An Improved Latent Low Rank Representation for Automatic Subspace Clustering

Ya-nan Han, Jian-wei Liu and **Xiong-lin Luo**
Department of Automation, China University of Petroleum, Beijing Campus(CUP), Beijing, China
liujw@cup.edu.cn

Abstract

There is growing interest in low rank representation (LRR) for subspace clustering. Existing latent LRR methods can exploit the global structure of data when the observations are insufficient and/or grossly corrupted, but it cannot capture the intrinsic structure due to the neglect of the local information of data. In this paper, we proposed an improved latent LRR model with a distance regularization and a non-negative regularization jointly, which can effectively discover the global and local structure of data for graph learning and improve the expression of the model.Then, an efficiently iterative algorithm is developed to optimize the improved latent LRR model. In addition, traditional subspace clustering characterizes a fixed numbers of cluster, which cannot efficiently make model selection. An efficiently automatic subspace clustering is developed via the bias and variance trade-off, where the numbers of cluster can be automatically added and discarded on the fly.

1 Introduction

Subspace clustering has been widely studied in the fields of computer vision and machine learning, such as motion segmentation , face segmentation and image clustering [Lai *et al*., 2017]. Vidal[Vidal, 2011] roughly divides subspace clustering algorithms into four groups: algebraic, statistical learning, iterative learning , and spectral clustering based methods. Spectral clustering based methods construct a similar matrix between different data points, and then apply spectral clustering to segment the samples. Spectral clustering can be viewed as the graph-based clustering because its performance is directly determined by the obtained graph[Jie *et al*., 2018]. As a successful example of spectral clustering, LRR is developed to learn the lowest-rank representation of data samples[Liu *et al*., 2010]. After that, a series of improvement approaches based on LRR are proposed, and have achieved certain success[Liu *et al*., 2011;Zhuang *et al*.,2017]. However, existing search methods still have clear limitations. For example, 1)LRR can effectively discovers the global structure of data. However, it is inability to identify local structure, which can affect the expression of the model;2)Usually, the observed sample matrix itself is selected as the dictionary, but it is inappropriate when the observations are insufficient and/or grossly corrupted.

Besides, when we apply spectral clustering to segment the samples, the cluster number is a pre-specified parameter in traditional subspace clustering which cannot adaptive determine the number of effective clusters and find out the optimal clustering scheme, and efficiency of subspace clustering approaches, in which first clustering and then enumerating all the possible cases, were greatly affected by the number of clusters.

Is there a solution that can address the aforementioned research challenge?In this paper, an improved latent LRR model and a novel automatic subspace clustering method are developed, which aim to address the aforementioned challenges. Although the observations are insufficient and/or grossly corrupted, the improved latent LRR method can better capture the underlying local structure information in dataset and enhance the interpretation of model. After that, we employ the automatic subspace clustering approach to segment the samples.

2 Contributions

2.1 Improve Latent LRR Model

In this work, we apply a effective Euclidean distance and non-negative constraint jointly to the latent LRR model. Euclidean distance constraint help to learn low dimensional representation of invariance of rotation and offset, and adaptive select few nearest neighbor samples for representation. Non-negative constraint can eliminate the undesired solution, and then we can obtain a better similar matrix between data points. After that, the sparse and local information can be preserved in this way. By jointly introduce the two constraint, the novel improved latent LRR model can be expressed as $\min_{Z} \sum_{i,j}^{\mathrm{n}} \left\| x_i - x_j \right\|_2^2 z_{ij} + \lambda \left\| Z \right\|_*$, subject to $X_O = [X_O, X_H]Z$, $diag(Z) = 0$ and $Z \geq 0$. Here, the concatenation of X_O and X_H is used as the dictionary, X_O and X_H represent the observed and unobserved sample matrix respectively. Note that the diagonal elements of graph is set to zero, which can remove the influence of self-representation. In practice, a fraction of the samples are grossly corrupted, so a sparse error term should be incorporated. We can derive the joint objective function as $\min_{Z,E}\sum_{i,j}^{n}\|x_i-x_j\|_2^2 z_{ij}+\lambda_1\|Z\|_*+\lambda_2\|E\|_1$, subject to $X_O=[X_O,X_H]Z$, $diag(Z)=0$ and $Z\geq 0$, where λ_1 and λ_2 are the trade-off parameters. Then, the optimization can be solved through an efficiently iterative algorithm. After that, we can obtain the graph Z , where Z is derived from a non-negative affinity matrix and can be used to cluster.

