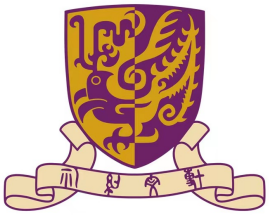
DDA4300 Final Project

120090715 Runze ZHAO
120090811 Yiwen LU

Submitted to
Prof. YE and the TA team of DDA4300



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Chinese University of Hongkong, Shenzhen
DATE SUBMITTED
[Latest Revised: April 25, 2023]

1 Introduction

1.1 Introduction to spectral clustering

Spectral clustering[7][11] is a prominent technique in the field of machine learning, particularly for unsupervised learning tasks, such as data clustering and image segmentation. The primary objective of this method is to partition data points into distinct groups or clusters to optimize intra-cluster similarity and inter-cluster dissimilarity[10]. Spectral clustering is a graph-based approach that utilizes eigenvectors of the similarity matrix to project data points into a lower-dimensional space, allowing for more effective separation of complex data structures and non-convex clusters.

Unlike centroid-based methods such as k-means clustering, which make assumptions about cluster shapes and sizes, spectral clustering can handle a variety of cluster shapes and structures. Therefore, it is considered a more versatile and robust solution to data clustering and image segmentation problems, especially when dealing with complex and non-convex datasets.

1.2 Paradigm to deal with spectral clustering

Briefly speaking, what spectral clustering do is to learn the label matrix Y from dataset X, with the key 'by-products' shown as follows:

$$X \to \phi(X) \to A \to L \to Y$$

where $\phi(X)$ is the data matrix after kernel transformation, A is the affinity matrix, L is the Laplacian matrix. We have summarized the symbols and formula in the appendix A.1 and A.2 respectively. The real process of spectral clustering can be roughly divided into two steps[2]:

i. Construct an affinity matrix A: Each element of the matrix A represents the similarity between two data points. Two predominant approaches are used to accomplish this idea.

   a. The first approach entails constructing the similarity matrix from the data matrix X using conventional transformations, such as the k-nearest neighbors (KNN) method and the Gaussian kernel method ($k(x, y) = \exp(-||x - y||^2/2\sigma^2)$). Here k, σ is hyper-parameter.

   b. The second approach involves adopting a self-expressive model [1] that learns the affinity matrix by utilizing an optimization function. Some of the previous researchers adopted Low-Rank Representation (LRR)[4] to deal with the task, and some used the least-square representation model (LSR) [5] to deal with the problem.

ii. Find the optimal cluster results Y based on A: Former researchers usually

   a. perform Eigenvalue Decomposition (EVD) on affinity matrix A and get the first c eigenvectors;

   b. apply k-means clustering on the first c eigenvectors. Here c refers to the number of clusters.

   Generally, the "optimality" of such divisions is evaluated using the normalized cut (NCuts)[10] and its variants. For example, some researchers deal with relative eigen-gap to improve the performance[2]. Additionally, some researchers[8] have summarized the process into two stages:

   a. Solve the optimization function $\underset{F^TF=I,W,b}{min} tr(F^T\tilde{L}F)$ to learn a relaxed cluster assignment F. Here $\tilde{L}$ represents the normalized Laplacian matrix.

   b. Use k-means clustering or spectral rotation on F to determine the cluster assignment.

2 Relation between our works and others'

We followed the main idea of conventional LSR Spectral Clustering, but replace the the regularization term with p-norm in constructing affinity matrix A.

2.1 Least Square Representation model (LSR)

The Least Square Representation model (LSR)[5] technique employs the method of least squares to build a model using the self-expressive process described by the equation

$$\min_{C} ||X - XC||_F^2 + \lambda R(C)$$

where λ is hyper-parameter and $R(C)$ denotes the regularization term. However, in the case of data exhibiting a non-linear relationship, linear regression cannot accurately capture the similarity between the pairwise columns of X. To address this, we can modify the equation by incorporating a kernel function $\phi(.)$, such that it becomes

$$\min_{C} ||\phi(X) - \phi(X)C||_F^2 + \lambda R(C)[2]$$

The kernel functions $\phi(.)$ is used in preprocessing to transform the data.

Our research differ from the original work by applying the Schatten p-norm ($0 < p < 1$) as a regularization term: $\lambda||C||_p^p$. By experimental results, this approach is suitable for capturing the non-linear relationship inherent in the data with the use of a kernel function.

