Non-negative Least Squares Regression (Collaborative Representation) for Subspace Clustering

July 1, 2017

Abstract

1 Introduction

The sparse subspace clustering (SSC) [1, 2] and low rank representation (LRR) [3, 4] methods construct the affinities matrix of an undirected graphs by sparse or low rank representations. However, the coefficients in these graphs can be negative, which allows the data points to "cancel each other out" by subtraction or include outliers which share opposite physical meaning, which is physically absurd for many visual analysis. For example, given a set of data points $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, if two of these points are just opposite to each other, i.e., $\boldsymbol{x}_i = -\boldsymbol{x}_j$. Then the representation of $\boldsymbol{x}_i$ over this set of data points would be $\boldsymbol{c}_i$, in which $\boldsymbol{c}_{ij} = -1$ while other coefficients would be zero. This is clearly not what we want since the two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are just opposite and would be problematic to represent each other. It is better to give a detailed example or give a figure to illustrate this point. The other methods also employ *self-expressiveness* in which each data point can be represented by a linear combination of other data points that generally involve complex cancellations between positive and negative samples. This combination form lacks intuitive meaning in physical sense. However, the non-negative LSR only allows non-negative combination of multiple data points to additively represent the given data point, which is compatible with the intuitive notion of combining parts into a whole.

1.1 Physically correct! Not only mathematically correct!

Each related data points are activated to generate a new data point additively. From this point, the model can be viewed as a generative model. Besides,since we add a contraint of "sum to one", hence the overall amount to be added is limited and the model will automatically select the most similar points from the same subspace. From this point, the model also enjoys discriminative ability. Hence, with non-negative and sum-to-one constraints, the NNLSR model simultaneously has generative and discriminative properties. Seen in this light, the proposed model can be guaranteed to perform well on representation based classification and clustering problems. Considering that the generative and discriminative property often need "sparse plus low rank" prior modeling, which is solved by ADMM algorithm with three variables [5], the proposed model is very simple and can be solved under the ADMM framework with only two variables, in which the convergency condition can be guaranteed.

In fact, non-negativity is more reasonable with the biological modeling of the visual data and often lead to better performance for data representation [6]. In many real-world problems, the underlying parameters which represent uantities can only take on non-negative values. Examples in this inlcude amounts of materials, chemical concentrations, pixel intensities, the compounds of endmenbers in hyperspectral images, to name a few. Our work is inherently benefit from the advantages of non-negative matrix factorization for parts-based facial analysis [6]

Non-negatie least squares has been studied in [7] for sparse recovery without regularization. The authors compare with the non-negative LASSO method [8] and found

that the proposed non-negative least squares model can achieve similar or even better performance on sparse recvoery problems. For $\boldsymbol{A} = (\boldsymbol{a}_{ij})$ we write $\boldsymbol{A} \geq 0$ if $\boldsymbol{a}_{ij} \geq 0$ for each i and j and we say $\boldsymbol{A}$ is a non-negative matrix. This notation can be naturally extended for vectors. Similarly we can define non-positive matrix, negative matrix, and negatvie matrix. The famous Perron-Frobenious theory are widely used to analysis non-negative matrices.

The proposed model can be solved under the framework of ADMM [9]. To make the method more efficient and scalable, we can also employ the linearized ADMM with adaptive penalty (LADMAP) [10], which uses less auxiliary variables and no matrix inversion.

2 Motivation

- Positive collaborative representation could achieve sparse representation since similar points are sparse while dissimilar points are dense.
- Positive supports are positive to self-reppresentation while negative supports are negative to self-representation.
- Due to the space of constraint is only half of the original LSR model, the searching for solution does not need too many iterations such as in SSC and LRR method, hence the proposed method keeps the efficiency of LSR and the speed of our method is faster than the others.
- In summmary, Physically feasible, higher accuray, and faster speed.

3 LSR Model

It is widely known that the least squares regression (LSR) model [11], which can be expressed as follows:

$$\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^2 + \lambda\|\boldsymbol{A}\|_F^2, \tag{1}$$

has a form of closed solution as

$$\hat{\boldsymbol{A}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}. \tag{2}$$

In subspace clustering community, the LSR method [12] proposed by Lu et al. can be formulated as follows:

$$\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^2 + \lambda\|\boldsymbol{A}\|_F^2 \text{ s.t. } \operatorname{diag}(\boldsymbol{A}) = \boldsymbol{0}. \tag{3}$$

Here we denote by $\operatorname{diag}(\boldsymbol{A})$ both a diagonal matrix whose diagonal elements are the diagonal entries of $\boldsymbol{A}$ and the vector consisted of the diagonal elements. According to [12], the above problem has the optimal solution as

$$\hat{\boldsymbol{A}} = -\boldsymbol{Z}(\operatorname{diag}(\boldsymbol{Z})) \text{ s.t. } \operatorname{diag}(\hat{\boldsymbol{A}}) = \boldsymbol{0}, \tag{4}$$

where $\boldsymbol{Z} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$.

