A Formal Explanation Space for the Simultaneous Clustering of Neurology Phenotypes

Raghu Yelugam[1*], Daniel B Hier, Tayo Obafemi- Ajayi, Michael D Carrithers, Donald C Wunsch II

[1*] Applied Computational Intelligence Laboratory, Missouri University of Science and Technology, Rolla, 65409, Missouri, USA,.
[2] Engineering Program, Missouri State University, Springfield, 65897, Missouri, USA.
[3] Neurology and Rehabilitation, University of Illinois at Chicago, Chicago, 60612, Illinois, USA.

*Corresponding author. E-mail: ry222@umsystem.edu
Contributing authors: hierd@umsystem.edu; tayoobafemiajayi@missouristate.edu; mcar1@uic.edu; dwunsch@umsystem.edu

Supplementary Materials

Simultaneous Clustering

Starting with Hartigan's *direct clustering* [1], simultaneous data matrix clustering of rows and columns has been described using various terms such as *two-way two-mode clustering*, *biclustering*, *subspace clustering*, and *coclustering* [2–5]. Though sequential clustering of rows and columns can yield useful outputs[1], the resulting submatrices may not be homogeneous. However, simultaneous clustering leads to homogeneous submatrices within a data matrix which could appear in various forms, including partitions of entire rows and columns or overlapping submatrices with partial rows and columns. We will use the term *bicluster* to refer to these homogeneous submatrices.

In this article, a data matrix $\mathbf{D}$ is represented as $\mathbf{D} = (R,\ C)$, where $R = \{r_1, \ldots, r_n\}$ is a set of n rows (or disease cases) and $C = \{c_1, \ldots, c_m\}$ is the set of m columns (or phenotypes). The element $d_{ij} \in \mathbf{D}$ denotes the magnitude of row i observed at column j. Various simultaneous clustering techniques have been tailored for biomedical use [5, 8, 9]. This article delves into spectral methods, notably *spectral coclustering* and *spectral biclustering*, which harness the eigenvectors of $\mathbf{DD^T}$ via the singular value decomposition (SVD) of $\mathbf{D}$ [10].

Spectral biclustering (SBC), conceptualized for genome expression data, is rooted in the assumption of a checkerboard structure in the data. Given a matrix $\mathbf{D}$ showcasing this pattern, when we apply a steplike classification vector x, the result is another steplike vector y. Reapplying $\mathbf{D^T}$ to y yields a vector akin to x. Thus, SBC finds eigenvectors (x and y) which satisfy the coupled eigenvalue equations

$$\mathbf{D^T D}x = \lambda^2 x,$$
$$\mathbf{DD^T}y = \lambda^2 y,$$

where both x and y share the same eigenvalue. This translates to obtaining the SVD of $\mathbf{D}$:

$$\mathbf{D} = U\Lambda V^T,$$

with Λ being a diagonal matrix with descending non-negative entries; $U_{n\times\min(n,m)}$ and $V_{m\times\min(n,m)}$ are, respectively, the right and left singular vectors associated with $\mathbf{DD^T}$ and $\mathbf{D^T D}$. Pre-eigenproblem-solving, SBC normalizes data to negate biases from differently scaled features. Bistochastisation and log-interaction normalization, introduced by [11], serve to prep data before SVD application. Post-SVD, if normalization is via bistochastisation or scaling, the first principal eigenvector from U and V is discarded for its uniform contribution to $\mathbf{D}$. Information for partitioning is extracted by sorting eigenvector entries, setting thresholds to split them, and then choosing eigenvectors that can be approximated by piecewise-constant vectors. The final clustering step involves using the k-means algorithm and normalized cut method on data projected to the top eigenvectors.

