Robust Manifold Non-Negative Matrix Factorization

Jin Huang. University of Texas at Arlington
Feiping Nie. University of Texas at Arlington
Heng Huang. University of Texas at Arlington
Chris Ding. University of Texas at Arlington

Non-negative matrix factorization (NMF) has been one of the most widely used clustering techniques for exploratory data analysis. However, since each data point enters the objective function with squared residue error, a few outliers with large errors easily dominate the objective function. In this paper, we propose a Robust Manifold Non-negative Matrix Factorization (RMNMF) method using $\ell_{2,1}$-norm and integrating NMF and spectral clustering under the same clustering framework. We also point out the solution uniqueness issue for the existing NMF methods, and propose additional orthonormal constraint to address this problem. With the new constraint, the conventional auxiliary function approach no longer works. We tackle this difficult optimization problem via a novel Augmented Lagrangian Method (ALM) based algorithm and convert the original constrained optimization problem on one variable into a multivariate constrained problem. The new objective function then can be decomposed into several subproblems that each has a closed-form solution. More important, we reveal the connection of our method with robust $K$-means and spectral clustering, and demonstrate its theoretical significance. Extensive experiments have been conducted on nine benchmark data sets and all empirical results show the effectiveness of our method.

Categories and Subject Descriptors: H.2.8 [Database Applications]: Data Mining

General Terms: Theory, Algorithms, Performance

Additional Key Words and Phrases: Clustering, Non-negative Matrix Factorization (NMF), Manifold, Robust NMF

ACM Reference Format:

1. INTRODUCTION

The applications of matrix factorization in data mining and information retrieval can be traced back to about 50 years ago. Salton et al. used matrix to represent a collection of document in their famous Vector Space Model (VSM) [Salton et al. 1975]. In the model, $n$ textual documents are changed into $n$ document vectors $d_1, \ldots, d_n$ and a term-by-document matrix $A = [d_1 \ldots d_n] \in \mathbb{R}^{m \times n}$ is formed. Given a test pseudo-document $q$, a document $d_i$ from $A$ is found via calculating the angular cosine measure. To make improvement to VSM, Foltz and Dumais proposed Latent Semantic Indexing (LSI) model [Foltz and Dumais 1992]. Assuming $A$ is of rank $r$ and can be decomposed via SVD as follows: $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$. LSI proposes to use $A_k = \sum_{i=1}^{k} \sigma_i \mu_i v_i^T$ in place of $A$. There are two potential advantages: (1) This method reduces storage when $k \ll r$; (2) This method also filters out unnecessary information, so that the performance of text mining tasks (e.g., query processing and clustering) improves. However, the interpretation of basis vectors $\mu_i$ is difficult due to mixed signs. With realization to this drawback, Lee and Seung proposed Non-negative Matrix Factorization (NMF) [Lee and Seung 2001], where non-negative $A_k = U_k \sum_k V_k^T = W_k H_k$. Here...
columns of $W$ are the underlying basis vectors, i.e., each of the $n$ columns of $A$ can be built from $k$ columns of $W$. Columns of $H$ give the weights associated with each basis vector.

The scope of research on matrix factorization has grown rapidly in recent years. Besides its primary application in data compression and connections with PCA and SVD, matrix factorization has been shown useful in a variety of data mining areas, including but not limited to microarray data analysis [Kim and Park 2007; Fogel et al. 2007], protein interaction prediction [Wang et al. 2013], signal processing [Hamza and Brady 2006], classification [Berry et al. 2006], multimedia analysis [Cooper and Foote 2000], matrix completion [Wright et al. 2009a; Nie et al. 2012b; Nie et al. 2012a], collaborative filtering [Rennie and Srebro 2005; Huang et al. 2013a; Huang et al. 2013b] and text mining (documents clustering and topic detection) [Berry et al. 1995; Wang et al. 2011; Wang et al. 2011]. In this paper, we put main focus on data matrices with additional constraint, i.e. all elements from the matrices are non-negative. Non-Negative Matrix Factorization (NMF) provides better interpretability [Lee and Seung 2001] for nonnegative elements matrices such as text and images. There have been quite a few optimization algorithms for NMF problems, such as multiplicative update rules [Lee and Seung 2001], gradient descent [Hoyer 2004], and alternative least square [Paatero and Tapper 1994].

Despite its solid mathematical theory foundations and encouraging performances, there are three major drawbacks for NMF: (1) It assumes the data matrix elements as non-negative, which limits its interpretation to the more general data sets. There are quite a few ways to get around this issue, such as relax the nonnegative constraint [Ding et al. 2010], pre-process the data and transform into the nonnegative one such as normalization; (2) It is prone to outliers since the error for each data point enters the objective function as a squared one. Previous solutions include outlier test at the pre-processing stage, and robust matrix factorization with $\ell_1$ norm [Ke and Kanade 2005] (a generalized NMF formulation) that alleviates the noise influence but fails to maintain feature rotation invariance; (3) NMF performs the factorization in the Euclidean space and learns a parts-based representation, but it fails to discover the intrinsic geometrical and discriminating structure of the data space, which is essentially useful for many real-world applications. A popular choice to incorporate the geometrical structure is to utilize the Laplacian regularization term, such as [Cai et al. 2011].

To address the above drawbacks in existing NMF approaches, we propose a Robust Manifold Non-negative Matrix Factorization (RMNMF) method with the structured sparsity-inducing norm based robust formulation. In the proposed RMNMF model, the mixed-norm $\ell_{2,1}$-norm is used to improve the model robustness, such that the new model is not sensitive to the data outliers and can be applied to practical data mining applications. Meanwhile, the $\ell_{2,1}$ norm based RMNMF model keeps the rotation invariance property to enhance the clustering performance. In addition to this, we also incorporate a manifold regularization term to encode the geometrical information existing in the data. The main contributions of this paper are summarized as follows.