The constraint of $\operatorname{diag}(\boldsymbol{A}) = \boldsymbol{0}$ in (1) could be removed and the LSR model achieves similar performance.

4 Least Squares Regression with Constraint $\operatorname{diag}(\boldsymbol{A}) = \boldsymbol{0}$

The LSR model can be reformulated as a collaborative representation model [13] for subspace clustering with an additional constraint of $\operatorname{diag}(\boldsymbol{A}) = \boldsymbol{0}$. The constraint of $\operatorname{diag}(\boldsymbol{A}) = \boldsymbol{0}$ is used to avoid the samples to represent themselves.

By introducing auxiliary variables into the optimization program, we can set $\boldsymbol{C} = \boldsymbol{A}$. The LSR model (1) can be transformed into

$$\begin{aligned} &\min_{\boldsymbol{A},\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^F + \lambda\|\boldsymbol{C}\|_F^2 \\ &\text{s.t. } \boldsymbol{C} = \boldsymbol{A} - \operatorname{diag}(\boldsymbol{A}), \end{aligned} \tag{5}$$

whose solution for $\boldsymbol{A}$ coincides with the solution of Eq. (1). By introducing a Lagrangian multipliers $\boldsymbol{\Delta}$ and a penalty parameter ρ, the Lagrangian function of the Eq. (35) can be written as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Delta}, \rho) = \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^F + \lambda\|\boldsymbol{C}\|_F^2 \\ &+ \langle\boldsymbol{\Delta}, \boldsymbol{C} - (\boldsymbol{A} - \operatorname{diag}(\boldsymbol{A}))\rangle + \frac{\rho}{2}\|\boldsymbol{C} - (\boldsymbol{A} - \operatorname{diag}(\boldsymbol{A}))\|_F^2 \end{aligned} \tag{6}$$

Denote by $(\boldsymbol{C}_k, \boldsymbol{A}_k)$ the optimization variables at iteration k, by $\boldsymbol{\Delta}_k$ the Lagrangian multipliers at iteration k, and by ρ the penalty parameter at iteration k. Taking detivatives of $\mathcal{L}$ with respect to the variables and setting the derivatives

to be zeros, we can alternatively update the variables as follows:

(1) Obtain $\boldsymbol{A}_{k+1}$ by minimizing $\mathcal{L}$ with respect to $\boldsymbol{A}$, while fixing $(\boldsymbol{C}_k, \boldsymbol{\Delta}_k)$. This is equivalent to solve the following problem:

$$\begin{aligned} &\boldsymbol{A}_{k+1} = \boldsymbol{J} - \text{diag}(\boldsymbol{J}), \\ &\boldsymbol{J} = (\boldsymbol{X}^\top \boldsymbol{X} + \frac{\rho}{2}\boldsymbol{I})^{-1}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\rho}{2}\boldsymbol{C}_k + \frac{1}{2}\boldsymbol{\Delta}_k) \end{aligned} \tag{7}$$

(2) Obtain $\boldsymbol{C}_{k+1}$ by minimizing $\mathcal{L}$ with respect to $\boldsymbol{C}$, while fixing $(\boldsymbol{A}_{k+1}, \boldsymbol{\Delta}_k)$. This is equivalent to solve the following problem:

$$\boldsymbol{C}_{k+1} = \arg\min_{\boldsymbol{C}} \frac{\rho}{2}\|\boldsymbol{C} - (\boldsymbol{A}_{k+1} - \rho^{-1}\boldsymbol{\Delta}_k)\|_F^2 + \lambda\|\boldsymbol{C}\|_F^2 \tag{8}$$

This is a least squares regression problem which has a closed-form solution as

$$\boldsymbol{C}_{k+1} = (\rho + 2\lambda)^{-1}(\rho\boldsymbol{A}_{k+1} - \boldsymbol{\Delta}_k). \tag{9}$$

(3) Obtain the Lagrangian multipliers $\boldsymbol{\Delta}_{k+1}$ while fixing $(\boldsymbol{C}_{k+1}, \boldsymbol{A}_{k+1})$:

$$\boldsymbol{\Delta}_{k+1} = \boldsymbol{\Delta}_k + \rho(\boldsymbol{C}_{k+1} - \boldsymbol{A}_{k+1}). \tag{10}$$

Convergency analysis?

