Journal of Data Science xx (x), 537–537
July 2020
DOI: xx.xxxx/xxxxxxxxx

Clustering US States by Time Series of COVID-19 New Case Counts in the Early Months with Non-negative Matrix Factorization

Jianmin Chen[1] and Panpan Zhang*[2]

[1]*Department of Statistics, Universityof Connecticut, Storrs, CT 06269, U.S.A.*
[2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.*

Summary

The supplement contains data and codes for reproducing the results presented in the article.

Data files

1. data_10_05.csv: This file contains the data from a public repository maintained by the Center for Systems Science and Engineering at the Johns Hopkins University (Dong et al., 2020). The data was retrieved on October 5, 2020. The case numbers may differ from those in the current version owing to possible modifications made after October 5, 2020.
2. nst-est2019-01.csv: This file contains the state-level population data, maintained by the US Census Bureau (https://www.census.gov). The data was released at the end of 2019.

Code files

1. pretreat.R: Codes for pre-processing the data (e.g., smoothing and scaling).
2. getnmfparameter.R: Codes for obtaining NMF ranks via the cross-validation method proposed in the paper.
3. model_fit.R: Codes for implementing the NMF method. The results of k-means clustering (including the selection of k) are given by running this file as well.
4. plotmaking.R: Codes for generating the figures in the paper.

References

Dong E, Du H, Gardner L (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5): 533–534.

*Corresponding author. Email: panpan.zhang@pennmedicine.upenn.edu

Received May, 2020; Accepted July, 2020