First, we propose a framework that combines $\ell_{2,1}$-norm NMF and spectral clustering methods in a novel way. In particular, we point out an important constraint that many previous papers fail to realize or solve in a satisfactory manner. In contrast, we explicitly include such constraint in our objective function. Note that such constraint turns our objective function into a totally different problem than the naive combination of two existing NMF and spectral clustering methods.

Second, the new constraint in our objective function makes the conventional auxiliary function optimization method no longer applicable. In this paper, we derive a new converged algorithm to optimize our objective function. We decompose the unconstrained problem into four subproblems that each has a closed-form solution. The overall computation cost for our algorithm is comparable to the existing methods.

Last, we show the theoretical connections between our framework and the combination of two classic clustering methods, which are of theoretical significance.

We perform experiments on nine benchmark data sets to validate the proposed new method. All empirical results demonstrate that the proposed method outperforms the related state-of-the-art clustering methods.
Notations. We summarize the notations and the definition of norms used in this paper. Matrices are written as uppercase letters and vectors are written as bold lowercase letters. Given a matrix $W = \{w_{ij}\}$, its $i$-th row, $j$-th column are denoted as $w^i, w^j$, respectively. The $\ell_p$-norm of the vector $v \in \mathbb{R}^n$ is defined as $\|v\|_p = \left( \sum_{i=1}^n |v|^p \right)^{1/p}$ for $p > 0$. The Frobenius norm of the matrix $W \in \mathbb{R}^{d \times m}$ is defined as $\|W\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^m w_{ij}^2} = \sqrt{\sum_{i=1}^d \|w^i\|_2^2}$. $\text{Tr}(W)$ means the trace operation for matrix $W$.

2. ORTHOGONAL MANIFOLD NMF

2.1. Semi-NMF via $\ell_{2,1}$ Norm

To overcome the drawbacks of NMF and other variations that use $\ell_2$ norm, we are proposing matrix factorization using $\ell_{2,1}$-norm. Given a data matrix $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$, where $p$ is the data dimension and $n$ is the number of samples. The $\ell_{2,1}$ norm of matrix $X$ is defined below

$$\|X\|_{2,1} = \sum_{i=1}^n \|x_i\|_2.$$  

(1)

Note that here data vectors are arranged in columns and $\ell_{2,1}$ norm calculates the $\ell_2$ norm for column vectors first. The corresponding matrix factorization task becomes

$$\min_{G \geq 0} \|X - FG^T\|_{2,1}.$$  

(2)

Note that the standard semi-NMF is the following equation

$$\min_{G \geq 0} \|X - FG^T\|_F^2 = \sum_{i=1}^n \|x_i - Fg_i^T\|_2^2.$$  

(3)

The error of each data point enters the objective function as the squared residue error. Thus, a few outliers with large errors easily dominate the objective function because of the squared errors. In contrast, $\ell_{2,1}$ norm simultaneously keeps the feature rotation invariance within each data vector and minimizes outlier impact between data vectors, such that our objective function is more robust than the original one. There is a similar work for $\ell_{2,1}$ NMF in [Kong et al. 2011]. We would highlight two key differences between these two work when we propose our complete objective function. Note that we drop the nonnegative constraint on $F$ to make our method applicable to mixed sign matrices.

2.2. Demonstration Example

In this subsection, we want to demonstrate the robustness of NMF using the $\ell_{2,1}$-norm over $\ell_2$-norm with respect to the noises in synthetic images. We use the images in the first 4 folders of AT&T1 [Samaria and Harter 1994], where each folder carries 10 images. Each image has been vectorized from size $28 \times 23$ and collected into the data matrix. With regard to the original data, if we apply the NMF clustering method using $\ell_{2,1}$-norm and $\ell_2$-norm respectively, we could get the perfect correct clustering results for both methods. In order to increase the clustering difficulty with simulating the real-world noisy data, we randomly occlude a $7 \times 7$ area for each image to simulate the noisy data, resize them into vectors, collect the data matrix, and repeat the clustering process again. The image clustering results of one round are plot in Fig. 1 and Fig. 2. The images with the red boxes are in the wrong clusters. Due to the occlusion, regular NMF now gets 4 images in the wrong clusters, while $\ell_{2,1}$-norm NMF still clusters every

---

1http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
Fig. 1. Clustering results comparison on occluded images. The $\ell_{2,1}$-norm NMF can cluster all images into the correct classes.

Fig. 2. Clustering results comparison on occluded images. The images with red square box are clustered into the incorrect group by $\ell_2$-norm NMF.

image into the correct group. Although the actual results could vary among different rounds due to the random occlusions, $\ell_{2,1}$-norm NMF almost always shows a better result. This demonstrates the $\ell_{2,1}$-norm can be more robust than the $\ell_2$-norm for noisy data vectors.

2.3. Manifold Regularized NMF

So far, our framework performs the parts-based representation learning in the Euclidean space without incorporating the geometric information. In many areas of data mining and machine learning, one is often confronted with intrinsically low-dimensional data lying in a very high-dimensional space [Belkin and Niyogi 2003; Roweis and Saul 2000]. A straightforward solution for the data sampled from a specific manifold is to construct a graph to discretely approximate the manifold, whose vertices correspond to the data samples and the edge weights represent the affinity between the data points. One common assumption about the affinity between data points is cluster assumption [Chapelle et al. 2006], which claims if two data samples are close to each other in the input space, then they are also close to each other in the embedding space. This assumption has been widely used in spectral clustering [Shi and Malik 2000; Ng et al. 2001]. To be specific, in this paper, we define the edge weight matrix $W$ as follows:

$$W_{ij} = \begin{cases} 
1 : & x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\
0 : & \text{otherwise}
\end{cases}$$
where \( N_k(x_i) \) denotes the set of \( k \) nearest neighbors of \( x_i \). It is easy to see \( W \) is symmetric. Laplacian graph is defined as \( L = D - W \), where \( D \) is a diagonal matrix whose entries are column sums of \( W \), \( D_{ii} = \sum_j W_{ij} \).