5 Non-Negative Least Squares Regression with $\text{diag}(\boldsymbol{A}) = \boldsymbol{0}$

This model enforces non-negative representation and hence produce sparse solutions, in the sense that it results only a few non-negative coefficients.

The performance of this method is much better than the original least squares regression (LSR) based subspace clustering method proposed by Lu et al. [12].

The LSR model in [12] can be reformulated as a collaborative representation model [13] for subspace clustering with an additional constraint of $\text{diag}(\boldsymbol{A}) = \boldsymbol{0}$. In this section, we want to mention that the coefficient matrix $\boldsymbol{C}$ with additional constraint could benefit the performance of subspace clustering. Motivated by the non-negative coefficient should share positive relationship while negative coefficients share negative relationship, we argue that non-negative representational coefficients should better represent the data points from the same subspace, while negative coefficients correspond to points from different subspaces. By this way, the negative coefficients will negatively influence the relationship among the points in the same subsapce and hence degrade the performance of the model on subspace clustering. Based on these observations, in this section, we propose to add an constraint on the coefficient matrix $\boldsymbol{A}$ that the elements in $\boldsymbol{A}$ should be non-negative, i.e., $\boldsymbol{A} \geq 0$. Hence, the proposed non-negative collaborative representation model can be formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^2 + \lambda\|\boldsymbol{A}\|_F^2 \\ &\text{s.t.} \quad \text{diag}(\boldsymbol{A}) = \boldsymbol{0}, \boldsymbol{A} \geq 0, \end{aligned} \tag{11}$$

where $\boldsymbol{A} \geq 0$ means that each element of $\boldsymbol{A}$ is non-negative.

By introducing auxiliary variables into the optimization program, we can set $\boldsymbol{C} = \boldsymbol{A}$. The LSR model (1) can be transformed into

$$\begin{aligned} &\min_{\boldsymbol{A},\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^F + \lambda\|\boldsymbol{C}\|_F^2 \\ &\text{s.t. } \boldsymbol{C} = \boldsymbol{A} - \text{diag}(\boldsymbol{A}), \boldsymbol{C} \geq 0, \end{aligned} \tag{12}$$

whose solution for $\boldsymbol{A}$ coincides with the solution of Eq. (9). By introducing a Lagrangian multipliers $\boldsymbol{\Delta}$ and a penalty parameter ρ, the Lagrangian function of the Eq. (35) can be written as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Delta}, \rho) = \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^F + \lambda\|\boldsymbol{C}\|_F^2 \\ &+ \langle\boldsymbol{\Delta}, \boldsymbol{C} - (\boldsymbol{A} - \text{diag}(\boldsymbol{A}))\rangle + \frac{\rho}{2}\|\boldsymbol{C} - (\boldsymbol{A} - \text{diag}(\boldsymbol{A}))\|_F^2 \end{aligned} \tag{13}$$

In this case, the matrix containing only zeros is a feasible starting point as it contains no negative values. After initializing the $\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Delta}$ as zero matrices, the ADMM algorithm iterates consist of 1) minimizing $\mathcal{L}$ with respect to $\boldsymbol{A}$ while fixing the other variables, and 2) minimizing $\mathcal{L}$ with respect to $\boldsymbol{C}$ subject to the constraint $\boldsymbol{C} \geq 0$ while fixing the other variables; 3) updating the Lagrangian variable $\boldsymbol{\Delta}$ while fixing the other variables. Specifically, denote by $(\boldsymbol{C}_k, \boldsymbol{A}_k)$ the optimization variables at iteration k, by $\boldsymbol{\Delta}_k$ the Lagrangian multipliers at iteration k. Taking detivatives of $\mathcal{L}$ with respect to the variables and setting the

derivatives to be zeros, we can alternatively update the variables as follows:

