A network-based algorithm for clustering multivariate repeated measures data

Matthew Koslovsky[1,*], John Arellano[2], Caroline Schaefer[2,*], Alan Feiveson[3], Millennia Young[3], and Stuart Lee[1]

[1]KBRwyle, Houston, TX; [2]MEI Technologies, Houston, TX;
[3]NASA Johnson Space Center, Houston, TX

Abstract

The National Aeronautics and Space Administration (NASA) Astronaut Corps is a unique occupational cohort for which vast amounts of measures data have been collected repeatedly in research or operational studies pre-, in-, and post-flight, as well as during multiple clinical care visits. In exploratory analyses aimed at generating hypotheses regarding physiological changes associated with spaceflight exposure, such as impaired vision, it is of interest to identify anomalies and trends across these expansive datasets. Multivariate clustering algorithms for repeated measures data may help parse the data to identify homogeneous groups of astronauts that have higher risks for a particular physiological change. However, available clustering methods may not be able to accommodate the complex data structures found in NASA data, since the methods often rely on strict model assumptions, require equally-spaced and balanced assessment times, cannot accommodate missing data or differing time scales across variables, and cannot process continuous and discrete data simultaneously. To fill this gap, we propose a network-based, multivariate clustering algorithm for repeated measures data that can be tailored to fit various research settings. Using simulated data, we demonstrate how our method can be used to identify patterns in complex data structures found in practice.

Objectives

- Design a multivariate, repeated measures clustering algorithm, CommClust, that is robust to the complex data structures found at NASA and flexible to other research settings
- Demonstrate its performance in simulation and apply it to data collected during a bed rest study
- Identify groups of individuals who behave similarly over time

Data Challenges

- Repeated measures
- Inconsistent time scales
- Missing data
- Unequally spaced, unbalanced assessment times
- Continuous and discrete measures

Contact Information

* matthew.d.koslovsky@nasa.gov
* caroline.m.schaefer@nasa.gov

A network-based algorithm for clustering multivariate repeated measures data

Matthew Koslovsky[1,*], John Arellano[2], Caroline Schaefer[2,*], Alan Feiveson[3], Millennia Young[3], and Stuart Lee[1]

[1]KBRwyle, Houston, TX; [2]MEI Technologies, Houston, TX;
[3]NASA Johnson Space Center, Houston, TX

Background

- There is a vast amount of repeated measures data collected on the National Aeronautics and Space Administration (NASA) Astronaut Corps
- Astronauts voluntarily participate in biomedical research studies before, during, and after flight
- Occupational surveillance data are collected yearly on astronauts, even after their retirement
- There is a focus to identify clusters of astronaut profiles that are at a greater risk for short and long term spaceflight associated health outcomes
- Clustering methods may not accommodate complex NASA data structures currently; models often rely on strict assumptions, equally-spaced and balanced assessment times, and cannot process continuous and discrete data simultaneously
- Parametric methods for clustering multivariate repeated measures data require specification of potentially complex relations between independent and dependent variables. Non-parametric clustering methods can require observations to be collected at the same time for each individual
- It may be advantageous to devise a multivariate clustering algorithm for repeated measures data that combines the strengths of previously developed machine learning and statistical methods
- This approach can require very few or no assumptions about the data and is extremely flexible
- We developed an algorithm, CommClust, to identify clusters using univariate clustering algorithms, networks analysis, and community detection methods

CommClust Methods

- Step 1: Perform univariate clustering for each repeated measure variable
  - For each subject $i = 1, \dots, N$, we observe $j = 1, 2, \dots, n_i$ assessments for each variable Y_p, $p = 1, 2, \dots, P$.
  - Use any univariate repeated measures clustering method to identify c_p clusters of individual trajectories
- Step 2: Construct the network
  - Build a network that connects individuals (nodes) that are clustered into the same group for each variable. Here, $E(i,k)$ is an edge between node i and k.

```
For p in 1 to P;
    For i in 1 to N;
        For k in i+1 to N;
            If c_p(i) = c_p(k), then E(i,k).
```

- Step 3: Identify communities in the network
  - Employ a community detection method to identify groups of individuals with similar trajectories across the measured variables

Simulation Study

- We compared CommClust with various specifications to another non-parametric approach, K-Means for Joint Longitudinal Data (KML3D) [2]
- Methods were compared using Jaccard index [3], convergence times, and average number of groups selected