2.2 p-norm

There are limited references available on matrix p-norm that can be utilized as a foundation for our study. Some existing literature has used the p-norm in step 2 of 1.2 within the framework of Low-Rank Representation (LRR) [3], but p-norm are not applied to the affinity learning.

As commonly acknowledged, the computation of the p-norm is challenging due to its non-convexity (for $0 < p < 1$), which may lead to sub-optimal solutions if the optimization algorithm becomes trapped in a local minimum rather than discovering the global minimum. In this paper, however, we employ two variants of the p-norm: the proximal p-norm and the Schatten p-norm. Doing so establishes a p-norm-constrained self-expressive approach as a theoretical framework for future research endeavors.

2.3 Summary our main contribution

This report presents a concise analysis of two methodologies that incorporate the p-norm as a regularization term during the data preprocessing phase of spectral clustering. The derivation processes for employing the proximal p-norm and Schatten p-norm within the affinity learning stage of spectral clustering are elucidated. Owing to computational constraints, only the performance of the Schatten p-norm can be assessed. Nonetheless, a functional spectral clustering framework is provided for future researchers, which displays encouraging outcomes. Acknowledging that this research domain has yet to be thoroughly investigated within the context of spectral clustering is crucial.

3 approach using proximal p-norm

The problem can be formulated as follows:

$$\min_{c} ||\phi(X) - \phi(X)C||_F^2 + \lambda||C||_p^p$$

$$s.t.\ 0 \leq C \leq 1$$

where the proximal p-norm is expressed as $||C||_p^p = \sum_{i=1}^{n}\sum_{j=1}^{n}|C_{ij}|^p$. The data matrix $X \in \mathbb{R}^{m\times n}$ represents the dataset, with m denoting the dimension of features, n signifying the number of observations, and p indicating the power of the norm (with $0 < p < 1$). The kernel function, $\phi(.)$, is represented by the Gaussian kernel, such that $\phi(X)_{ij} = exp(-||X_i - X_j||^2/2\sigma^2)$, where σ is a hyperparameter.

Following the framework of Least Square Representation (LSR), the output is $C^* \in \mathbb{R}^{n\times n}$. Subsequently, the affinity matrix A is constructed as $A = (C + C^T)/2 \in \mathbb{R}^{n\times n}$. The Laplacian matrix $\tilde{L}$ is formulated as $L = I - D^{-1/2}AD^{-1/2}$, where D refers to the degree matrix derived from the affinity matrix using $D = diag(A \cdot e)$, and $e = [1; 1; \ldots; 1]$.

To generate the algorithm, we employ the concept of the Alternating Direction Method of Multipliers (ADMM) to iteratively obtain the optimized result as follows (for a detailed explanation, please refer to the content in appendixB.1):

Algorithm 1 Updating proximal p-norm using ADMM

Input: $X, \lambda, p, \rho, \gamma, \alpha$, tolerance Initialize $C^{(0)}, Z^{(0)}, U^{(0)}$ **while** *not converged* **do**

C-update: $C^{(k+1)} \leftarrow (2X^TX + \frac{1}{\rho}I_n)^{-1}(2X^TX + U^k + \frac{1}{\rho}Z^k)$

Z-update: Use a solver to compute Z: $\frac{\partial L}{\partial z_{ij}} = \lambda pzij^{p-1} + \gamma I(z_{ij} - 1) + \frac{1}{\rho}(z_{ij} - c_{ij}) + u_{ij} = 0$

U-update: $U^{k+1} = U^k + \rho(Z^{k+1} - C^{k+1})$

Convergence: if (primal) $C^{t+1} - Z^{t+1} \leq$ tol and (dual) $\rho(Z^{t+1} - Z^t) \leq$ tol: **break**

end

Nonetheless, the computation proves to be too slow due to the following aspects:

- Matrix inversion: Inverting $(2X^TX + \frac{1}{\rho}I_n)^{-1}$ has a computational complexity of $O(n^3)$.
- Solver for Z-update: The golden-section solver is chosen to update Z_{ij}, resulting in an approximate computational complexity of $\log(\frac{1}{\epsilon})$ for each entry [6], which is quite high.
- Potential complex solutions may arise when performing the element-wise update.

