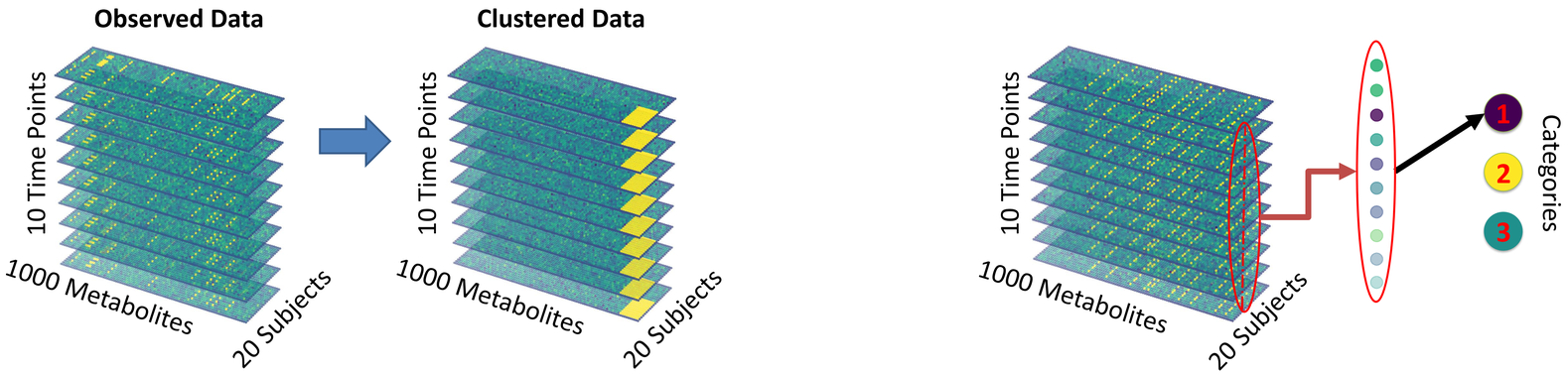
Biclustering Matrices with Categorical Entries

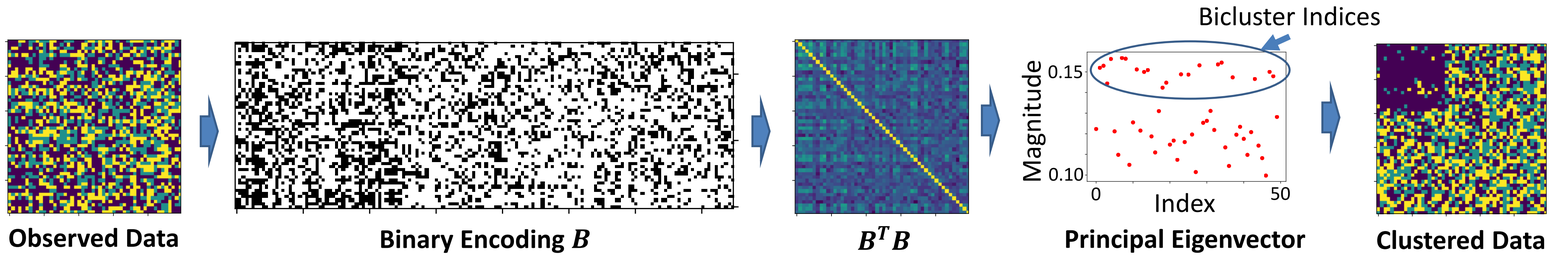
Anthony Degleris, Soheil Feizi, and David Tse
Stanford University, Department of Electrical Engineering

Motivating Problem: Clustering Tensor Data



- Modern data is often represented as tensors (many dimensions).
- Can we find clusters with a similar trajectory in time?
- **Idea:** Categorize each time vector.
- Categories reduce tensor to matrix.

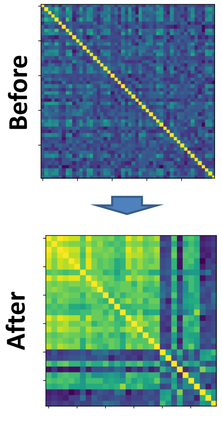
Our Contribution: Biclustering Categorical Data



Spectral Analysis of Categorical Data Matrix

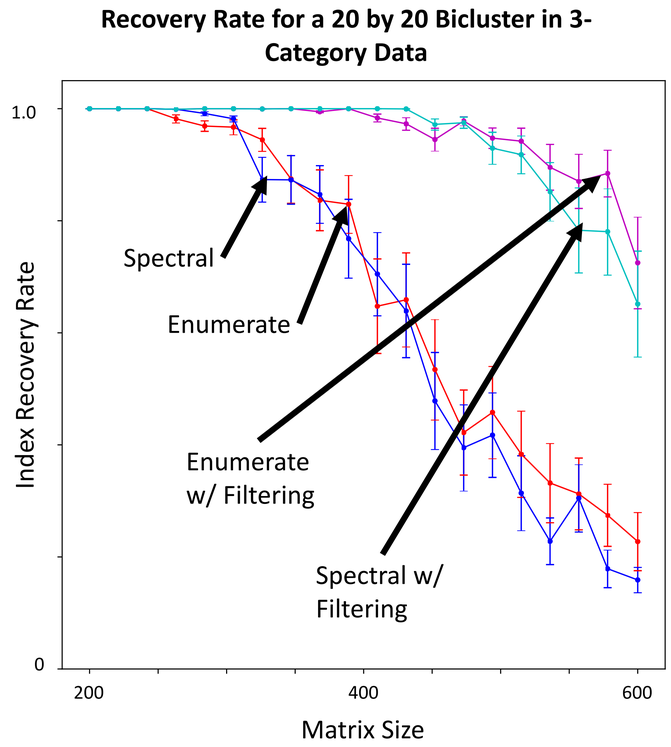
- *Biclustering* permutes rows and columns to locate similar entries.
- Well studied for continuous data (hidden submatrix) and binary data (hidden clique).
- What about categorical data?

- **Method:** Encode data in a binary matrix.
- Use spectral (eigenvector) method to estimate bicluster.
- Theoretically guarantee success for $k \propto \sqrt{n}$ (bicluster size k and matrix size n).
- *(Optional)* Iterate several times to filter noisy rows/columns and improve estimate.



Filtering Step

Numerical Results



- Enumeration counts frequency of each category to determine bicluster.
- Enumeration and spectral analysis perform similarly.
- Filtering step notably improves performance.

References

1. Cai, T. T., Liang, T., & Rakhlin, A. (2017). Computational and statistical boundaries for submatrix localization in a large noisy matrix. The Annals of Statistics, 45(4), 1403-1430.
2. Cheng, Y., & Church, G. M. (2000, August). Biclustering of expression data. In *Ismb* (Vol. 8, No. 2000, pp. 93-103).

Acknowledgements

This project was funded by the Research Experience for Undergraduates program in the Department of Electrical Engineering at Stanford University.