negative matrices $F$ and $G$ whose product can well approximate original matrix $X$, i.e., $X \approx FG$. NMF can be formulated as

$$
\min_{F \geq 0, G \geq 0} J(X \| FG) = \| X - FG \|_F^2
$$

where $J(X \| FG)$ is some divergence function that measures dissimilarity between $X$ and $FG$. There are a variety of divergence functions with corresponding multiplicative update rules [9]. In this paper, we focus on the simple Least Squares Error divergence function as follows,

$$
\min_{F \geq 0, G \geq 0} J(X \| FG) = \| X - FG \|_F^2
$$

with the corresponding multiplicative update rules,

$$
G_{ki} \leftarrow G_{ki} \frac{(F^T X)_{ki}}{(F^T FG)_{ki}} (3)
$$

$$
F_{jk} \leftarrow F_{jk} \frac{(XG^T)_{jk}}{(FGG^T)_{jk}} (4)
$$

Motivated by the great success of GNMF methods, we employ graph regularization term into objective function to preserve local structures in new feature space as follows,

$$
\min_{F \geq 0, G \geq 0} J(X \| FG) = \| X - FG \|_F^2 + \lambda \text{tr}(GLG^T) (5)
$$

where $L$ is graph Laplacian, $L = D - W$, $W$ is weight matrix of the constructed nearest neighbor graph, $D$ is diagonal matrix whose entries are sums of $W$.

When real world data contain noises or outliers, especially some non-Gaussian noises exist, the error for both features and samples are squared, thus a few noisy features or outliers with large errors will dominate objective function. Then traditional NMF or GNMF can’t handle these kinds of non-Gaussian noises effectively.

We employ joint $L_{2,1}$-norm regularization into objective function to handle noisy and sparse data effectively. $L_{2,1}$-norm loss function based on NMF enables our model to effective handle non-Gaussian noises or outliers and $L_{2,1}$-norm regularization enables our model to effective handle sparse data.

We give a view of $L_{2,1}$-norm [10][11][12] as follows. For a matrix, $L_{2,1}$-norm is defined as,

$$
\| W \|_{2,1} = \sum_{i=1}^m \sqrt{\sum_{j=1}^d w_{ij}^2} = \sum_{i=1}^m \| w_i \|_2. (6)
$$

where $w_{ij}$ is $i$-th row of $W$. We give an intuitive explanation of $L_{2,1}$-norm. First, we compute $L_2$-norm of rows $w_i$, and then compute $L_1$-norm of vector $b(W) = (|w_1|_2, |w_2|_2, \ldots, |w_m|_2)$. The magnitudes of components of vector $b(W)$ indicate how important each dimension is. $L_{2,1}$-norm favors a small numbers of nonzero rows in $W$, thereby ensuring dimension reduction will be achieved.

In summary, given a data matrix $X$, we can get the final formulation of RSGNMF model, which takes robust loss function, sparse regularization and local structures of data into consideration simultaneously,

$$
\min_{F \geq 0, G \geq 0} \| X - FG \|_{2,1} + \alpha \| G \|_{2,1} + \beta \text{tr}(GLG^T) (7)
$$

where $X \in R^{m \times n}$, $F \in R^{m \times p}$, $G \in R^{p \times n}$.

It is worthy to point why this formulation can make success in handling high-dimensional, noisy and sparse data simultaneously. There are four reasons as follows,

1. Real data may contain noises or outliers, while the first $L_{2,1}$-norm loss function term is designed to alleviate impact of noises or outliers. Then we can get more clear data to further process.

2. The clear data get in above procedure may be sparse, i.e., not all the features are important to learning procedure, while the second $L_{2,1}$-norm regularization term is designed to generate row sparsity to get final sparse solution.

3. After handling noisy and sparse data, we wish to preserve local structures of original data in new feature space, which is employed as third graph regularization term.

4. Our optimization algorithm iterative optimizes the above three procedure until convergence, which is benefit to process this kind of complex data effectively.

5. Experimental results demonstrate three constraints enable to build more effective models especially on high-dimensional, sparse and noisy data sets.

B. Solution for RSGNMF model

We present the solution for RSGNMF model via iterative updating algorithm below, and convergence and correctness of algorithm is given in next two sections.

$$
G_{ki} \leftarrow G_{ki} \frac{(F^T XD_1)_{ki}}{(F^T FGD_1 + \alpha GD_2 + 2\beta GLG^T)_{ki}} (8)
$$

$$
F_{jk} \leftarrow F_{jk} \frac{(XGD_1 G^T)_{jk}}{(FGGD_1 G^T)_{jk}} (9)
$$

where $D_1$ and $D_2$ are diagonal matrix with diagonal elements given by

$$
(D_{1i})_1 = \frac{1}{m} \sum_{j=1}^m (X - FG)_{ji} = 1/\|x_i - FG\|_1 (10)
$$

$$
(D_{1i})_2 = \frac{1}{m} \sum_{j=1}^m G_{ji}^2 = 1/\|g_i\|_2 (11)
$$
Then we get new parts-based representations of original data $X$, which contain much useful information of data. Graph regularization term makes the learning procedure preserving local structures of data. $L_{2,1}$-norm loss function make our model be robust to noises or outliers. $L_{2,1}$-norm regularization forces most of rows in $G$ shrink to zero, which implies the corresponding features of these zero rows are not important to new representations.