4 approach using Schatten p-norm

The Schatten p-norm is defined as $||X||_{S_p} = (\sum_{i=1}^{n} \sigma_i^p(X))^{1/p}$, where $\sigma_i(X)$ represents the i^{th} singular value of X (obtained through singular value decomposition).

Drawing from prior research[9], for $\frac{1}{2} < p < 1$ and $d \geq rank(C)$, the following holds true:

$$||C||_p = \min_{U\in R^{n\times d}, V\in R^{n\times d}, C=UV^T} \frac{||U||_F^2 + ||V||_F^2}{2}$$

Thus, for $\frac{1}{2} < p < 1$, we define $L = ||\phi(X) - \phi(X)UV^T||_F^2 + \frac{\lambda}{2}||U||_F^2 + \frac{\lambda}{2}||V||_F^2$ as a relaxation of the previous optimization problem without loss of generality. Employing gradient descent, we design the algorithm as follows (for more details, please refer to the content in appendixB.2):

Algorithm 2 Update Schatten p-norm using decomposition

repeat

Update U using the grdient descent:

$$\nabla_U L = 2\phi(X)^T\phi(X)(UV^T - I)V + \lambda U$$

Update V using the closed-form solution for V:

$$V = 2\phi(X)^T\phi(X)U(2U^T\phi(X)^T\phi(X)U + \lambda I)^{-1}$$

until *until convergence:* $||U - U_{old}||_F^2 < tol$;

return *Affinity matrix:* $A = (C + C^T)/2$

5 Performance

We perform a comparison between various p-norm spectral clustering algorithms, conventional spectral clustering algorithms, and similar algorithms using generated data:

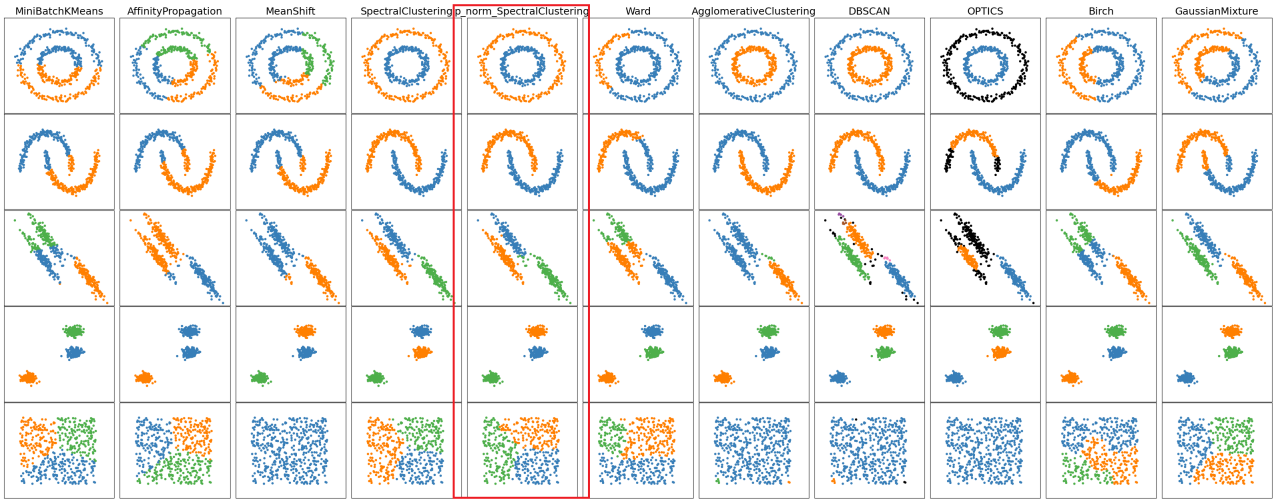


Figure 1: Comparison for large-scale data

Our findings indicate that our algorithm performs effectively on a sample size of 500. Moreover, when applied to small-scale data (with a sample size of 100), our algorithm demonstrates superior performance compared to conventional spectral clustering:

