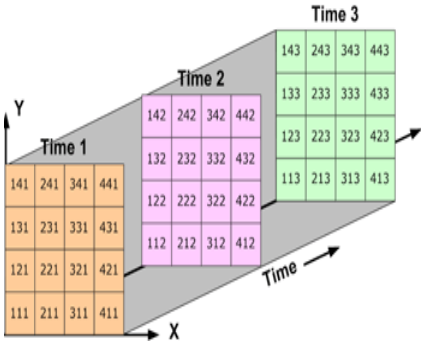
Tensor and Unsupervised Learning

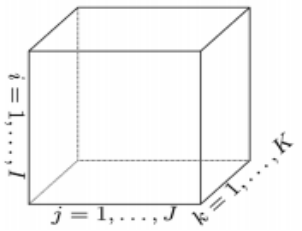
Supervisors : Joseph Bengeloun (LIPN-CALIN),
Mustapha Lebbah (HdR, LIPN-A3),

1 Topic

Clustering analysis has become a fundamental tool in statistics and machine learning. Many clustering algorithms have been developed, with the general idea of seeking groups among different individuals in all space of features. Biclustering consists of simultaneous partitioning of a set of observations and a set of their features into subsets often called bicluster. Consequently, a subset of rows exhibiting significant coherence within a subset of columns in the matrix can be extracted, which corresponds to a specific coherent pattern [9]. Nowadays, there is a new type of data collection, in which we may collect data by *individual-feature* pair at multiple times Fig.1(a). The variation of a couple *individual-feature* at different times is called trajectory. This data can be represented as a tensor of three dimensional $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times m}$ where n_1, n_2 and m represent the size of observations, features and times respectively Fig.1(b). Many tools on tensor manipulation already exist in literature and help us to solve this tensor biclustering problem [1, 5–7].



(a) Data matrix collected at different times



(b) A third-order tensor with $I = n_1$, $J = n_2$ and $K = m$

Figure 1: Data description, (a) and (b) represent the collection of data and the all data set respectively.

Complete the missing features and grouping the trajectory of the tensor according to the correlation or simularity between them is still a very challenging topics. Many researcher in data sciences working to solve this problem in large data set for example, in [8] the authors proposed different methods based on two methods: first they are the spectral decomposition of matrix and second the length of trajectory which is applied on countinuous data set and provide only one bicluster.

The tensor folding and spectral method in [8] is extended and improved by Dina Faneva et al. [2] and they provide two methods to select many bicluster in tensor datasets. On another setting, Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye [4], proposed one method of filling the missing values in tensor of visual datasets.

2 Problems

(1) Low dimensional clustering is not to partition all data into different clusters, but to look for clusters in projections to lower dimensions [3, 10]. In tensor biclustering methods [8], if there is one bicluster, we need to define the parameter k_1 which is the cardinality of the subset of individuals and k_2 which is the cardinality of the subset of features. The same problem remains if we increase the size of low dimensional clustering and partitioning by many biclusters.

(2) The core problem of the missing value estimation lies on how to build up the relationship between the known elements and the unknown ones. In [4], the authors proposed a method of tensor completion using a convex optimization algorithm, called Low Rank Tensor Completion (LRTC) and block coordinate descent (BCD). It solves only the problem in small the datasets. We want to generalize it to a tensor completion with large real datasets.

(3) Tensors are well-known objects in representation theory of the general linear group [11]. Their decomposition in irreducible representations turns out to be useful in many situations to simplify calculations. We may ask the question whether such a decomposition in irreducible representations may be imported and useful to the search of bi or higher dimensional clusters.

(4) Concerning the process of collection of times series data, for each period of times, the fix couple *(individuals, features)* has a new data entry. So, we aim at finding a probabilistic model predicting the next value of the trajectory [8] using Markov chain.

3 Work plan

Our goal is the improvement of biclustering algorithms, elaborating new biclustering method and to provide idea for filling missing value in high dimensional datasets.

We have the following plan:

(1) First year: understand all the concepts related to our problem and provide a new approach to fill the missing value in high dimensional datasets.

(2) Second year: provide improvements and new biclustering methods by starting with three dimensional datasets. To generalize low subspace clustering of $n-$dimensional datasets.

(3) Third year: build a probabilistic model predicting the next value of the scalability of each algorithm in large size dataset. Writing the thesis manuscript.

References

[1] Andrea Montanari, Daniel Reichman and Ofer Zeitouni. On the limitation of spectral methods: From the gaussian hidden clique problem to rank-one perturbations of gaussian tensors. *In Advances in Neural Information Processing Systems*, 2015.

[2] Andriantsiory Dina Faneva, Mustapha Lebbah, Hanane Azzag and Beck Gael. Algorithms for an efficient tensor biclustering. *arXiv:1903.04042v1*, 2019.

[3] Frank Klawonn, Frank Hoppner, Balasubramanian Jayaram. What are clusters in high dimensions and are they difficult to find? *Revised Selected Papers of the First International Workshop on Clustering High–Dimensional Data - Volume 7627*, pages 14–33, 2015.

[4] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. 2018.

[5] Emile Richard and Andrea Montanari. A statistical model for tensor pca. *In Advances in Neural Information Processing Systems*, 2014.

[6] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. *In COLT*, 2015.

[7] Samuel B Hopkins, Tselil Schramm, Jonathan Shi and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. *arXiv preprint arXiv*, 2015.

[8] Soheil Feizi, Hamid Javadi, David Tse. Tensor biclustering. *Advances in Neural Information Processing Systems*, 30:1311–1320, 2017.

[9] Amos Tanay, Roded Sharan, and Ron Shamir. Biclustering algorithms: A survey. In *In Handbook of Computational Molecular Biology Edited by: Aluru S. Chapman Hall/CRC Computer and Information Science Series*, 2005.

[10] Widia Sembiring, Jasni Mohamad Zain and Abdullah Embong. Clustering high dimensional data using subspace and projected clustering algorithms. *arXiv preprint arXiv:1009.0384*, 2010.

[11] Fulton Williams and Harris Joe. *Representation theory. A first course. Graduate Texts in Mathematics, Readings in Mathematics*. Springer-Verlag, 1991. 129.