(1) Obtain $\boldsymbol{A}_{k+1}$ by minimizing $\mathcal{L}$ with respect to $\boldsymbol{A}$, while fixing $(\boldsymbol{C}_k, \boldsymbol{\Delta}_k)$. This is equivalent to solve the following problem:

$$\begin{aligned} &\boldsymbol{A}_{k+1} = \boldsymbol{J} - \text{diag}(\boldsymbol{J}), \\ &\boldsymbol{J} = (\boldsymbol{X}^\top \boldsymbol{X} + \frac{\rho}{2}\boldsymbol{I})^{-1}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\rho}{2}\boldsymbol{C}_k + \frac{1}{2}\boldsymbol{\Delta}_k) \end{aligned} \tag{14}$$

(2) Obtain $\boldsymbol{C}_{k+1}$ by minimizing $\mathcal{L}$ with respect to $\boldsymbol{C}$, while fixing $(\boldsymbol{A}_{k+1}, \boldsymbol{\Delta}_k)$. This is equivalent to solve the following problem:

$$\begin{aligned} \boldsymbol{C}_{k+1} = \arg\min_{\boldsymbol{C}} &\|\boldsymbol{C} - (2\lambda + \rho)^{-1}(\rho \boldsymbol{A}_{k+1} - \boldsymbol{\Delta}_k)\|_F^2 \\ \text{s.t.} \quad &\boldsymbol{C} \geq 0. \end{aligned} \tag{15}$$

This is a non-negative least squares problem which can be solved by many solves developed via active set method [14] or specifically, the algorithm of Lawson and Hanson [].

(3) Obtain the Lagrangian multipliers $\boldsymbol{\Delta}_{k+1}$ while fixing $(\boldsymbol{C}_{k+1}, \boldsymbol{A}_{k+1})$:

$$\boldsymbol{\Delta}_{k+1} = \boldsymbol{\Delta}_k + \rho(\boldsymbol{C}_{k+1} - \boldsymbol{A}_{k+1}). \tag{16}$$

Convergency analysis?

The above solution is slow when the number of column of $\boldsymbol{X}$ is much larger than its number of rows, i.e., when $N > d$. where d is the dimension of features for each sample and N is the number of samples in $\boldsymbol{X}$. Hence, we propose the following solution via *Woodbury Identity* to reduce the computational cost for the inversion of the solution in Eq. (12).

6 Large Scale Subset Selection Via Woodbury Identity

The Woodbury Identity is

$$(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1}+\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}. \tag{17}$$

Then the first step of updating $\boldsymbol{A}$ can be formulated as follows;

(1) Obtain $\boldsymbol{A}_{k+1}$ by minimizing $\mathcal{L}$ with respect to $\boldsymbol{A}$, while fixing $(\boldsymbol{C}_k, \boldsymbol{\Delta}_k)$. This is equalivalently to solve the following problem

$$\begin{aligned} &\boldsymbol{A}_{k+1} = \boldsymbol{J} - \text{diag}(\boldsymbol{J}), \\ &\boldsymbol{J} = (\boldsymbol{X}^\top \boldsymbol{X} + \frac{\rho}{2}\boldsymbol{I})^{-1}(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\rho}{2}\boldsymbol{C}_k + \frac{1}{2}\boldsymbol{\Delta}_k) \end{aligned} \tag{18}$$

Since the matrices $\boldsymbol{X}^\top \boldsymbol{X}$ is of $N \times N$ dimension. It is computational expensive when N is very large. By employing the Woodburry Identity mentioned above, we can have

$$(\frac{\rho}{2}\boldsymbol{I} + \boldsymbol{X}^\top \boldsymbol{X})^{-1} = \frac{2}{\rho}\boldsymbol{I} - (\frac{2}{\rho})^2 \boldsymbol{X}^\top (\boldsymbol{I} + \frac{2}{\rho}\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}. \tag{19}$$

and transform this problem as

$$\begin{aligned} \boldsymbol{J} = &(\frac{2}{\rho}\boldsymbol{I} - (\frac{2}{\rho})^2 \boldsymbol{X}^\top (\boldsymbol{I} + \frac{2}{\rho}\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}) \\ &* (\boldsymbol{X}^\top \boldsymbol{X} + \frac{\rho}{2}\boldsymbol{C}_k + \frac{1}{2}\boldsymbol{\Delta}_k) \end{aligned} \tag{20}$$

which will save a lot of computational costs.

The other parts are just the same as the previous section.

7 Removing the Constraint of $\text{diag}(\boldsymbol{A}) = 0$

We can eliminate the constraint of $\text{diag}(\boldsymbol{A}) = \boldsymbol{0}$. And the solution for each subproblem is obtained by removing the term of $\boldsymbol{A} = \boldsymbol{J} - \text{diag}(\boldsymbol{J})$. The physical meaning is that we can allow each data point to represente itself, so there always exists reasonable solutions even with outliers or when the data is insufficient.

