



Jayesh K. Gupta, Sumanik Singh, Nishchal K. Verma

Department of Electrical Engineering, Indian Institute of Technology, Kanpur

Introduction

Biclustering is a popular approach to analyze patterns in a dataset, especially those of biological origin such as gene expression data. Biclustering performs better than classical clustering techniques under certain data sets, since it can simultaneously cluster both rows and columns of the matrix unlike the latter. As a result, submatrices exhibiting unique patterns can be revealed helping us to better understand the relationship between row and column variables. We introduce a new Matlab toolbox designed to perform a variety of biclustering algorithms under a common user interface. Although some implementations are available for the proposed biclustering algorithms, each program is accompanied by a different user interface and use different input-output formats. MTBA tries to fill this gap by providing multiple functionalities for data handling, preprocessing, biclustering and visualization.

Algorithms

- Cheng and Church [1]
- Bipartite Spectral Graph Partitioning [2]
- OPSM [3]
- ISA [4]
- Spectral Biclustering [5]
- Information Theoretic Learning [6]
- xMOTIF [7]
- Plaid [8]
- FLOC [9]
- Bimax [10]
- Bayesian Biclustering [11]
- LAS [12]

Implementation

A single wrapper function `biclust`, can be used to call any algorithm along with their respective parameters to act on the data matrix. E.g.

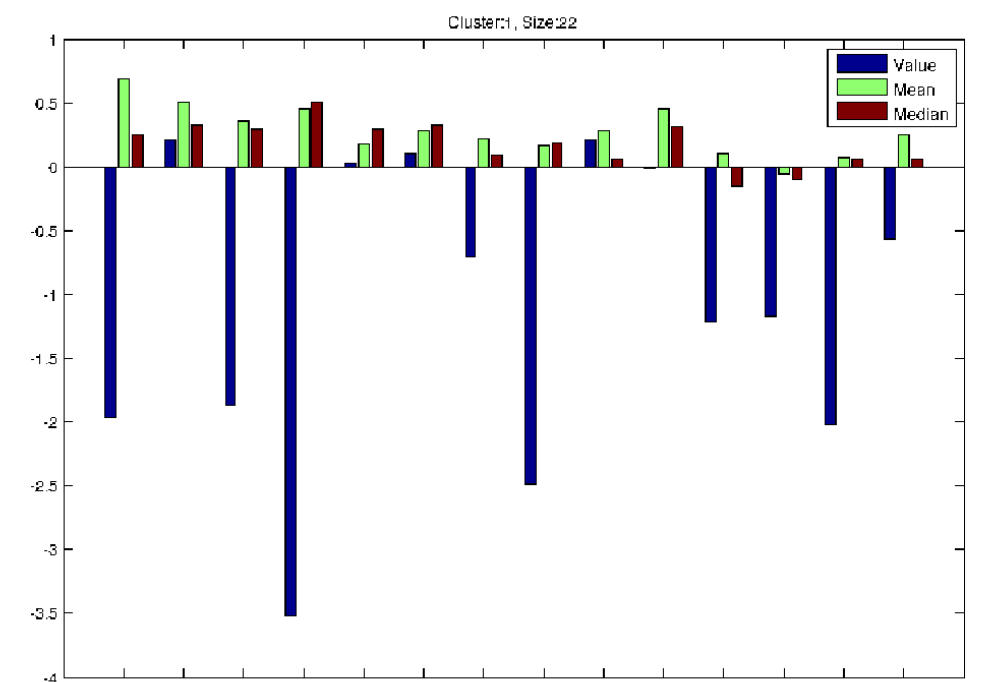
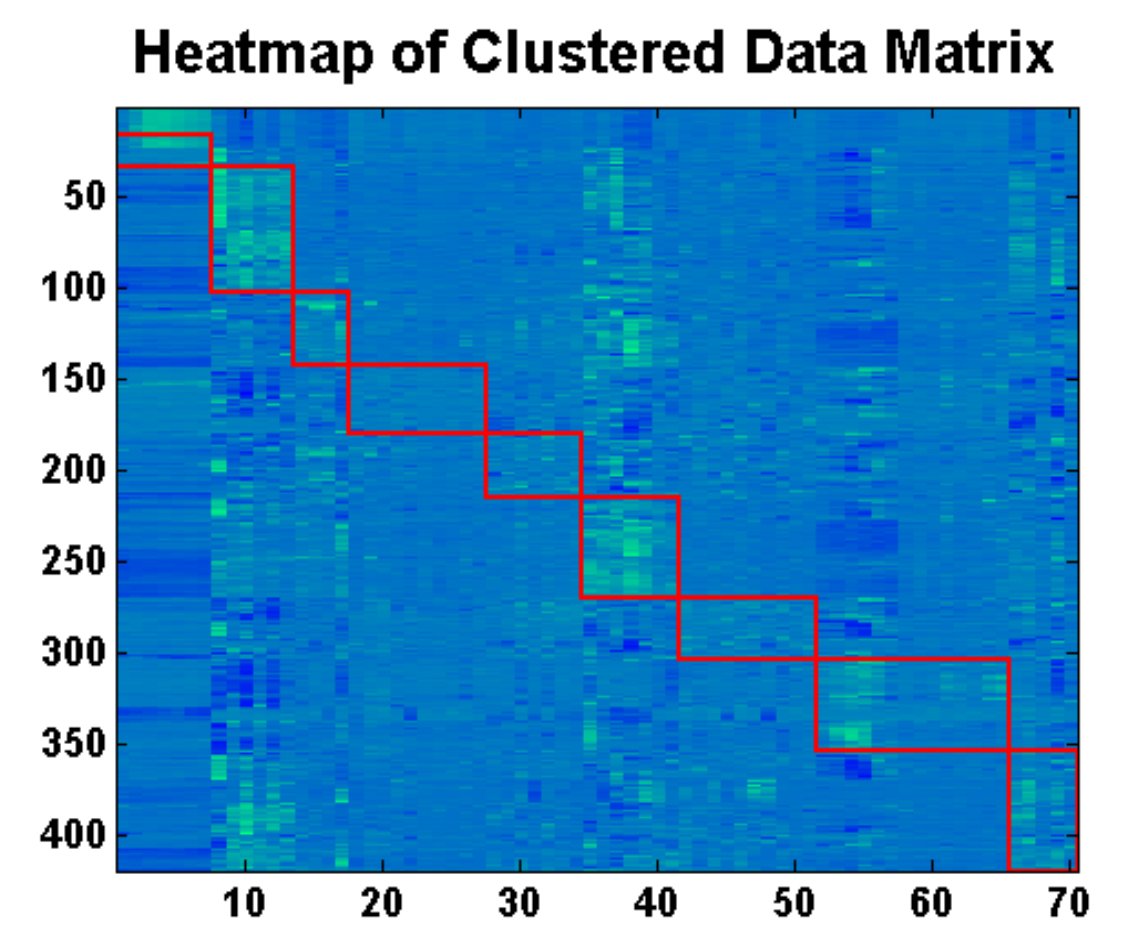
```
res = biclust(data, 'plaid');
```

