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

June 19, 2017


## Abstract


## 1 Introduction

The sparse subspace clustering (SSC) [**?** ] and low rank representation (LRR) [**?** ] 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.

### 1.1 Physically correct! Not only mathematically correct!

In fact, non-negativity is more reasonable with the biological modeling of the visual data and often lead to better performance for data representation [] and graph construction []. 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.

Non-negatie least squares has been studied in [1] for sparse recovery without regularization. The authors compare with the non-negative LASSO method [2] 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 [**?** ]. To make the method more efficient and scalable, we can also employ the linearized ADMM with adaptive penalty (LADMAP) [**?** ], 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 [3], 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 [4] proposed by Lu et al. can be formulated as follows:

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

Here we denote by $\mathrm{diag}(\boldsymbol{A})$ both a diagonal matrix whose diagonal elements are the diagonal entries of $\boldsymbol{A}$ and the vector consisted of the diagonal elements. According to [4], the above problem has the optimal solution as

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

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

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

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

The LSR model can be reformulated as a collaborative representation model [5] for subspace clustering with an additional constraint of $\mathrm{diag}(\boldsymbol{A})=\boldsymbol{0}$. The constraint of $\mathrm{diag}(\boldsymbol{A})=\boldsymbol{0}$ 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

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

whose solution for $\boldsymbol{A}$ coincides with the solution of Eq. (1). By introducing a Lagrangian multipliers $\boldsymbol{\Delta}$ and a penalty parameter $\rho$, the Lagrangian function of the Eq. (30) can be written as

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

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 $\rho$ 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:

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

(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 $\mathrm{diag}(\boldsymbol{A})=\boldsymbol{0}$

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. [4].

The LSR model in [4] can be reformulated as a collaborative representation model [5] 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 $\rho$, the Lagrangian function of the Eq. (30) 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 [6] 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

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

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 $\text{diag}(A) = 0$

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 Robust Large Scale Subset Selection via Dissimilarity based Outlier Detection?

We can also introduce a dissimiarlity based matrix $\boldsymbol{D}$ to replace the $\ell_p$ 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 $\ell_1$ 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

$$\min_{\boldsymbol{A}} \|(\boldsymbol{X} - \boldsymbol{XA})\boldsymbol{D}\|_F^2 + \lambda\|\boldsymbol{A}\|_{p,1}. \tag{21}$$

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

$$\min_{\boldsymbol{A},\boldsymbol{C}} \|(\boldsymbol{X} - \boldsymbol{XA})\boldsymbol{D}\|_F^2 + \lambda\|\boldsymbol{C}\|_{p,1} \text{ s.t. } \boldsymbol{C} = \boldsymbol{A}. \tag{22}$$

By introducing a Lagrangian multiplier $\boldsymbol{\Delta}$, the Lagrangian function of the Eq. (30) can be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Delta}, \rho) =& \|(\boldsymbol{X} - \boldsymbol{XA})\boldsymbol{D}\|_F^2 + \lambda\|\boldsymbol{C}\|_{p,1} \\ &+ \langle\boldsymbol{\Delta}, \boldsymbol{C} - \boldsymbol{A}\rangle + \frac{\rho}{2}\|\boldsymbol{C} - \boldsymbol{A}\|_F^2 \end{aligned} \tag{23}$$

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 $\rho$ 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:

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

which is equalivalently to solve the following problem

$$\boldsymbol{X}^\top\boldsymbol{XADD}^\top + \frac{\rho}{2}\boldsymbol{A} = \boldsymbol{X}^\top\boldsymbol{XDD}^\top + \frac{\rho}{2}(\boldsymbol{C}_k - \rho^{-1}\boldsymbol{\Delta}_k) \tag{25}$$

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:

$$\min_{\boldsymbol{C}} \frac{1}{2}\|(\boldsymbol{A}_{k+1} + \rho^{-1}\boldsymbol{\Delta}k) - \boldsymbol{C}\|_F^2 + \frac{\lambda}{\rho}\|\boldsymbol{C}\|_{p,1}. \quad (26)$$