Cai et al. incorporated Laplacian graph into the NMF framework and proposed the following objective function [Cai et al. 2011]:

\[
\min_{F \geq 0, G \geq 0} \| X - FG^T \|_F^2 + \lambda \text{Tr}(G^T LG),
\]

where \( \lambda > 0 \) is the regularity parameter and \( \text{Tr} \) is the trace operator on the matrix. They called this method as Generalized Non-negative Matrix Factorization (GNMF) and empirically show the effectiveness of their method. However, this method may not get the optimal solution due to the nature of NMF. We will discuss this problem in next several paragraphs.

In this paper, we propose the following objective function:

\[
\min_{G \geq 0, G^TG = I} \| X - FG^T \|_{2,1} + \lambda \text{Tr}(G^T LG),
\]

where \( \lambda \geq 0 \) is the regularization parameter and \( \text{Tr}(\cdot) \) is the trace of a matrix.

We would now explain the two important difference between our work and the \( \ell_{2,1} \)-norm NMF in [Kong et al. 2011].

The first is the obvious Laplacian regularization term, which improves the clustering performance as we will demonstrate.

The second is the subtle additional constraint \( G^TG = I \) here. It has two primary purposes. The first one is to guarantee the uniqueness of our solution. Suppose \( F^* \) and \( G^* \) are the solutions to Eq. (5), then for any given non-zero constant \( c > 1 \), \( cF^* \) and \( G^*/c \) would give the identical value in the first term and lower value for the second term, this is true no matter \( F^* \) and \( G^* \) are local or global optimum solutions. Note that Eq. (4) suffers from this issue as we mentioned above. To circumvent this, Cai et al. normalizes columns of \( F \) after convergence as a remedial measure. Therefore, it is necessary to explicitly include the orthonormal constraint in our framework to avoid such an ad-hoc step. The second purpose is to significantly reduce the computation cost for the optimization algorithm, which we would give details later.

3. COMPUTATIONAL ALGORITHM

The objective function in Eq. (5), as well as many NMF objective functions, is not convex in both \( F \) and \( G \). Therefore, it is unrealistic to expect an algorithm to find the global minimum solution. In the literature, quite a few papers related to NMF made efforts to find the appropriate auxiliary function [Lee and Seung 2001; Kong et al. 2011; Cai et al. 2011] to justify the monotone decreasing property of the objective function. But such technique is difficult to apply here due to the analytical difficulty of \( \ell_{2,1} \) norm and the orthonormal constraint. Here we would tackle this problem with a novel approach, called Augmented Lagrangian Multiplier (ALM) [Bertsekas 2003]. It was originally proposed for convex problems [Gabay and Mercier 1976] and recently extended to non-separable, non-convex problems [Xu et al. 2011; Cai et al. 2013]. The main idea is to eliminate equality constraints and instead add a penalty term to the cost function that assigns a very high cost to the infeasible points. ALM differs from other penalty-based approaches by simultaneously estimating the optimal solution and Lagrange multipliers in an iterative fashion.

Note there are two variables \( F \) and \( G \), we first introduce two auxiliary variables \( E = X - FG^T \) and \( H = G \), then we rewrite the objective function in Eq. (5) to get the following task.

\[
\min_{F,E,H,G} \| E \|_{2,1} + \lambda \text{Tr}(G^T LH)
\]

\[
+ \frac{\mu}{2} \| X - FG^T - E + \frac{1}{\mu} \Lambda \|_F^2 + \frac{\mu}{2} \| G - H + \frac{1}{\mu} \Sigma \|_F^2
\]

\[
\text{s.t. } H \geq 0, G^TG = I
\]

\[\tag{6}\]

ACM Transactions on Knowledge Discovery from Data, Vol. V, No. N, Article A, Publication date: January 2013.
Here $\mu$, $\Lambda$ and $\Sigma$ are all framework parameters for ALM. $\mu$ is the regularity coefficient which determines the penalty for infeasibility, $\Lambda$ and $\Sigma$ each penalizes the gap between the target variable and the auxiliary variable. $\mu$ is a scalar and $\Lambda$, $\Sigma$ are two matrices of size $n \times K$, where $K$ is the desired number of clusters. Now we have an objective function which carries four variables, it is natural to apply alternative optimization method to minimize the objective function, optimizing the objective function with respect to one variable while fixing the other variables. As you will see, the formidable Eq. (6) is reduced to four manageable subproblems. Each subproblem yields a closed-form solution. This procedure repeats until convergence. The following subsections are going to provide optimization details for each subproblem.

3.1. Initialization

Since at each iteration $\mu$ would be updated, it is usually beneficial to start with a small value. The initial value usually ranges in $10^{-5}$ to $10^{-3}$, which depends on the dataset. $\Lambda$ and $\Sigma$ are also updated to approximate the gaps between target variables and auxiliary variables, they can be set to 0 matrices for simplicity. We apply $K$-Means clustering of columns of $X$ to get cluster membership indicator matrix $G$ and initialize $F = XG$. Note that $F$ may contain negative elements if matrix $X$ contains negative elements. In other words, we do not restrict $X$ to be non-negative, this gives more flexibility of $X$.