This results in a consistent output of structure `'biclustResult'` which contains the information about the output biclusters as the row and column indices, along with their logical matrix counterpart. Also any extra information from the algorithms is stored in this structure.

References: [1] Y. Cheng, G.M. Church, "Biclustering of Expression Data", *Proc Int Conf Intell Syst Mol Biol*, 2000. 8: p. 93-103. [2] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. *IProceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '01)* 2001. ACM, New York, NY, USA, 269-274. [3] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. "Discovering local structure in gene expression data: The order-preserving submatrix problem", *Proc. International Conference on Computational Biology*, pages 49-57, 2002. [4] Bergmann S, Ihmels J, Barkai N: Iterative signature algorithm for the analysis of large-scale gene expression data *Phys Rev E Stat Nonlin Soft Matter Phys*. 2003 Mar;67(3 Pt 1):031902. Epub 2003, Mar 11. [5] Kluger et al., "Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions", *Genome Research*, 2003, vol. 13, pages 703-716 . [6] I.S. Dhillon, S. Mallela, D.S. Modha, (2003, August). "Information-theoretic co-clustering", *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 89-98). ACM. [7] Murali, T. & Kasif, S. Extracting Conserved Gene Expression Motifs from Gene Expression Data. *Pacific Symposium on Biocomputing*, sullivan.bu.edu, 2003, 8, 77-88 [8] Heather Turner et al, "Improved biclustering of microarray data demonstrated through systematic performance tests", *Computational Statistics and Data Analysis*, 2003, vol. 48, pages 235-254. [9] Yang, Jiong, et al. "An improved biclustering method for analyzing gene expression profiles." *International Journal on Artificial Intelligence Tools* 14.05 (2005): 771-789. [10] Prelic, A.; Bleuler, S.; Zimmermann, P.; Wil, A.; Buhlmann, P.; Gruissem, W.; Hennig, L.; Thiele, L. & Zitzler, E. A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data. *Bioinformatics*, Oxford Univ Press, 2006, 22, 1122-1129 [11] Gu, Jiajun, and Jun S. Liu. "Bayesian biclustering of gene expression data." *BMC genomics* 9.Suppl 1 (2008): S4. [12] Shabalin, Andrey A., et al. "Finding large average submatrices in high dimensional data." *The Annals of Applied Statistics* (2009): 985-1012. [13] Chia, Burton Kuan Hui, and R. Krishna Murthy Karuturi. "Research Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms." (2010).



1. Easy to use API
2. Documentation embedded in Matlab help system
3. Preprocessing features like data normalization, discretization and binarization.
4. Heatmap and profile visualization.
5. Postprocessing and analysis of overlapping biclusters, Chia and Karuturi scores [13]
6. Free to use and remix (MIT Licence)



Acknowledgements

We would like to thank all members of Intelligent Informatics Group at IIT-K for valuable discussions and suggestions. We are also grateful to Andrey Shabalin for graciously releasing the LAS code under LGPL and allowing us to port it to our toolbox.

Get MTBA
online:



<http://home.iitk.ac.in/~jayeshkg/mtba/>