### C. Convergence and correctness analysis of the model

In order to prove convergence of the model, we should prove the following Theorem 1 and Theorem 2.

**Theorem 1:** Updating $G$ using rule of 8 while fixing $F$, the objective function of 7 monotonically decreases, i.e.,

$$
\|X - FG_{t+1}\|_{2,1} + \alpha\|G_{t+1}\|_{2,1} + \beta \text{tr}(G_{t+1}LGT_{t+1})
- \|X - FG_{t}\|_{2,1} - \alpha\|G_{t}\|_{2,1} - \beta \text{tr}(G_{t}LGT_{t}) \leq 0
$$

where $t$ is the iteration number of times.

**Theorem 2:** Updating $F$ using rule of 9 while fixing $G$, the objective function of 7 monotonically decreases, i.e.,

$$
\|X - F_{t+1}G\|_{2,1} + \|X - FG_{t}\|_{2,1} \leq 0
$$

where $t$ is the iteration number of times.

In order to prove Theorem 1, we need the following Lemma 1 and Lemma 2. In order to prove Theorem 2, similar with [8], we need the following Lemma 3 and Lemma 4.

**Lemma 1:** under updating rule of 8, inequation holds

$$
\text{tr}((X - FG_{t+1})D_{1}(X - FG_{t+1})^{T}) + \text{tr}(G_{t+1}(\alpha D_{2} + \beta L)G_{t+1}) \leq \text{tr}((X - FG_{t})D_{1}(X - FG_{t})^{T}) + \text{tr}(G_{t}(\alpha D_{2} + \beta L)G_{t})
$$

**Lemma 2:** under updating rule of 8, inequation holds

$$
\|X - FG_{t+1}\|_{2,1} + \alpha\|G_{t+1}\|_{2,1} + \beta \text{tr}(G_{t+1}LGT_{t+1})
- \|X - FG_{t}\|_{2,1} - \alpha\|G_{t}\|_{2,1} - \beta \text{tr}(G_{t}LGT_{t}) \leq \frac{1}{2}\text{tr}((X - FG_{t+1})D_{1}(X - FG_{t+1})^{T}) + \text{tr}(G_{t+1}(\alpha D_{2} + \beta L)G_{t+1}) - \text{tr}((X - FG_{t})D_{1}(X - FG_{t})^{T}) - \text{tr}(G_{t}(\alpha D_{2} + \beta L)G_{t})
$$

**Lemma 3:** under updating rule of 9, inequation holds

$$
\text{tr}(X - F_{t+1}G)D_{1}(X - F_{t+1}G)^{T} \leq \text{tr}(X - F_{t}G)D_{1}(X - F_{t}G)^{T}
$$

**Lemma 4:** under updating rule of 9, inequation holds

$$
\|X - F_{t+1}G\|_{2,1} - \|X - FG_{t}\|_{2,1} \leq \frac{1}{2}\text{tr}((X - F_{t+1}G)D_{1}(X - F_{t+1}G)^{T}) - \text{tr}((X - F_{t}G)D_{1}(X - F_{t}G)^{T})
$$

Now, we have proved convergence of our model. Next we should prove the converged solution is the correct optimal solution. We derive two Theorems to prove correctness of algorithms.

**Theorem 3:** The converged $G^*$ of the updating rule of 8 satisfies the KKT condition of the optimization theory.

**Theorem 4:** The converged $F^*$ of the updating rule of 9 satisfies the KKT condition of the optimization theory.

Due to space limitations, we skip detailed proof of Theorem 1-4 and Lemma 1-4.

## IV. EXPERIMENTAL RESULTS

### A. Data sets descriptions and comparison algorithms

We use four image data sets to evaluate effectiveness of RSGNMF method. USPS [13] contains 9298 images in 10 categories and MNIST [14] contains 10000 images in 10 categories. They are represented as raw pixel intensities with $28 \times 28$ pixel image. UMIST [15] consists of 564 images from 20 individuals and each image is cropped to $28 \times 23$ pixels. YALE [16] consists of 165 images from 15 individuals and we resize them to $32 \times 32$ pixel image.

We give a general view of comparison algorithms in Table 1. They can only handle one or more data characters separately, while our model can handle high-dimensional, sparse and noisy data simultaneously.