Spectral coclustering (SCC), a bipartite graph partitioning technique, was proposed to co-cluster the word-by-document matrix $\mathbf{D}$, where rows correspond to words and columns to documents. Given that the matrix entry D_{ij} symbolizes the occurrence of word i in document j, $\mathbf{D}$ is inherently sparse. For a given $\mathbf{D}$, a graph $G = (V,\ E)$ is defined, where vertices $V = \{1, 2, \ldots, n\}$ represent documents and words, and edges

[1]For instance, see the heat map widget in Orange and the cluster map application in Seaborn [6, 7]

$E = \{i, j\}$ represent word i appearing in document j (or vice versa). An adjacency matrix M is constructed as

$$M = \begin{bmatrix} 0 & \mathbf{D} \\ \mathbf{D^T} & 0 \end{bmatrix},$$

with $M_{ij} = w_{ij}$ (the edge weight) for every $\{i, j\} \in E$. A *cut*, defined as $cut(V_1, V_2) = \sum_{i \in V_1, j \in V_2} M_{ij}$, provides insight into how clusters are interconnected. To ensure cluster balance, a normalized-cut objective function is employed, defined as

$$N(V_1, V_2) = \frac{cut(V_1, V_2)}{weight(V_1)} + \frac{cut(V_1, V_2)}{weight(V_2)},$$

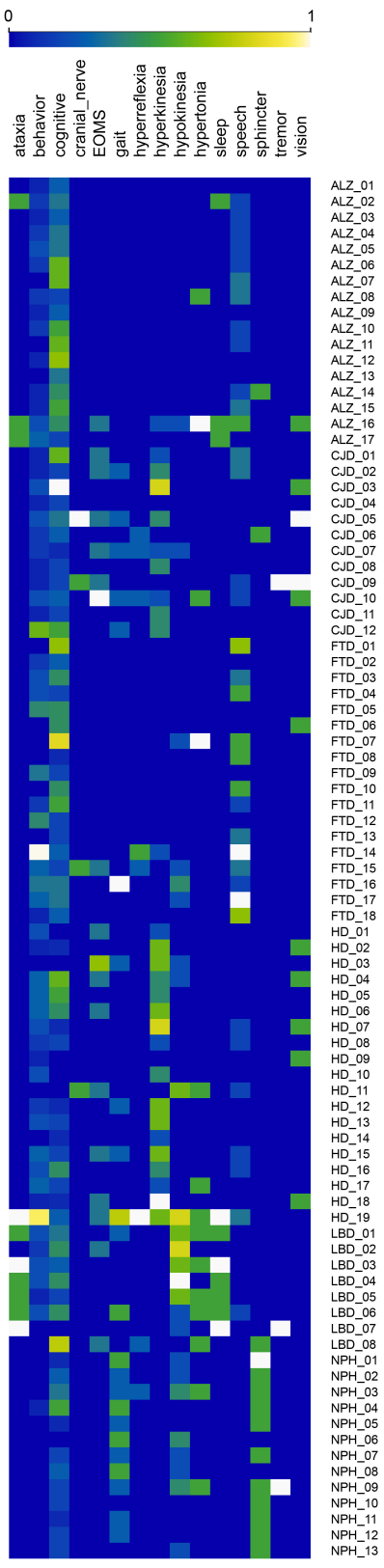
where the weight is $weight(V) = \sum_{i \in V} \sum_k E_{ik}$. This objective function is tackled using SVD to pinpoint the $l = \log_2(k)$ singular vectors $U = [u_2, \ldots, u_{l+1}]$ and $W = [w_2, \ldots, w_{l+1}]$ of the normalized matrix $\mathbf{D_n} = D_1^{-\frac{1}{2}} \mathbf{D} D_2^{-\frac{1}{2}}$. Here, D_1 and D_2 are diagonal degree matrices. The derived singular vectors construct a dataset of dimension l,

$$Z = \begin{bmatrix} D_1^{-\frac{1}{2}} U \\ D_2^{-\frac{1}{2}} W \end{bmatrix},$$