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

$$\min_{\boldsymbol{C}} \sum_{i=1}^{M} \frac{1}{2}\|(\boldsymbol{A}_{k+1})_{i*} + \rho^{-1}(\boldsymbol{\Delta}_k)_{i*} - \boldsymbol{C}_{i*}\|_2^2 + \frac{\lambda}{\rho}\|\boldsymbol{C}_{i*}\|_p, \quad (27)$$

where $\boldsymbol{F}_{i*}$ is the $i$th 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 $(\boldsymbol{C}_{k+1}, \boldsymbol{A}_{k+1})$:

$$\boldsymbol{\Delta}_{k+1} = \boldsymbol{\Delta}_k + \rho(\boldsymbol{C}_{k+1} - \boldsymbol{A}_{k+1}). \quad (28)$$

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

# 9 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

$$\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\|_F^2 + \lambda\|\boldsymbol{A}\|_{p,1} \quad \text{s.t.} \quad \text{diag}(\boldsymbol{A}) = \boldsymbol{0}. \quad (29)$$

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

$$\begin{aligned} &\min_{\boldsymbol{A},\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \lambda\|\boldsymbol{A}\|_{p,1} \\ &\text{s.t. } \boldsymbol{C} = \boldsymbol{A} - \text{diag}(\boldsymbol{A}), \end{aligned} \quad (30)$$

whose solution for $\boldsymbol{A}$ coincides with the solution of Eq. (30). By introducing two Lagrangian multipliers $\boldsymbol{\Delta}$, the Lagrangian function of the Eq. (30) can be written as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Delta}, \rho) = \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \lambda\|\boldsymbol{A}\|_{p,1} \\ &+ \langle\boldsymbol{\Delta}, \boldsymbol{C} - (\boldsymbol{A} - \text{diag}(\boldsymbol{A}))\rangle + \frac{\rho}{2}\|\boldsymbol{C} - (\boldsymbol{A} - \text{diag}(\boldsymbol{A})\|_F^2 \end{aligned} \quad (31)$$

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 $\rho$ 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:

$$\begin{aligned} &\boldsymbol{A}_{k+1} = \boldsymbol{J} - \text{diag}(\boldsymbol{J}), \\ &\boldsymbol{J} = \arg\min_{\boldsymbol{J}} \frac{1}{2}\|\boldsymbol{C}_k + \rho^{-1}\boldsymbol{\Delta}_k - \boldsymbol{J}\|_F^2 + \frac{\lambda}{\rho}\|\boldsymbol{J}\|_{p,1}. \end{aligned} \quad (32)$$

(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:

$$\min_{\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \frac{\rho}{2}\|\boldsymbol{C} - \boldsymbol{A}_{k+1} + \frac{1}{\rho}\boldsymbol{\Delta}_k\|_F^2 \quad (33)$$

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

$$\boldsymbol{C}_{k+1} = (\boldsymbol{X}^\top\boldsymbol{X} + \frac{\rho}{2}\boldsymbol{I})^{-1}(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\rho}{2}\boldsymbol{A}_{k+1} - \frac{1}{2}\boldsymbol{\Delta}_k). \quad (34)$$

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

$$\boldsymbol{\Delta}_{k+1} = \boldsymbol{\Delta}_k + \rho(\boldsymbol{C}_{k+1} - \boldsymbol{A}_{k+1}). \quad (35)$$

# 10 Experiments

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

### 10.1 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 $32 \times 32$ pixels.
- The USPS contains 9298 images for digit numbers form 0 to 9. Each of the image in USPS is resized into $16 \times 16$ pixels. We use all the images in USPS for experiments.
- The MNIST contains 60000 training images for digit numbers form 0 to 9.

### 10.2 Subste Selection

### 10.3 Image Classification

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

### 10.4 Complexity Analysis and Comparison

## References


[1] 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

[2] Seung-Jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky. An interior-point method for large-scale $\ell_1$-regularized least squares. *IEEE journal of selected topics in signal processing*, 1(4):606–617, 2007. 1

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

[4] 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

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

[6] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006. 3