<table>
<thead>
<tr>
<th>Comparison algorithms</th>
<th>Dimension reduction</th>
<th>Sparse regularization</th>
<th>Graph regularization</th>
<th>Robust to noises</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>GNMF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RNMF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ERFS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RSGNMF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

We select four kinds of algorithms as the baselines.

1. K-means clustering. We directly perform k-means clustering without dimension reduction. By comparing this method, we want to know whether dimension reduction technique can improve clustering performance on high-dimensional data.

2. Graph regularized Non-negative Matrix Factorization (GNMF) [4]. GNMF takes the local structures of data into consideration. By comparing GNMF algorithm, we want to know whether the graph regularization term is necessary.

3. Robust Non-negative Matrix Factorization (RNMF) [8]. RNMF reformulated NMF as a robust version via $L_{2,1}$-norm loss function. We directly perform RNMF on data to verify whether robust loss function term can handle noises or outliers.

4. Efficient Robust Feature Selection (ERFS) [12]. ERFS is a robust feature selection method via joint $L_{2,1}$-norm minimization which considered not only robust loss function but also sparse structures of data. By comparing ERFS, we want to know whether the joint $L_{2,1}$-norm minimization term is necessary.

In re-achieving the results of the comparison algorithms, we set the parameters following the constructions described in their own papers. For example, for algorithm GNMF, which used graph regularization to preserve the local structures of data, need to construct a $k$-nearest neighbor graph. We set the distance metric as Euclidean distance, the value of $k$ as 5, the weight mode as a heat kernel in LPP.

The number of $k$ in our RSGNMF is an important parameter to discuss. Clearly, $k$ can be bigger or smaller than the dimensionality $d$ of original data. For simplicity like, we set $k = d$ in our methods.
B. Overall comparison results

In this paper, we perform three groups of experiments. First, we compare clustering results by the derived clustering Accuracy (ACC) and Normalized Mutual Information (NMI). Second, we test convergence property of our algorithm. Third, we study sensitivity of parameters.

In first group, the evaluations are conducted with different numbers of clusters varying from 2 to 10. Figure 1-4 give the results separately. Afterwards we fix the number of clusters as real number of classes in original data separately. Table 2-3 give clustering results separately. We run 20 repetitions and report means and standard deviations. The numbers showed in bold are the best results.

From Figure 1-4 and Table 2-3, we can get several observations.

1). K-means clustering method are the worst among five algorithms. This indicates that dimension reduction technique is necessary to improve clustering performance on handling high-dimensional data.

2). GNMF and RSGNMF algorithms, which take local structures into consideration, perform better than other algorithms. This indicates that preserving local structures via graph regularization in dimension reduction model is crucial.

3). ERFS and RSGNMF algorithms, which take robust loss function and sparse regularization term into consideration, perform better than other algorithms. This indicates the joint $L_{2,1}$-norm regularization is important in improving clustering performance especially on noisy and sparse data sets.

4). Our RSGTL method performs best among five algorithms on four data sets. This demonstrates the effectiveness of our model.

Above all, we can get a conclusion that more constraints enable to build more effective models especially on high-dimensional, sparse and noisy data sets.
In the second group, we use two data sets, i.e., USPS and YALE, to test convergence property of RSGNMF model. Figure 5 gives convergence curve of RSGNMF model.

![Convergence Curve](image)

**Fig. 5.** The convergence curve of RSGNMF model.

From Figure 5, we can see that our models converge fast and the numbers of iteration is less than 15.

In the third group, we use two data sets, i.e., MNIST and UMIST, to test parameter sensitivity of RSGNMF model. There are two parameters $\alpha$ and $\beta$ to select. If we set $\alpha$ larger, we can get larger numbers of zero rows in $G$, correspondingly $F$ can contain more meaningful sparse information. But if we set $\alpha$ too large, the second term will dominate objective function, most of rows in $G$ will be zeros, which is not benefit for clustering either. In a word, we should select suitable $\alpha$ to make clustering effectively. We first select them in a grid search and then change them in certain ranges. Figure 6 gives parameter sensitivity of RSGNMF model.

![Parameter Sensitivity](image)

**Fig. 6.** The parameter sensitivity of RSGNMF model.

From Figure 6, we can see that when parameters change within a certain range, the performances of RSGNMF model changes within a certain range.

V. CONCLUSION AND DISCUSSION

In this paper, we propose a robust NMF via joint sparse and graph regularization model to handle high-dimensional, sparse and noisy data simultaneously. Efficient iterative update algorithms with rigorous convergence and correctness analysis are also given. Experimental results demonstrate the effectiveness of our model. In the future, we plan to extend our RSGNMF framework to handle non i.i.d. data, i.e., the data follows different distributions.

REFERENCES