3.2. Computation of $E$

Optimizing Eq. (6) with respect to $E$ is reduced to the following equation

$$
\min_E \| E \|_{2,1} + \frac{\mu}{2} \| X - FG^T - E + \frac{1}{\mu} \Lambda \|_F^2.
$$

(7)

We need the following Lemma to solve Eq. (7). This Lemma has been also presented as proposition 1 in [Yuan and Lin 2006] without detailed proof.

**Lemma 3.1.** Given a matrix $W = [w_1, \cdots, w_n] \in \mathbb{R}^{m \times n}$ and a positive scalar $\lambda$, then $X^*$ is the optimal solution of

$$
\min \frac{1}{2} \| X - W \|_F^2 + \lambda \| X \|_{2,1},
$$

(8)

and the $i$-th column of $X^*$

$$
X^*(, i) = \begin{cases} \frac{\| w_i \| - \lambda}{\| w_i \|} w_i, & \text{if } \lambda < \| w_i \| \\ 0, & \text{otherwise} \end{cases}
$$

(9)

**Proof.** Note that the objection function in Eq. (8) is equivalent to the following equation.

$$
\sum_{i=1}^n \| x_i - w_i \|_2^2 + \lambda \sum_{i=1}^n \| x_i \|_2
$$

(10)

It is easy to see the above equation can be solved in a decoupled manner for each $x_i$.

$$
\min_{x_i} \frac{1}{2} \| x_i - w_i \|_2^2 + \lambda \| x_i \|_2
$$

(11)

After taking derivative with respect to $x_i$, we get

$$
\frac{\partial \| x_i \|_2}{\partial x_i} = \begin{cases} r, & x_i = 0 \\ \sqrt{x_i}, & \text{otherwise} \end{cases}
$$

(12)

where $r$ is a subgradient vector and $\| r \|_2 \leq 1$. 

ACM Transactions on Knowledge Discovery from Data, Vol. V, No. N, Article A, Publication date: January 2013.
For the case $x_i = 0$, we get
\[-w_i + \lambda r = 0,\] (13)
which implies $\lambda \geq \|w_i\|$. For the case $x_i \neq 0$, we get
\[x_i - w_i + \lambda \frac{x_i}{\sqrt{x_i^T x_i}} = 0.\] (14)
Then we get
\[x_i = \alpha w_i,\] (15)
where $\alpha = \frac{\|x_i\|^2}{\|x_i\|^2 + \lambda} > 0$. It is easy to see each entry of $x_i$ must be of the same sign of the corresponding position of $w_i$. It trivial to plug Eq. (15) back to Eq. (14), solve $\alpha$ and substitute back into Eq. (15). After all these steps, we get
\[x_i = (1 - \frac{\lambda}{\|w_i\|})w_i.\] (16)
Then we know $\lambda < \|w_i\|$ for this case.

The Eq. (7) can be written in the following form:
\[
\min_E \frac{1}{2} \|E - Y\|_F^2 + \frac{1}{\mu} \|E\|_{2,1},
\] (17)
where
\[Y = X - FG^T + \Lambda \mu.\] (18)
According to Lemma 3.1, The solution to Eq. (7) is
\[E(:, i) = \begin{cases} \frac{\|y_i\| - \lambda}{\|y_i\|} y_i, & \text{if } \frac{1}{\mu} < \|y_i\| \\ 0, & \text{otherwise} \end{cases},\] (19)
where $y_i$ is the $i$-th column of $Y$.

### 3.3. Computation of $F$

Optimizing Eq. (6) with respect to $F$ can be simplified to the following one:
\[
\min_{F, G^T G = I} \left\| X - FG^T - E + \frac{1}{\mu} \Lambda \right\|_F.\] (20)
It is not difficult to recognize this is a classical regression problem. We denote:
\[J = X - E + \frac{\Lambda}{\mu}.\] (21)
The solution of $F$ becomes:
\[F = JG,\] (22)
due to $G^T G = I$. This implies the necessity to impose the orthonormal constraint of $G$ in addition to guaranteeing the solution uniqueness, as otherwise solving $F$ requires inverting the large size matrix. It is easy to see the computation cost has been significantly reduced. This is the second purpose we impose the orthonormal constraint $G^T G = I$ in our objective function.
3.4. Computation of $H$

Optimizing Eq. (6) with respect to $H$ becomes the next equation:

$$\min_{H \geq 0} \lambda \text{Tr}(G^T L H) + \frac{\mu}{2} \left\| G - H + \frac{1}{\mu} \Sigma \right\|_F^2 .$$

(23)

After expanding the objective function in the above equation and dropping the terms independent of $H$, we get:

$$\min_{H \geq 0} \text{Tr}(H H^T - 2 H G^T - \frac{2}{\mu} H \Sigma^T + \frac{2\lambda}{\mu} G^T L H) .$$

(24)

With a little algebra, we arrive at:

$$\min_{H \geq 0} \| H - (G + \frac{\Sigma}{\mu} - \frac{\lambda}{\mu} G^T L) \|_F^2 .$$

(25)

Let

$$M = G + \frac{\Sigma}{\mu} - \frac{\lambda}{\mu} G^T L ,$$

(26)

and it is easy to see the solution of $H$

$$H_{ij} = \max(M_{ij}, 0), \quad i = 1, \ldots, n, \quad j = 1, \ldots, K .$$

(27)

3.5. Computation of $G$

Optimizing Eq. (6) with respect to $G$ comes to the equation below:

$$\min_{G^T G = I} \lambda \text{Tr}(G^T L H) + \frac{\mu}{2} \left\| X - F G^T - E + \frac{1}{\mu} \Lambda \right\|_F^2 ,$$

$$+ \frac{\mu}{2} \left\| G - H + \frac{1}{\mu} \Sigma \right\|_F^2 .$$

(28)

With the similar procedure as last subsection and let

$$N = H - \frac{\Sigma}{\mu} + \frac{\lambda}{\mu} L H + (X - E + \frac{\Lambda}{\mu})^T F \in \mathbb{R}^{n \times K} ,$$

(29)

we can write the optimization problem in a compact form:

$$\min_{G^T G = I} \| G - N \|_F^2 .$$

(30)

With the constraint, the above Eq. (30) is equivalent to

$$\max_{G^T G = I} \text{Tr}(G^T N) .$$

(31)

and we use a lemma to solve the above equation. This lemma is related to the classical orthogonal procrustes problem and the approach is similar to theorem 1 in [Yu and Shi 2003].