8 Non-negative Least Squares

Motivated by the fact that the non-negativity constraint can automatically maintain the sparsity property, we can remove the regularization term on the NNLSR model. The resulting non-negative least squares (NNLS) model could be formulated as follows:

$$\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^2 \quad \text{s.t.} \quad \boldsymbol{A} \geq 0. \tag{21}$$

This model can have a lot of benefits in complex appliactions. First, it is parameter free, and therefore we can avoid the burden of tuning paramters especially on large datasets. This will make this model suitable for large scale applications. Besides, since the non-negativity can automatically boost the sparsity on the model, the model can be viewed as "Self-Regularized" without any explicit regulariztion on the coding matrix $\boldsymbol{A}$.

This model is also easy to solve. After introducing an auxiliary variable $\boldsymbol{C}$, we can resort to the variable splitting method to reformulate the problem (21) into a inequality constraint problem as:

$$\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^2 \quad \text{s.t.} \quad \boldsymbol{C} = \boldsymbol{A}, \boldsymbol{C} \geq 0. \tag{22}$$

The Lagrangian function of the problem can be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Delta}, \rho) =& \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^2 + \langle \boldsymbol{\Delta}, \boldsymbol{C} - \boldsymbol{A}\rangle \\ &+ \frac{\rho}{2}\|\boldsymbol{C} - \boldsymbol{A}\|_F^2.\end{aligned} \tag{23}$$

This model can be solved by the ADMM algorithm, just similar to those problems in previous sections. Each variable can be updated by taking derivatives of the Lagrangian function w.r.t. the variables $\boldsymbol{A}$ and $\boldsymbol{C}$. We ignore the detail steps and only introduce the major updating steps as follows:

Updating $\boldsymbol{A}$ while fixing $\boldsymbol{C}$:

$$\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^2 + \frac{\rho}{2}\|\boldsymbol{A} - (\boldsymbol{C}_k - \rho^{-1}\boldsymbol{\Delta}_k)\|_2^F. \tag{24}$$

This is a standard LSR problem which can be solved in closed form.

Updating $\boldsymbol{C}$ while fixing $\boldsymbol{A}$:

$$\min_{\boldsymbol{C}} \|\boldsymbol{C} - (\boldsymbol{A} + \rho^{-1}\boldsymbol{\Delta}_k)\|_F^2 \quad \text{s.t.} \quad \boldsymbol{C} \geq 0. \tag{25}$$

This is a standard non-negative least squares (NNLS) constraint problem which can be solved by the active set methods such as the Lawson and Hanson algorithm [15]. In this problem, since the measurement matrix is an identity matrix, this problem is much easier than the standard NNLS problem. In fact, the solution of $\boldsymbol{C}$ can be obtained by setting each element $\boldsymbol{C}_{ij} = \max(0, (\boldsymbol{A} + \rho^{-1}\boldsymbol{\Delta}_k)_{ij})$.

Improvement: We can also add constraints of $\text{diag}(\boldsymbol{C}) = \boldsymbol{0}$ and/or $\boldsymbol{1}^\top \boldsymbol{C} = \boldsymbol{1}^\top$ (each column vector is "sum to one") to further enhance the performance of the NNLS model. This model only has two parameters, i.e., the penalty parameter ρ and the number of iterations, which are quite robust to different tasks.

9 Robust Large Scale Subset Selection via Dissimilarity based Outlier Detection?

We can also introduce a dissimiarlity based matrix $\boldsymbol{D}$ to replace the ℓ_p or $\ell_{2,1}$ norms to ensure robustness. This can also remore the additional term $\boldsymbol{Z}$ on modeling the outliers with the restriction of ℓ_1 norm. The matrix $\boldsymbol{D}$ should better be diagonal matrix. How to design the matrix $\boldsymbol{D}$ is another problem need to be solved.

