

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

June 17, 2017

Abstract

1 Introduction

Non-negative 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 recovery problems. For $\mathbf{A} = (\mathbf{a}_{ij})$ we write $\mathbf{A} \geq 0$ if $\mathbf{a}_{ij} \geq 0$ for each i and j and we say \mathbf{A} is a non-negative matrix. This notation can be naturally extended for vectors. Similarly we can define non-positive matrix, negative matrix, and negative matrix.

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-representation while negative supports are negative to self-representation.
- Better performance. Faster?

3 LSR Model

It is widely known that the least squares regression (LSR) model [3], which can be expressed as follows:

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{X}\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_F^2, \quad (1)$$

has a form of closed solution as

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

In subspace clustering community, the LSR method [4] proposed by Lu et al. can be formulated as follows:

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

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

$$\hat{\mathbf{A}} = -\mathbf{Z}(\text{diag}(\mathbf{Z})) \text{ s.t. } \text{diag}(\hat{\mathbf{A}}) = \mathbf{0}, \quad (4)$$

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

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

4 Collaborative Representation based Clustering with Constraint $\text{diag}(\mathbf{A}) = \mathbf{0}$

The LSR model can be reformulated as a collaborative representation model [5] for subspace clustering with an additional constraint of $\text{diag}(\mathbf{A}) = \mathbf{0}$. The constraint of $\text{diag}(\mathbf{A}) = \mathbf{0}$ is used to avoid the samples to represent themselves.

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

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

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

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

Denote by $(\mathbf{C}_k, \mathbf{A}_k)$ the optimization variables at iteration k , by $\mathbf{\Delta}_k$ the Lagrangian multipliers at iteration k , and by ρ the penalty parameter at iteration k . Taking derivatives 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 \mathbf{A}_{k+1} by minimizing \mathcal{L} with respect to \mathbf{A} , while fixing $(\mathbf{C}_k, \mathbf{\Delta}_k)$. This is equivalent to solve the following problem:

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

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

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

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

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

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

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

Convergency analysis?

5 Non-Negative Collaborative Representation

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 $\text{diag}(\mathbf{A}) = \mathbf{0}$. In this section, we want to mention that the coefficient matrix \mathbf{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 subspace 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 \mathbf{A} that the elements in \mathbf{A} should be non-negative, i.e., $\mathbf{A} \succeq 0$. Hence, the proposed non-negative collaborative representation model can be formulated as follows:

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

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

By introducing auxiliary variables into the optimization program, we can set $\mathbf{C} = \mathbf{A}$. The LSR model (11) can be transformed into

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

whose solution for \mathbf{A} coincides with the solution of Eq. (9). By introducing a Lagrangian multipliers $\mathbf{\Delta}$ and a penalty parameter ρ , the Lagrangian function of the Eq. (30) can be written as

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

After initializing the $\mathbf{A}, \mathbf{C}, \mathbf{\Delta}$ as zero matrices, the ADMM algorithm iterates consist of 1) minimizing \mathcal{L}

with respect to \mathbf{A} while fixing the other variables, and 2) minimizing \mathcal{L} with respect to \mathbf{C} subject to the constraint $\mathbf{C} \succeq 0$ while fixing the other variables; 3) updating the Lagrangian variable Δ while fixing the other variables. Specifically, denote by $(\mathbf{C}_k, \mathbf{A}_k)$ the optimization variables at iteration k , by Δ_k the Lagrangian multipliers at iteration k . Taking derivatives 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 \mathbf{A}_{k+1} by minimizing \mathcal{L} with respect to \mathbf{A} , while fixing (\mathbf{C}_k, Δ_k) . This is equivalent to solve the following problem:

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

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

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

This is a non-negative least squares problem which can be solved by many solves developed via active set method [6].

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

$$\Delta_{k+1} = \Delta_k + \rho(\mathbf{C}_{k+1} - \mathbf{A}_{k+1}). \quad (16)$$

Convergency analysis?

The above solution is slow when the number of column of \mathbf{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 \mathbf{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

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{U} (\mathbf{C}^{-1} + \mathbf{V} \mathbf{A}^{-1} \mathbf{U})^{-1} \mathbf{V} \mathbf{A}^{-1}. \quad (17)$$

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

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

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

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

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

and transform this problem as

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

which will save a lot of computational costs.

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

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

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

Then the proposed model can be formulated as

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

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

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

By introducing a Lagrangian multiplier Δ , the Lagrangian function of the Eq. (30) can be written as

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

Denote by $(\mathbf{A}_k, \mathbf{C}_k)$ the optimization variables at iteration k , by Δ_k the Lagrangian multiplier at iteration k , and by ρ the penalty parameter at iteration k . Taking derivatives 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 \mathbf{A}_{k+1} by minimizing \mathcal{L} with respect to \mathbf{A} , while fixing (\mathbf{C}_k, Δ_k) . This is equivalent to solve the following problem:

$$\min_{\mathbf{A}} \|(\mathbf{X} - \mathbf{X}\mathbf{A})\mathbf{D}\|_F^2 + \frac{\rho}{2} \|\mathbf{A} - (\mathbf{C}_k - \rho^{-1}\Delta_k)\|_F^2, \quad (24)$$

which is equivalently to solve the following problem

$$\mathbf{X}^\top \mathbf{X} \mathbf{A} \mathbf{D} \mathbf{D}^\top + \frac{\rho}{2} \mathbf{A} = \mathbf{X}^\top \mathbf{X} \mathbf{D} \mathbf{D}^\top + \frac{\rho}{2} (\mathbf{C}_k - \rho^{-1}\Delta_k) \quad (25)$$

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

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

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

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

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

where \mathbf{F}_{i*} is the i th row of the matrix \mathbf{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 (Δ_{k+1}) while fixing $(\mathbf{C}_{k+1}, \mathbf{A}_{k+1})$:

$$\Delta_{k+1} = \Delta_k + \rho(\mathbf{C}_{k+1} - \mathbf{A}_{k+1}). \quad (28)$$

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

8 Large Scale Subset Selection Via Row-Column Separation

We can also restrict that $\text{diag}(\mathbf{A}) = \mathbf{0}$ 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

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

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

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

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

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

Denote by $(\mathbf{C}_k, \mathbf{A}_k)$ the optimization variables at iteration k , by Δ_k the Lagrangian multipliers at iteration k , and by ρ the penalty parameter at iteration k . Taking derivatives 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 \mathbf{A}_{k+1} by minimizing \mathcal{L} with respect to \mathbf{A} , while fixing $(\mathbf{C}_k, \mathbf{\Delta}_k)$. This is equivalent to solve the following problem:

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

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

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

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

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

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

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

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.
- [2] Seung-Jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dmitry 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.
- [3] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, 1st edition, 2006.
- [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.
- [5] Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? *ICCV*, pages 471–478, 2011.

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