**Lemma 3.2.** Given the objective function in Eq. (31), the optimal $G$ is defined as:

$$G = U V^T ,$$

(32)

where $U$ and $V$ are left and right singular values of SVD decomposition of $N$, which is defined in Eq. (47).

**Proof.** Let $N = U \Gamma V^T$ according to SVD decomposition, here $U \in \mathbb{R}^{n \times n}$, $\Gamma \in \mathbb{R}^{n \times k}$, $V \in \mathbb{R}^{k \times k}$, here $k < n$.

$$\text{Tr}(G^T N) = \text{Tr}(G^T U \Gamma V^T) = \text{Tr}(\Gamma V^T G^T U) = \text{Tr}(\Gamma Z) = \sum_i \sigma_{ii} z_{ii} ,$$

(33)
where $Z = V^T G U \in \mathbb{R}^{k \times n}$. It is easy to see $Z$ is orthonormal as:
\[
ZZ^T = V^T G^T U U^T G V = I,
\]
therefore $-1 \leq z_{ij} \leq 1$. Note that $\sigma_{ii} \geq 0$, therefore
\[
\text{Tr}(G^T N) = \sum_i \sigma_{ii} z_{ii} \leq \sum_i \sigma_{ii}.
\]
The equality holds when $Z = I_{k,n}$, i.e., $V^T G U = I_{k,n}$. It is then obvious that
\[
G = UV^T
\]

3.6. Updating $\mu$, $\Sigma$ and $\Lambda$

After updating the variables, we need to update these ALM method parameters.

\[
\Lambda = \Lambda + \mu (X - FG^T - E)
\]
\[
\Sigma = \Sigma + \mu (G - H)
\]
\[
\mu = \rho \mu
\]

Here $\rho > 1$ is a parameter to control the convergence speed. The larger $\rho$ is, the fewer iterations required to get the convergence but meanwhile may lose some precision of the final objective function value.

3.7. Algorithm Complexity Analysis

We evaluate the computation costs of our algorithm as follows.

The computation cost for $E$ involves calculation of $Y$ and the update, which is $O(pn + K^3)$ and $O(pn)$ respectively.

The computation cost for $F$ is $O(K^2)$ since $G$ is sparse and there is only one non-zero in each column.

The computation cost for $H$ is $O(K^2)$, same as $F$, due to the sparse property of $G$ and $L$.

The main computation cost for $G$ includes the calculation of $N$ and its SVD decomposition, which are of order $O(p^3)$ and $O(nK^2)$, respectively.

The overall cost for each iteration is then of order $O(p^3 + pn)$. In real data experiments, we find our computation speed is comparable to classic clustering algorithms.

3.8. Summarization of Algorithm

We first give a summary of our method so far in Inexact ALM Algorithm 1. Note that $\mu$ grows exponentially as the number of iterations increases, and it can be observed that objective function in Eq. (6) asymptotically converges to the original objective function in Eq. (5). To see this, note that the third and fourth term in Eq. (6) have to be both 0 when $\mu$ is infinity to keep the objective function value finite. Therefore given sufficient number of iterations, the convergence of our algorithm is guaranteed, this completes the intuitive sketch of our algorithm. The convergence of our algorithm relies on the convergence of ALM framework. The convergence of ALM framework was proved and discussed in previous papers and books [Lin et al. 2009; Bertsekas 1996; Hestenes 1969; Powell 1969; Bertsekas 1982]. In particular, [Lin et al. 2009] presents the detailed proof of inexact ALM to convex objective functions. The rigorous mathematical argument about our algorithm convergence is very difficult and it remains unknown in literature so far. However, it is clear that as the number of iterations increases, $\mu$ grows exponentially and this ensures $H$ converges to $G$ given sufficient number of iterations. Our empirical experiments also confirm the convergence of our algorithm, for the benchmark data sets used in this paper. One thing to note is that Eq. (6) is non-convex, the solution yielded via alternative optimization from Eq. (6) is not a global optimal one. In fact, it is a local solution, however, it is also a unique one due to the orthonormal constraint. This solution
converges to the solution corresponding to Eq. (5) asymptotically and therefore the solution to
Eq. (5) is also a local but unique solution.

We set the convergence criteria as \( \frac{|J_{t+1} - J_t|}{J_t} < \varepsilon \), where \( J_t \) is the objective function value in
Eq. (6) at \( t \)-th iteration and \( \varepsilon \) is set at \( 10^{-4} \). The maximal number of iterations is empirically set at
500, which is usually more than enough. Our experiments find our algorithm converges within 20
iterations for all our benchmark data. If users prefer to have the algorithm converged within a certain
number of iterations when the convergence criteria is not satisfied, they may consider increasing the
\( \rho \) value a little bit.

Finally, we want to mention that the all ALM parameters such as \( \mu \), \( \rho \) generally only affect the
convergence speed, and have little effect on the clustering results if they are given reasonable initial
values. In this paper, we set these ALM framework parameters empirically according to the literature
without tuning them.