Then the proposed model can be formulated as

$$\min_{\boldsymbol{A}} \|(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A})\boldsymbol{D}\|_F^2 + \lambda\|\boldsymbol{A}\|_{p,1}. \tag{26}$$

By introducing an auxiliary variable $\boldsymbol{C}$ into the optimization program, we can get

$$\min_{\boldsymbol{A},\boldsymbol{C}} \|(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A})\boldsymbol{D}\|_F^2 + \lambda\|\boldsymbol{C}\|_{p,1} \text{ s.t. } \boldsymbol{C} = \boldsymbol{A}. \tag{27}$$

By introducing a Lagrangian multiplier $\boldsymbol{\Delta}$, the Lagrangian function of the Eq. (35) can be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Delta}, \rho) =& \|(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A})\boldsymbol{D}\|_F^2 + \lambda\|\boldsymbol{C}\|_{p,1} \\ &+ \langle \boldsymbol{\Delta}, \boldsymbol{C} - \boldsymbol{A}\rangle + \frac{\rho}{2}\|\boldsymbol{C} - \boldsymbol{A}\|_F^2\end{aligned} \tag{28}$$

Denote by $(\boldsymbol{A}_k, \boldsymbol{C}_k)$ the optimization variables at iteration k, by $\boldsymbol{\Delta}_k$ the Lagrangian multiplier at iteration k, and by ρ the penalty parameter at iteration k. Taking detivatives of $\mathcal{L}$ with respect to the variables and setting the derivatives to be zeros, we can alternatively update the variables as follows:

(1) Obtain $\boldsymbol{A}_{k+1}$ by minimizing $\mathcal{L}$ with respect to $\boldsymbol{A}$, while fixing $(\boldsymbol{C}_k, \boldsymbol{\Delta}_k)$. This is equivalent to solve the following problem:

$$\min_{\boldsymbol{A}} \|(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A})\boldsymbol{D}\|_F^2 + \frac{\rho}{2}\|\boldsymbol{A} - (\boldsymbol{C}_k - \rho^{-1}\boldsymbol{\Delta}_k)\|_F^2, \tag{29}$$

which is equalivalently to solve the following problem

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}\boldsymbol{D}\boldsymbol{D}^\top + \frac{\rho}{2}\boldsymbol{A} = \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}^\top + \frac{\rho}{2}(\boldsymbol{C}_k - \rho^{-1}\boldsymbol{\Delta}_k) \tag{30}$$

Since the matrices $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{D}^\top\boldsymbol{D}$ are positive semi-definite and positive definite, respectively. The above equation is a standard Sylvester equation which has a unique solution.

(2) Obtain $\boldsymbol{C}_{k+1}$ by minimizing $\mathcal{L}$ with respect to $\boldsymbol{C}$, while fixing $(\boldsymbol{A}_{k+1}, \boldsymbol{\Delta}_k)$. This is equivalent to solve the following problem:

$$\min_{\boldsymbol{C}} \frac{1}{2}\|(\boldsymbol{A}_{k+1} + \rho^{-1}\boldsymbol{\Delta}k) - \boldsymbol{C}\|_F^2 + \frac{\lambda}{\rho}\|\boldsymbol{C}\|_{p,1}. \quad (31)$$

Since the $\ell_{p,1}$ norm is separable with respect to each row, we can write the above problem as

$$\min_{\boldsymbol{C}} \sum_{i=1}^{M} \frac{1}{2}\|(\boldsymbol{A}_{k+1})_{i*} + \rho^{-1}(\boldsymbol{\Delta}_k)_{i*} - \boldsymbol{C}_{i*}\|_2^2 + \frac{\lambda}{\rho}\|\boldsymbol{C}_{i*}\|_p, \quad (32)$$

where $\boldsymbol{F}_{i*}$ is the ith row of the matrix $\boldsymbol{F}$. Since this step is separable w.r.t. each row, we can employ parallel processing resources and reduce its computational time.

(3) Obtain the Lagrangian multipliers $(\boldsymbol{\Delta}_{k+1})$ while fixing $(\boldsymbol{C}_{k+1}, \boldsymbol{A}_{k+1})$:

$$\boldsymbol{\Delta}_{k+1} = \boldsymbol{\Delta}_k + \rho(\boldsymbol{C}_{k+1} - \boldsymbol{A}_{k+1}). \quad (33)$$

(5) Update the penalty parameter ρ as $\rho = \mu\rho$, where $\mu > 1$.

10 Large Scale Subset Selection Via Row-Column Separation

We can also restrict that $\text{diag}(\boldsymbol{A}) = \boldsymbol{0}$ to avoid the samples to be self-represented. However, I want to mention that the proposed model solved by ADMM algorithm with three variables and does not have convergence results.