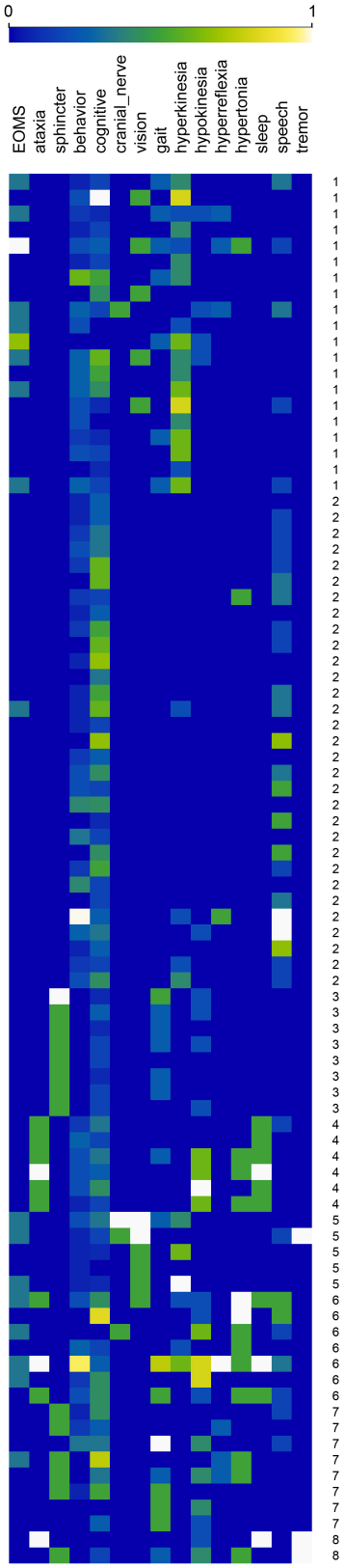
which is then clustered using the k-means method to yield k clusters.

Both SBC and SCC are spectral-based but make different assumptions regarding the underlying structure in $\mathbf{D}$. SBC assumes a checkerboard pattern, whereas SCC views $\mathbf{D}$ as a sparse matrix aiming to discover a diagonal block structure. Moreover, SCC results in the same number of row and column clusters, leading to an equal number of submatrices, whereas SBC can yield a different number of row and column clusters, producing a distinct set of submatrices.

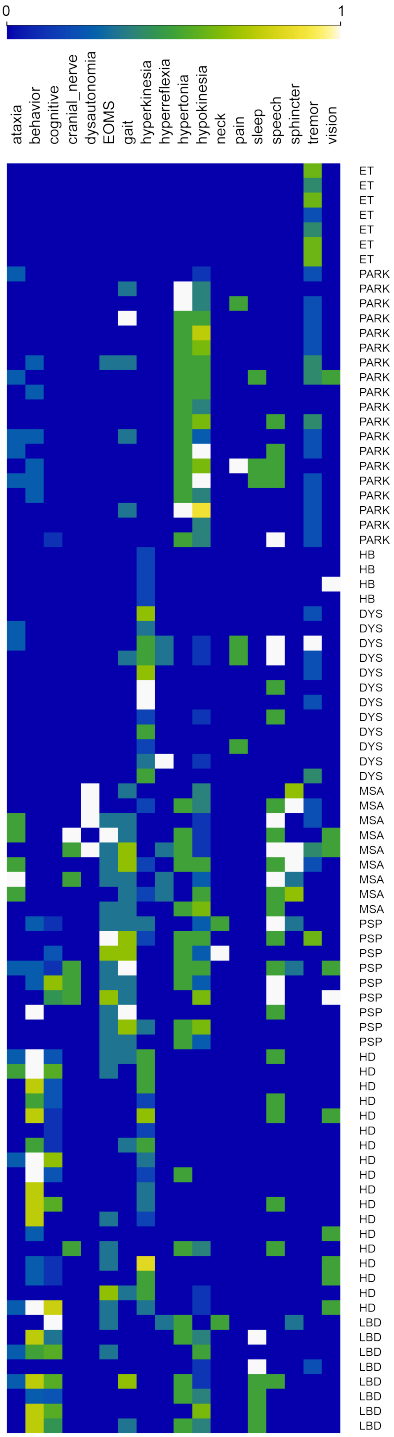
Data Visualizations Before and After Simultaneous Clustering