4. CONNECTIONS TO OTHER CLUSTERING METHODS

In this section, we will show theoretically our framework is a combination of two classic clustering
methods, \( K \)-Means clustering and spectral clustering. Therefore, when a proper \( \lambda \) value is chosen,
our method inherits the merits of both methods.

4.1. Connection to \( K \)-Means

First, we want to show the relation between Eq. (6) and \( K \)-Means clustering. The objective function
of \( K \)-Means clustering is:

\[
J = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - f_k\|_2^2,
\]

where \( f_k \) is the centroid of the \( k \)-th cluster \( C_k \).

The outlier issue for standard NMF also exists for \( K \)-Means. A robust variation of the \( K \)-Means
clustering can be formulated as:

\[
J' = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - f_k\|_2,
\]

where the errors are not squared, and therefore cluster centroids will not drift too much towards
outliers in this robust formulation.

When \( \lambda = 0 \) for our method, we now show that given non-negative \( X \), it is a relaxed form of
the robust \( K \)-Means clustering Eq. (52). \( G \) represents the solution of the clustering: \( G_{ki} = 1 \) if \( x_i \)
belongs to cluster \( C_k \), \( G_{ki} = 0 \) otherwise. We have:

\[
J' = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - f_k\|_2
= \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - F g_i\|_2
= \sum_{i=1}^{n} \|x_i - F g_i\|_2 = \|X - FG\|_{2,1}.
\]

Thus our method implicitly performs a robust \( K \)-Means clustering.

4.2. Connection to Spectral Clustering

Spectral clustering, a clustering method based on the Laplacian graph, looks for a partition of the
graph such that the weights between different groups have a very low weight (which means that
ALGORITHM 1: Robust Manifold Orthogonal NMF

**Input:** $X$, the number of data clusters $K$, $\mu, \Gamma, \rho$, maximum number of iterations $T$

**Output:** Converged $F$ and $G$

Initialize $F$ and $G$ using K-means as described.

**while** not converged and iteration less than $T$ **do**

Step (A). Update $E$ with the formula

$$E(:,i) = \begin{cases} \frac{\|y_i\| - \lambda}{\|y_i\|} y_i, & \text{if } \frac{1}{\mu} < \|y_i\| \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (40)

where

$$Y = X - FG^T + \frac{\Lambda}{\mu}.$$  \hspace{1cm} (41)

Step (B). Update $F$ with the formula

$$F = JG$$  \hspace{1cm} (42)

where

$$J = X - E + \frac{\Lambda}{\mu}.$$  \hspace{1cm} (43)

Step (C). Update $H$ with the formula

$$H_{ij} = \max(M_{ij}, 0), \quad i = 1, \ldots, n, \quad j = 1, \ldots, K.$$  \hspace{1cm} (44)

where

$$M = G + \frac{\Sigma}{\mu} - \frac{\lambda}{\mu} G^T L.$$  \hspace{1cm} (45)

Step (D). Update $G$ with the formula

$$G = UV^T$$  \hspace{1cm} (46)

where $U$ and $V$ are left, right singular matrices of the SVD decomposition of $N$

$$N = H - \frac{\Sigma}{\mu} + \frac{\lambda}{\mu} LH + (X - E + \frac{\Lambda}{\mu})^T F \in \mathbb{R}^{n \times K},$$  \hspace{1cm} (47)

Step (E). Update $\Lambda, \Sigma$ and $\mu$ using

$$\Lambda = \Lambda + \mu(X - FG^T - E)$$  \hspace{1cm} (48)

$$\Sigma = \Sigma + \mu(G - H)$$  \hspace{1cm} (49)

$$\mu = \rho \mu.$$  \hspace{1cm} (50)

end

points in different clusters are dissimilar from each other). When $\lambda \to \infty$, our objective function can be approximated by:

$$\min_{G \geq 0, G^T G = I} Tr(G^T L G).$$  \hspace{1cm} (54)

This is exact the objective function of spectral clustering. Depends on whether $L$ is normalized or not, Shi et al. and Ng et al. proposed two different versions of normalized spectral clustering [Shi and Malik 2000; Ng et al. 2001]. There is a detailed discussion about these three different approaches in terms of graph cut theory and perturbation theory respectively in [Luxburg 2007].
5. EXPERIMENTAL RESULTS

In this section, we will evaluate the performance of the proposed method on the benchmark data set. We compare our method with $K$-Means, Normalized Cut (NCut) \cite{Ng et al. 2001}, NMF \cite{Lee and Seung 2001}, GNMF \cite{Cai et al. 2011} and $\ell_{2,1}$-norm based NMF \cite{Kong et al. 2011}.

5.1. Evaluation Metrics

To evaluate the clustering results, we adopt the three widely used clustering performance measures which are defined below.

**Clustering Accuracy** discovers the one-to-one relationship between clusters and classes and measures the extent to which each cluster contained data points from the corresponding class. Clustering Accuracy is defined as follows:

$$\text{Acc} = \frac{\sum_{i=1}^{n} \delta(\text{map}(r_i), l_i)}{n}, \quad (55)$$

where $r_i$ denotes the cluster label of $x_i$ and $l_i$ denotes the true class label, $n$ is the total number of documents, $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and $\text{map}(r_i)$ is the permutation mapping function that maps each cluster label $r_i$ to the equivalent label from the data set.

**Normalized Mutual Information** (NMI) is used for determining the quality of clusters. Given a clustering result, the NMI is estimated by

$$\text{NMI} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} n_{i,j} \log \frac{n_{i,j}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^{c} n_i \log \frac{n_i}{n})(\sum_{j=1}^{c} \hat{n}_j \log \frac{\hat{n}_j}{n})}}, \quad (56)$$

where $n_i$ denotes the number of data contained in the cluster $C_i (1 \leq i \leq c)$, $\hat{n}_j$ is the number of data belonging to the $L_j (1 \leq j \leq c)$, and $n_{i,j}$ denotes the number of data that are in the intersection between cluster $C_i$ and the class $L_j$.