Then the model above can be

$$\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^2 + \lambda\|\boldsymbol{A}\|_{p,1} \quad \text{s.t.} \quad \text{diag}(\boldsymbol{A}) = \boldsymbol{0}. \quad (34)$$

By introducing an auxiliary variable $\boldsymbol{C}$ into the optimization program, we can get

$$\begin{aligned} &\min_{\boldsymbol{A},\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \lambda\|\boldsymbol{A}\|_{p,1} \\ &\text{s.t. } \boldsymbol{C} = \boldsymbol{A} - \text{diag}(\boldsymbol{A}), \end{aligned} \quad (35)$$

whose solution for $\boldsymbol{A}$ coincides with the solution of Eq. (35). By introducing two Lagrangian multipliers $\boldsymbol{\Delta}$, the Lagrangian function of the Eq. (35) can be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Delta}, \rho) &= \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \lambda\|\boldsymbol{A}\|_{p,1} \\ &+ \langle\boldsymbol{\Delta}, \boldsymbol{C} - (\boldsymbol{A} - \text{diag}(\boldsymbol{A}))\rangle + \frac{\rho}{2}\|\boldsymbol{C} - (\boldsymbol{A} - \text{diag}(\boldsymbol{A})\|_F^2 \end{aligned} \quad (36)$$

Denote by $(\boldsymbol{C}_k, \boldsymbol{A}_k)$ the optimization variables at iteration k, by $\boldsymbol{\Delta}_k$ the Lagrangian multipliers at iteration k, and by ρ the penalty parameter at iteration k. Taking detivatives of $\mathcal{L}$ with respect to the variables and setting the derivatives to be zeros, we can alternatively update the variables as follows:

(1) Obtain $\boldsymbol{A}_{k+1}$ by minimizing $\mathcal{L}$ with respect to $\boldsymbol{A}$, while fixing $(\boldsymbol{C}_k, \boldsymbol{\Delta}_k)$. This is equivalent to solve the following problem:

$$\begin{aligned} &\boldsymbol{A}_{k+1} = \boldsymbol{J} - \text{diag}(\boldsymbol{J}), \\ &\boldsymbol{J} = \arg\min_{\boldsymbol{J}} \frac{1}{2}\|\boldsymbol{C}_k + \rho^{-1}\boldsymbol{\Delta}_k - \boldsymbol{J}\|_F^2 + \frac{\lambda}{\rho}\|\boldsymbol{J}\|_{p,1}. \end{aligned} \quad (37)$$

(2) Obtain $\boldsymbol{C}_{k+1}$ by minimizing $\mathcal{L}$ with respect to $\boldsymbol{C}$, while fixing $(\boldsymbol{A}_{k+1}, \boldsymbol{\Delta}_k)$. This is equivalent to solve the following problem:

$$\min_{\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \frac{\rho}{2}\|\boldsymbol{C} - \boldsymbol{A}_{k+1} + \frac{1}{\rho}\boldsymbol{\Delta}_k\|_F^2 \quad (38)$$

This is a least squares regression problem which has a closed-form solution as

$$\boldsymbol{C}_{k+1} = (\boldsymbol{X}^\top\boldsymbol{X} + \frac{\rho}{2}\boldsymbol{I})^{-1}(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\rho}{2}\boldsymbol{A}_{k+1} - \frac{1}{2}\boldsymbol{\Delta}_k). \quad (39)$$

(3) Obtain the Lagrangian multipliers $(\boldsymbol{\Delta}_{k+1})$ while fixing $(\boldsymbol{C}_{k+1}, \boldsymbol{A}_{k+1})$:

$$\boldsymbol{\Delta}_{k+1} = \boldsymbol{\Delta}_k + \rho(\boldsymbol{C}_{k+1} - \boldsymbol{A}_{k+1}). \quad (40)$$

11 Experiments

In this section, we apply the proposed Non-negative Least Squares Regression (NNLSR) model into subspace clustering, subset selection, and image classification problems.

For these experiments, we fixed the parameters for each dataset, and choose the parameters with the lowest error rates. In fact, if we further tune the parameters for individual case of each dataset, the performance of our method would be even better.

11.1 Subspae Clustering

We evaluate the advantages of the proposed NNLSR method on subspace clustering problem. We compare with the state-of-the-art methods on four datasets including the Hopkins 155 dataset, the Extended Yale B dataset, the USPS and MNIST datasets.

- The Hopkins 155 motion dataset contains 156 video sequences, 155 of which have two or three moving objects, and 1 video sequence has 5 moving onbjects. The motion trajectives of each object is viewed as a subspace.