2.2 Automatic Subspace Clustering

Modeling the cluster number in the optimization framework is intractable especially on the complex data space.In this work, we can model the cluster number automatically via the bias and variance trade-off. Automatic subspace clustering features an open cluster-number in the clustering phase, where the cluster number can be automatically added and discarded on the fly.Specifically, automatic subspace clustering is capable of initiating its cluster number from scratch without a fixed number. Its number automatically evolves in respect of the NS formula which produces an estimate of the data bias and variance[Andri *et al*., 2019]. Here, the NS formula is defined as the expectation of the squared predictive errors $E[(C-C)^2]$, where C represents the true class label and the C represents the predictive class label . After several mathematical derivation steps, we can obtain the bias formulas $(E[C]-C)^2$ and variance formulas $(E[C^2]-E[C]^2)$ respectively. Then we can monitor the quality of the subspace clustering automatically and not rely on the pre-specified cluster-number. In other words, the cluster number automatically augments $K=K+1$ if it appear a high bias or decreases $K=K-1$ if it signifies a high variance. It is noteworthy that the computational cost of this method is inexpensive especially on the complex data space, because the NS formulas is comparatively simple and can be calculated recursively.

Finally,we verify the performance of proposed latent LRR and automatic subspace clustering on both synthetic and real benchmark datasets by clustering accuracy(Acc) and normalized mutual information (NMI), and compares with the state-of-art baseline methods in the above aspects. We also verifies the convergence of the algorithm and analysis the computation complexity. However, the trade-off parameters about improve latent LRR need to be set respectively based on different datasets in order to obtained the desired performance.

3 Related Work

Subspace clustering base on LRR has been studied extensively in the fields of computer vision and machine learning. Liu propose latent LRR model base on the traditional LRR model, which aims to address the insufficiency of the observed data[Liu *et al*., 2011]. Recently, many researchers shown the locality preserving is critical to the expression of the model.Although NNLRS model achieve adaptive select few samples for data representation, the sparsity of this model doesn't guarantee[Zhuang *et al*., 2017]. Jie et al. proposed the LRR with adaptive graph regularization and can efficiently capture the local and sparse information of data simultaneously, but the performance of model is poor when the observations are insufficient and/or grossly corrupted[Jie *et al*., 2018]. In this paper, an improve latent LRR is presented to address the aforementioned issues, and then we can obtain a better graph for cluster. Traditional subspace clustering characterizes a fixed numbers of cluster, which isn't beneficial to make model selection automatically. We proposed an automatic subspace cluster approach via the bias and variance trade-off, which can identify the cluster number automatically.

4 Future Work

Future work should consider that the performance of proposed latent LRR depends on compute speed, memory demand and accuracy, so we envision that a more complex nonlinear model could be constructed in order to bridge that gap. Besides, the noise has much more complex statistical structures in practice, so we can construct a novel model under the Bayesian framework, which can fit a wide range of noises such as Gaussian, Laplacian, sparse noises and any combinations of them. The proposed latent LRR also should be applied to other tasks, such as large image classification, feature extraction and background modeling, etc.

References

[Lai *et al*., 2017] Taotao Lai, Hanzi Wang et al. Motion Segmentation Via a Sparsity Constraint. *IEEE Trans. Intelligent Transportation Systems*,18(4):973-983, 2017.

[Vidal, 2011] René Vidal. Subspace clustering.*IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

[Liu *et al*., 2011] Guangcan Liu, Shuichen Yan. Latent low-rank representation for subspace segmentation and feature extraction, In *IEEE International Conference on Computer Vision*, 2011, pp. 1615 – 1622.

[Zhuang *et al*., 2017] Zhuang, L., Gao, H., Lin, Z., *et al*., Non-negative low rank and sparse graph for semisupervised learning.In *IEEE Conference on CVPR*. IEEE, pp. 2328 – 2335, 2017.

[Jie *et al*., 2018] Jie Wen, Xiaozhao Fang et al. , Low-rank representation with adaptive graph regularization. *Neural Networks,* 108: 83-96, 2018.

[Liu *et al*., 2010] G. Liu, Z. Lin, and Y. Yu, Robust subspace segmentation by low-rank representation, in *International Conference of Machine Learning*, 2010.

[Andri *et al*., 2019] Andri Ashfahani, Mahardhika Pratama, Edwin Lughofer, Yew-Soon Ong: DEVDAN: Deep Evolving Denoising Autoencoder. *CoRR*, 2019.