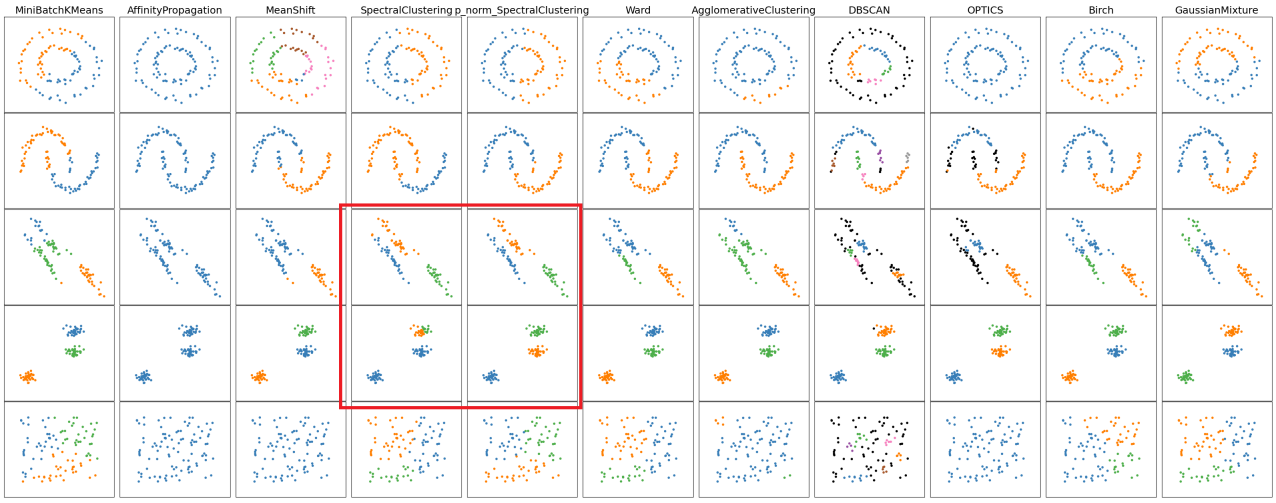


Figure 2: Comparison for small-scale data

6 Conclusion

In summary, we have proposed a p-norm spectral clustering algorithm. Our experiments demonstrate that this algorithm performs effectively on datasets of average size and yields superior results when applied to small-scale data, making it suitable for providing a robust cold start for spectral clustering models. Also, it is essential to note that our work is limited to the Least Squares Representation (LSR) form of spectral clustering, as there are various alternative approaches to spectral clustering yet considered in this study.

References

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

[2] Fan, J., Tu, Y., Zhang, Z., Zhao, M., and Zhang, H. A simple approach to automated spectral clustering. *Advances in Neural Information Processing Systems 35* (2022), 9907–9921.

[3] Lerman, G., and Zhang, T. Robust recovery of multiple subspaces by geometric lp minimization.

[4] Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence 35*, 1 (2012), 171–184.

[5] Lu, C.-Y., Min, H., Zhao, Z.-Q., Zhu, L., Huang, D.-S., and Yan, S. Robust and efficient subspace segmentation via least squares regression. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12* (2012), Springer, pp. 347–360.

[6] Luenberger, D. G. Linear and nonlinear programming second edition. *Columbus, Ohio: Addison-Wesley* (1984).

[7] Ng, A., Jordan, M., and Weiss, Y. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems 14* (2001).

[8] Nie, F., Xu, D., Tsang, I. W., and Zhang, C. Spectral embedded clustering. In *IJCAI* (2009), pp. 1181–1186.

[9] Shang, F., Liu, Y., and Cheng, J. Unified scalable equivalent formulations for schatten quasi-norms. *arXiv preprint arXiv:1606.00668* (2016).

[10] Shi, J., and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence 22*, 8 (2000), 888–905.

[11] Weiss, Y. Segmentation using eigenvectors: a unifying view. In *Proceedings of the seventh IEEE international conference on computer vision* (1999), vol. 2, IEEE, pp. 975–982.