- Extended Yale B dataset contains 38 human subjects, each subject has around 64 near frontal images taken under different illuminational conditions. The images are projected on a 60 dimensions space by PCA. Each class or subspace includes 64 imgaes which are resized into 32×32 pixels.

- The USPS contains 9298 images for digit numbers form 0 to 9. Each of the image in USPS is resized into 16×16 pixels. We use all the images in USPS for experiments.

- The MNIST contains 60000 training images for digit numbers form 0 to 9.

11.1.1 Motion Segmentation

11.1.2 Face Clustering

11.2 Subste Selection

11.3 Image Classification

We compare the proposed method with ScSPM [] and LLC [] on UIUC psorts dataset [] and Scene15 [] dataset.

11.4 Complexity Analysis and Comparison

12 Future Work

NNLSR or ANNLSR should have closed-fprm solutions. However, these two problems can also be solved by ADMM algorithm in an iterative manner. So we can compare the solutions obtaied by closed-form and ADMM to evaluate the quality of the ADMM algorithm proposed.

We can also remove the regularization term and propose a "self-regularized"' model with non-negativity and affinity constraints. We believe this new model will share the advantages of the ANNLSR model with several better properties for subspace clustering and image claasification.

References

[1] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. pages 2790–2797. IEEE, 2009. 1

[2] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013. 1

[3] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670, 2010. 1

[4] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on*

Table 1: Clustering error (%) of different algorithms on the Hopkins 155 dataset [**?**] with the 12-dimensional data points obtained by applying PCA.

Algorithms	**SSC[?]**	**LRR[?]**	**LRSC[?]**	**LSR1**[12]	**NNLSR**	**NPLSR**
Mean	2.18	2.56		2.97		
Median	0.00	0.00	0.00	0.28		
Time						
Algorithms	**NNLSRd0**	**NPLSRd0**	**ANNLSR**	**ANPLSR**	**ANNLSRd0**	**ANPLSRd0**
Mean						
Median	0.00		0.00			
Time						

Pattern Analysis and Machine Intelligence, 35(1):171–184, 2013. 1

[5] Liansheng Zhuang, Haoyuan Gao, Zhouchen Lin, Yi Ma, Xin Zhang, and Nenghai Yu. Non-negative low rank and sparse graph for semi-supervised learning. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2328–2335, 2012. 1

[6] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 1

[7] Martin Slawski, Matthias Hein, et al. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7:3004–3056, 2013. 1

[8] Seung-Jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE journal of selected topics in signal processing*, 1(4):606–617, 2007. 1

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, January 2011. 2

[10] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. *Advances in Neural Information Processing Systems 24*, pages 612–620, 2011. 2

[11] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, 1st edition, 2006. 2

[12] Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. *ECCV*, pages 347–360, 2012. 2, 3, 7, 9, 10

[13] Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? *ICCV*, pages 471–478, 2011. 2, 3

[14] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006. 4

[15] Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995. 5

Table 2: Clustering error (%) of different algorithms on the Extended Yale B dataset [] with the 12-dimensional data points obtained by applying PCA.

No. of Subject	Algorithms	**SSC**[**?**]	**LRR**[**?**]	**LRSC**[**?**]	**LSR2**[12]	**NNLSR**	**NPLSR**	**ANNLSR**	**ANPLSR**
2	Mean	1.85	2.13						
(163 Instances)	Median								
	Time								
3	Mean								
(416 Instances)	Median								
	Time								
5	Mean								
(812 Instances)	Median								
	Time								
8	Mean								
(136 Instances)	Median								
	Time								
10	Mean								
(3 Instances)	Median								
	Time								

Table 3: Accracy (%) of different algorithms as a function of the frction of selected samples from each class (η) on the MNIST dataset [**?**].

Algorithms	**SSC**[**?**]	**LRR**[**?**]	**LRSC**[**?**]	**LSR**[12]
Mean	2.18			
Median				
Max				
Time				
Algorithms	**NNLSR**	**NPLSR**	**ANNLSR**	**ANPLSR**
Mean				
Median				
Max				
Time				

Table 4: Clustering error (%) of different algorithms on the Hopkins 155 dataset [**?**] with the 12-dimensional data points obtained by applying PCA.

Algorithms	**SSC**[**?**]	**LRR**[**?**]	**LRSC**[**?**]	**LSR**[12]
Mean	2.18			
Median				
Max				
Time				
Algorithms	**NNLSR**	**NPLSR**	**ANNLSR**	**ANPLSR**
Mean				
Median				
Max				
Time				