**Purity** measures the extent to which each cluster contained data points from primarily one class. The purity of a clustering is observed by the weighted sum of individual cluster purity values, given as follows:

$$\text{Purity} = \frac{\sum_{i=1}^{K} n_i}{n} P(S_i), \quad P(S_i) = \frac{1}{n_i} \max_j (n_{i,j}^2) \quad (57)$$

where $S_i$ is a particular cluster size of $n_i$, $n_{i,j}^2$ is the number of the $i$-th input class that was assigned to the $j$-th cluster. $K$ is the number of the clusters and $n$ is the total number of the data points.

Among all the three measures, they are generally positive correlated and a larger value indicates a better clustering solution.

5.2. Data Sets Description

There are in total nine data sets used in our experiments, which includes four image ones: AR \cite{Martinez et al. 2001; Martinez and Benavente 1998}, AT&T \cite{Samaria and Harter 1994}, subset of JAFFE \cite{Lyons et al. 1998}, subset of PIE \cite{Sim et al. 2002}. The other five non-image data sets are all from UCI machine learning repository \cite{Frank and Asuncion}: Abalone, Ionosphere, Movement, Scale and Wine Quality (WineQ).

AR contains cropped images from 100 individuals, and each individual has 26 images. Images from two sessions feature frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). Participants may wear clothes and glasses, have different make-up and hair styles.

\footnote{1http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html, other downloads 1.}

\footnote{2http://www.zjucadcg.cn/dengcai/Data/data.html, we use first 10 images in each class.}
Table I. Description of Data Sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number of Samples</th>
<th>Dimensions</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>2600</td>
<td>792</td>
<td>100</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>400</td>
<td>168</td>
<td>40</td>
</tr>
<tr>
<td>JAFFE</td>
<td>800</td>
<td>256</td>
<td>68</td>
</tr>
<tr>
<td>PIE</td>
<td>680</td>
<td>100</td>
<td>15</td>
</tr>
<tr>
<td>Abalone</td>
<td>4177</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Movement</td>
<td>560</td>
<td>90</td>
<td>15</td>
</tr>
<tr>
<td>Scale</td>
<td>625</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>WineQ</td>
<td>4898</td>
<td>11</td>
<td>3</td>
</tr>
</tbody>
</table>

AT&T carries 10 different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions and facial details. All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position.

JAFFE carries 7 different emotional frontal faces of 15 Japanese female models. The 7 facial expressions consist of 6 basic facial expressions plus 1 neutral. The images are taken against a homogeneous background with almost no lighting variation.

PIE carries 10 frontal images from 68 subjects. They are of different poses, illumination conditions and expressions. The images we choose from each subject are extremely noisy due to the above variations.

Abalone has 8 physical measurements to predict abalone’s age. The ground truth age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope.

Ionosphere consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. “Good” radar returns are those showing evidence of some type of structure in the ionosphere. “Bad” returns are those that do not; their signals pass through the ionosphere.

Movement contains 15 classes of 24 instances each, where each class references to a hand movement type in LIBRAS. In the video pre-processing stage, a time normalization is carried out selecting 45 frames from each video. In each frame, the centroid pixels of the segmented objects are found, which compose the discrete version of the curve F with 45 points. All curves are normalized in the unitary space.

In the Scale data set, each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The attributes are the left weight, the left distance, the right weight, and the right distance. The class label is the greater of (left-distance * left-weight) and (right-distance * right-weight). If they are equal, it is balanced.

WineQ contains 12 attributes of wine, including fixed acidity, volatile acidity etc, for classification purpose. 4898 red and white wine samples, from the north of Portugal are given scores between 0 and 10.

Among 4 image data sets, AR and PIE are relatively noisy because they contain a lot of occluded images or images with varying light conditions. This makes clustering more especially difficult considering the large number of classes each one has. AR has images from 100 different persons and PIE has 68 different ones. In contrast, AT&T and JAFFE have fewer classes and clean images.

To make our data sets have more diversity, we also include 5 non-image ones. Note that Ionosphere contains negative elements.

Table 1 summarizes the characteristics of the data sets used in the experiments.

5.3. Parameters Setting

Since most algorithms have one or more parameters to be tuned, to compare these algorithms fairly, we run these algorithms with different parameter settings, and select the best average result to com-
We then set the regularization parameter by searching the grid \( K \) set the neighborhood size \( \ell \). The parameters have little influence on the clustering results, so we set the initial values empirically. We compare with each other. We set the number of clusters equal to the ground truth class number for all the data sets and clustering algorithms.

For Ncut and GNMF, we construct the Laplacian graph by searching the grid from \( \{1, 3, 5, 7, 9\} \). In addition to that, for GNMF, we set the regularization parameter by searching the grid \( \{0, 0.1, 1, 10, 100\} \).

For our method RMNMF, we initialize \( \Gamma, \Sigma \in 0^{n \times p} \), \( \mu = 10^{-3} \), \( \rho = 1.05 \). As mentioned, these parameters have little influence on the clustering results, so we set the initial values empirically. We set the neighborhood size \( K \) equal to 5 here for simplicity and will investigate its influence empirically. We then set the regularization parameter by searching the grid \( \{10^{-3}, 10^{-2}, \cdots, 1\} \).

Note that no parameter selection is needed for \( K \)-Means and NMF, given the number of clusters.

Under each parameter setting of each method mentioned above, we repeat clustering 20 times and compute the average result. We report the best average result for each method.