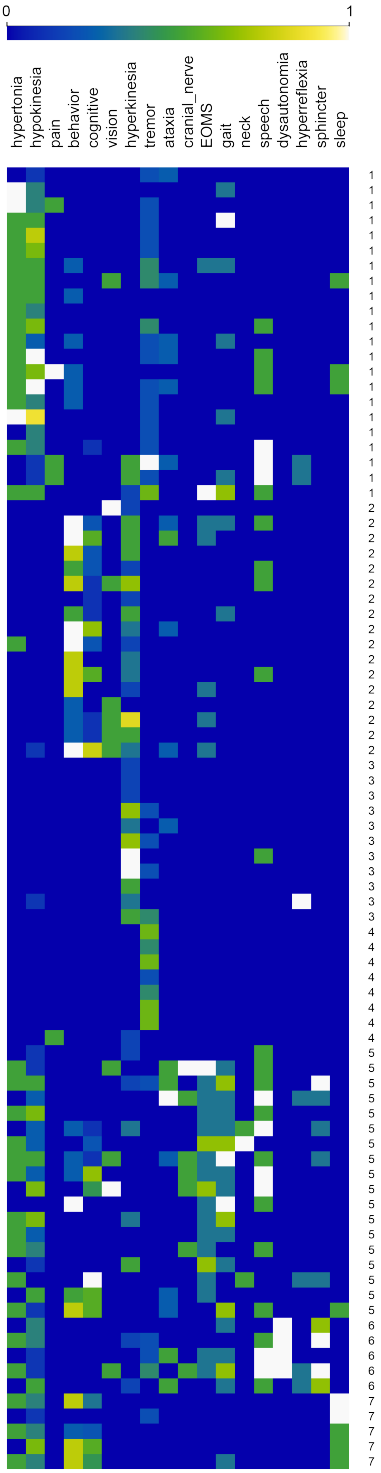
Supplementary Fig. 1: Data structure for the dementia dataset before coclustering. There are 15 columns (features) and 87 rows (cases). The heat map shows the mean feature score for each row normalized to the interval [0,1].



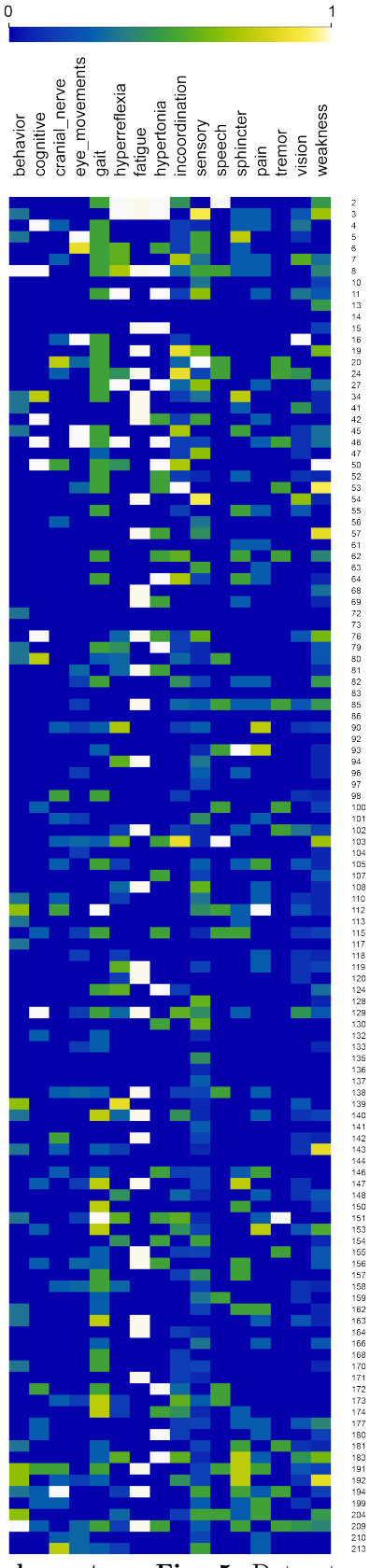
Supplementary Fig. 2: Data structure for dementia dataset after coclustering. Columns and rows have been re-arranged according to cluster membership. Coclustering brings out a diagonal pattern in the heat map.