A Summary of symbols and formula

A.1 Symbol used for this algorithm

- X: data matrix, $X \in R^{n \times m}$, where n refers to the number of observations, m refers to the number of features of samples.
- C: similarity matrix learned using the self-expressive algorithm. $C \in R^{n \times n}$
- A: affinity matrix, which performs spectral clustering. $A = (C + C^T)/2 \in R^{n \times n}$
- D: Degree matrix, constructed from the affinity matrix by $D = diag(A \cdot e)$
- L: Laplacian matrix, $L \in R^{n \times n}$
- F: Continuous indication matrix. $F \in R^{n \times c}$

A.2 Formula used for this algorithm

- $||.||_F^2$ refers to the Frobenius norm, where $||X||_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}^2$ for matrix $X \in R^{n \times m}$
- Proximal p-norm: $||X||_p^p = \sum_{i=1}^{n} \sum_{j=1}^{m} |X_{ij}|^p$ for matrix $X \in R^{n \times m}$
- Schatten-p quasi-norm: $||X||_{S_p} = (\sum_{i=1}^{n} \sigma_i^p(X))^{1/p}$, where $\sigma_i(X)$ denotes the i^{th} singular value of X (approached by singular value decomposition)
- $\phi(.)$ is a kernel function, here we adopt Gaussian kernel, where $\phi(X)_{ij} = exp(-||X_i - X_j||^2/2\sigma^2)$, σ being hyper-parameter. $\phi(.) : \mathbb{R}^{n \times m} \mapsto \mathbb{R}^{n \times n}$

B Detailed Derivations

B.1 Derivation for proximal p-norm

- Augmented Lagrangian is proposed as:

$$\mathcal{L}_\rho(C, Z, U) = ||\phi(X) - \phi(X)C||_F^2 + \lambda||Z||_p^p + \gamma \sum_{i,j} max(z_{ij} - 1, 0)^2 + Tr(U^T(Z - C)) + \frac{1}{2\rho}||Z - C||_F^2$$

  Here, U is the dual variable, denoted as the Lagrange multiplier and $\rho > 0$ is a penalty parameter.

- We then propose the ADMM process as follows:

  * C-update: Update the variable C by minimizing the augmented Lagrangian with respect to C:

$$\begin{aligned} C^{k+1} &= arg\min_C L(C, Z^k, U^K) \\ &= (2\phi(X)^T\phi(X) + \frac{1}{\rho}I_n)^{-1}(2\phi(X)^T\phi(X) + U^k + \frac{1}{\rho}Z^k) \end{aligned}$$

  * Z-update: Unfortunately, finding a closed-form solution for Z-update is not straightforward due to the non-convex nature of the $||Z||_p^p$ term and the penalty term. Thanks to the format of our chosen objective function, we can relax the matrix optimization to an element-wise optimization without loss of generality. Below, we show our deriving process.

$$\begin{aligned} L_2 &= \lambda||Z||_p^p + \gamma \sum_{i,j} max(z_{ij} - 1, 0) + Tr(U^T(Z - C)) + \frac{1}{2\rho}||Z - C||_F^2 \\ &= \lambda \sum_{i,j}(z_{ij})^p + \gamma \sum_{i,j} max(z_{ij} - 1, 0) + \sum_{i,j} \frac{1}{2\rho}(z_{ij} - c_{ij})^2 + u_{ij}(z_{ij} - c_{ij}) \\ &= \sum_{i,j}(\lambda z_{ij}^p + \gamma I(z_{ij} - 1) + \frac{1}{2\rho}(z_{ij} - c_{ij})^2 + u_{ij}(z_{ij} - c_{ij})) \end{aligned}$$

$$\frac{\partial L}{\partial Z} = \lambda p z_{ij}^{p-1} + \gamma I(z_{ij} - 1) + \frac{1}{\rho}(z_{ij} - c_{ij}) + u_{ij} = 0$$

  Using solver to solve.

  * U-update:

$$U^{k+1} = U^k + \rho(Z^{k+1} - C^{k+1})$$

B.2 Derivation for Schatten p-norm

We first take the derivative of L to V and U, respectively.

1. Derivative to U:

$$\frac{\partial L}{\partial U} = 2\phi(X)^T\phi(X)(UV^T - I)V + \lambda U$$

2. Derivative to V:

$$\frac{\partial L}{\partial V} = 2VU^T\phi(X)^T\phi(X)U - 2\phi(X)^T\phi(X)U + \lambda V$$

Following the first-order necessary condition, we can derive that the close form solution for V is

$$V = 2\phi(X)^T\phi(X)U(2U^T\phi(X)^T\phi(X)U + \lambda I)^{-1}$$

Unfortunately, close form solution for U is hard to derive straightly. Thus, we borrow the idea of GD and construct the following algorithm to update the value of U.