### 5.4. Clustering Results on Non-Negative Data Sets

Fig. (3) summarizes the clustering performance for each method on 8 out of 9 data sets we mentioned, which are non-negative. Although the standard deviations of all methods are not displayed in the figure, we tested the statistical difference between our method and other competitive methods via one-tailed T-test. The performance differences between our method and other methods are all significant on all the benchmark data sets.

In addition, GNMF also performs NMF on these data sets which confirms that the added geometric regularity term in our objective function incorporates the geometric and manifold information. In addition, RMNMF arises to be non-negative, we truncate corresponding negative values to 0 to apply them. From Table 2, we observe that RMNMF outperforms \( K \)-means and normalized spectral clustering.

### 5.5. Clustering Results on Data with Negative Values

We would like to show our method also applies to the data set containing negative values. Due to the scarcity of appropriate data sets for clustering, here we are only able to demonstrate on the data set Ionosphere mentioned earlier. Since NMF, GNMF and \( \ell_{2,1} \)-norm NMF require the data matrix to be non-negative, we truncate corresponding negative values to 0 to apply them. From Table 2, we can observe our method still outperforms \( K \)-means and normalized spectral clustering.

### 5.6. Computation Time Comparison

In this section, we list the average computation time in Table III for all methods on these benchmark data sets. The Experiments are conducted are 8 core Linux server with each core CPU at 2.5GHz, the total memory is 16G. All codes are written in Matlab or use Matlab toolbox function. For each method, we decide the optimal parameter setting in terms of optimal clustering accuracy. We can observe that RMNMF is slower than \( K \)-Means but in line with other methods on most data sets.
Table III. Computation Time for All Methods (in seconds)

<table>
<thead>
<tr>
<th>Data set</th>
<th>RMNMF</th>
<th>K-Means</th>
<th>NMF</th>
<th>GNMF</th>
<th>Ncut</th>
<th>$\ell_{2,1}$-norm NMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>32.24</td>
<td>9.51</td>
<td>21.21</td>
<td>17.24</td>
<td>15.63</td>
<td>23.14</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>2.14</td>
<td>0.47</td>
<td>6.23</td>
<td>2.29</td>
<td>2.43</td>
<td>2.32</td>
</tr>
<tr>
<td>JAFFE</td>
<td>3.87</td>
<td>1.17</td>
<td>5.34</td>
<td>2.57</td>
<td>4.07</td>
<td>4.35</td>
</tr>
<tr>
<td>PIE</td>
<td>1.14</td>
<td>0.19</td>
<td>0.15</td>
<td>0.24</td>
<td>0.65</td>
<td>0.86</td>
</tr>
<tr>
<td>Abalone</td>
<td>1.14</td>
<td>0.16</td>
<td>0.96</td>
<td>1.02</td>
<td>1.16</td>
<td>1.17</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>1.67</td>
<td>0.22</td>
<td>1.46</td>
<td>1.53</td>
<td>1.62</td>
<td>1.84</td>
</tr>
<tr>
<td>Movement</td>
<td>2.01</td>
<td>0.18</td>
<td>0.04</td>
<td>1.37</td>
<td>3.24</td>
<td>2.28</td>
</tr>
<tr>
<td>Scale</td>
<td>3.57</td>
<td>0.42</td>
<td>8.47</td>
<td>2.03</td>
<td>2.37</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Fig. 3. Clustering Results on Different Data Sets
5.7. Study on the Regularity $\lambda$

In this subsection, we check the influence with respect to the regularity parameter $\lambda$. Fig. (4) shows the clustering results of RMNMF as $\lambda$ varies from 0.001 to 1. It can be observed that overall our method is not very sensitive to the choice of $\lambda$ values. In particular, as $\lambda$ value varies, the performance of our method also approximates from the corresponding metric measure from $\ell_{2,1}$-norm to spectral clustering.

5.8. Study on the Neighborhood Size

In this subsection, we investigate the influence with respect to the neighborhood size $k$. When we vary the value of neighborhood size $k$, we keep the other parameters fixed at the optimal value. We find the overall performance of RMNMF is stable, but it decreases slightly as $k$ increases.

As we described, RMNMF incorporates a $k$-nearest Laplacian graph to capture the local geometric structure of the data distribution, which assumes the two neighboring points share the same label. The assumption could fail when $k$ increases, this explains the previous discovery.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a Robust Manifold Nonnegative Matrix Factorization (RMNMF) method based on semi-nonnegative matrix factorization with manifold regularization. The main motivations of our paper are to alleviate the outlier issue and incorporate the geometric information. RMNMF objective function is solved via the Augmented Lagrangian Multiplier alternating manner. The original problem is decomposed into 4 subproblems, each provides a closed form solution. The convergence of our algorithm is obvious and theoretically founded. Experiments of clustering on nine benchmark data sets demonstrate that the proposed method outperforms many state-of-the-art clustering methods.

In our future work, we will investigate other forms of manifold regularization such as feature regularity to enable co-clustering [Gu and Zhou 2009; Dhillon et al. 2008], which considers the features of the high-dimensional data also underly on the specific manifold.

7. ACKNOWLEDGEMENT

This research was partially supported by NSF DMS-0915228, IIS-1117965, IIS-1302675, IIS-1344152.

REFERENCES


X. Cai, F. Nie, and H. Huang. 2013. Exact Top-k Feature Selection via $\ell_{2,0}$-Norm Constraint. In International Joint Conference on Artificial Intelligence (IJCAI). 1240–1246.


Fig. 4. Clustering Results with Varying λ Values
Fig. 5. Clustering Results with Varying $k$ Values


H. Wang, H. Huang, C. Ding, and F. Nie. 2013. Predicting Protein-Protein Interactions from Multimodal Biological Data Sources via Nonnegative Matrix Tri-Factorization. In International Conference on Research in Computational Molecular Biology (RECOMB), 302–313.