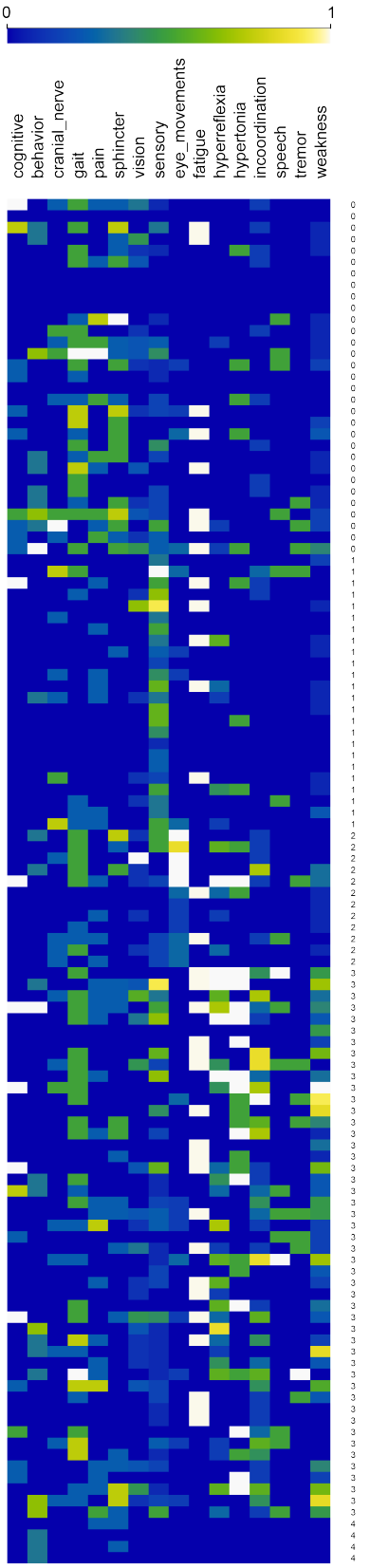
Supplementary Fig. 3: Data structure for the movement disorders dataset before coclustering. There are 18 columns (features) and 86 rows (cases). The heat map shows the mean feature score for each row normalized to the interval [0,1].



Supplementary Fig. 4: Data structure for movement disorders dataset after coclustering. Columns and rows have been rearranged according to cluster membership. Coclustering produces a diagonal pattern in the heat map.



Supplementary Fig. 5: Data structure for the MS dataset before coclustering. There are 16 columns (features) and 119 rows (cases). The heat map shows the mean feature score for each row normalized to the interval [0,1].



Supplementary Fig. 6: Data structure for MS dataset after coclustering. Columns and rows have been rearranged according to cluster membership. Coclustering produces a diagonal pattern in the heat map.

References

[1] Hartigan, J. A. Direct clustering of a data matrix. *Journal of the American Statistical association* **67**, 123–129 (1972).

[2] Van Mechelen, I., Bock, H.-H. & De Boeck, P. Two-mode clustering methods: a structured overview. *Statistical methods in medical research* **13**, 363–394 (2004).

[3] Madeira, S. C. & Oliveira, A. L. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1**, 24–45 (2004).

[4] Vidal, R. Sub clustering. *IEEE Signal Processing Magazine* **28**, 52–68 (2011). URL 10.1109/MSP.2010.939739.

[5] Govaert, G. & Nadif, M. *Co-clustering: models, algorithms and applications* (John Wiley & Sons, 2013).

[6] Waskom, M. L. Seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021). URL https://doi.org/10.21105/joss.03021.

[7] Demšar, J. *et al.* Orange: Data mining toolbox in python. *Journal of Machine Learning Research* **14**, 2349–2353 (2013). URL http://jmlr.org/papers/v14/demsar13a.html.

[8] Padilha, V. A. & Campello, R. J. A systematic comparative evaluation of biclustering techniques. *BMC bioinformatics* **18**, 55 (2017).

[9] Parsons, L., Haque, E. & Liu, H. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter* **6**, 90–105 (2004).

[10] Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing* **17**, 395–416 (2007).

[11] Kluger, Y., Basri, R., Chang, J. T. & Gerstein, M. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research* **13**, 703–716 (2003).