Table 1: Specifications of CommClust for simulation study

Method	Univariate Clustering Algorithm	# of Clusters	Community Detection	Label
1	KML Univariate [2]	Calinski Harabatz [1]	Leading eigenvector [7]	KML-lec
2			Walktrap [8]	KML-walk
3	longclustEM [6]	BIC [11]	Leading eigenvector	Longclust-lec
4			Walktrap	Longclust-walk
5	traj [5]	Cubic clustering criterion [10]	Leading eigenvector	Traj-lec
6			Walktrap	Traj-walk
7	Non-parametric model and PAM [4]	Silhouette Width [9]	Leading eigenvector	Non-lec
8			Walktrap	Non-walk
9	Scagnostics and PAM [13]	Silhouette Width	Leading eigenvector	Scag-lec
10			Walktrap	Scag-walk

Simulations

- Bivariate, repeated measures data, $\boldsymbol{Y}_{ij} = (Y_{1ij}{}^{(g)}, Y_{2ij}{}^{(g)})$, were simulated in 4 different scenarios, similar to [2]
- Scenarios included random individual $i = 1, 2, \dots, N_g$ and measurement error, had various numbers of true clusters ($g = 1, \dots, G$), and varied in the patterns of trajectories (e.g., linear, curvilinear)
- In each scenario, we assessed $N_g = 10$ and 50 individuals in each cluster $n_i = n = 11$ equally spaced assessment times $t_{ij} (j = 1, 2, \dots, n_i)$ in $[0, 10]$
- Scenario 3 was additionally analyzed with randomly spaced assessment times (3^a)
- Algorithm performance for Scenario 4 was evaluated with 20% and 50% missing data (4^a and 4^b)

Example Scenario 1:

$$\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix}^{(1)} = \begin{matrix} 0 + b_i + \varepsilon_{ij} \\ 0 + b_i + \varepsilon_{ij} \end{matrix}$$

$$\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix}^{(2)} = \begin{matrix} 0 + b_i + \varepsilon_{ij} \\ t_{ij} + b_i + \varepsilon_{ij} \end{matrix}$$

$$\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix}^{(3)} = \begin{matrix} t_{ij} + b_i + \varepsilon_{ij} \\ 0 + b_i + \varepsilon_{ij} \end{matrix}$$

where $b_i \sim N(0, \sigma^2)$, σ ranges from 1 to 8 by 0.05, and $\varepsilon_{ij} \sim N(0, 1)$.

Simulation Results

Table 2: Best performing* model for each scenario

Scenario	N_g	Model	Jaccard Median (Min-Max)	Time, s Mean (sd)	# of Groups Mean (sd)
1	10	Scag-walk	0.73 (0.23-1.00)	0.27 (0.01)	3.49 (1.02)
	50	Longclust-walk	1.00 (0.50-1.00)	25.92 (4.01)	3.08 (0.32)
2	10	Non-lec	0.27 (0.14-1.00)	0.64 (0.11)	3.26 (0.58)
	50	KML3D	0.34 (0.14-1.00)	3.27 (0.22)	3.32 (0.98)
3	10	Non-lec	0.38 (0.13-0.62)	0.89 (0.11)	3.89 (0.92)
	50	Longclust-walk	0.44 (0.19-0.62)	51.08 (7.49)	3.31 (0.87)
3^a	10	Non-walk	0.39 (0.12-0.53)	1.09 (0.17)	3.73 (1.09)
	50	Non-lec	0.39 (0.15-0.55)	14.21 (4.47)	3.96 (0.91)
4	10	Non-lec	0.28 (0.11-0.69)	0.92 (0.11)	3.53 (0.76)
	50	Longlust-walk	0.45 (0.27-0.64)	36.64 (6.26)	3.69 (1.54)
4^a	10	Non-lec	0.26 (0.11-0.69)	0.88 (0.11)	3.30 (0.72)
	50	Non-walk	0.26 (0.13-0.60)	7.11 (0.92)	3.45 (1.07)
4^b	10	Non-lec	0.22 (0.11-0.53)	0.84 (0.09)	3.23 (0.70)
	50	KML3D	0.23 (0.13-0.71)	9.72 (0.58)	2.56 (0.78)

*Perfomance determined by highest Jaccard index [0,1] and closest to correct number of groups (3 or 5)

Discussion

- We developed a flexible clustering algorithm for multivariate repeated measures data
- This method can handle various analytical challenges, including missing data, unequally spaced and unbalanced assessment times, and different time scales
- The CommClust approach can be used for both repeated measures and cross-sectional clustering
- Clustering can be run in parallel to reduce computational times in high-dimensional settings
- The CommClust model's usefulness was demonstrated with different combinations of single variable, repeated measures cluster algorithms and community detection models
- Using a network-based approach, the CommClust algorithm additionally provided an intuitive output that aids researchers better understand the relationships between subjects

Contact Information

* matthew.d.koslovsky@nasa.gov
* caroline.m.schaefer@nasa.gov

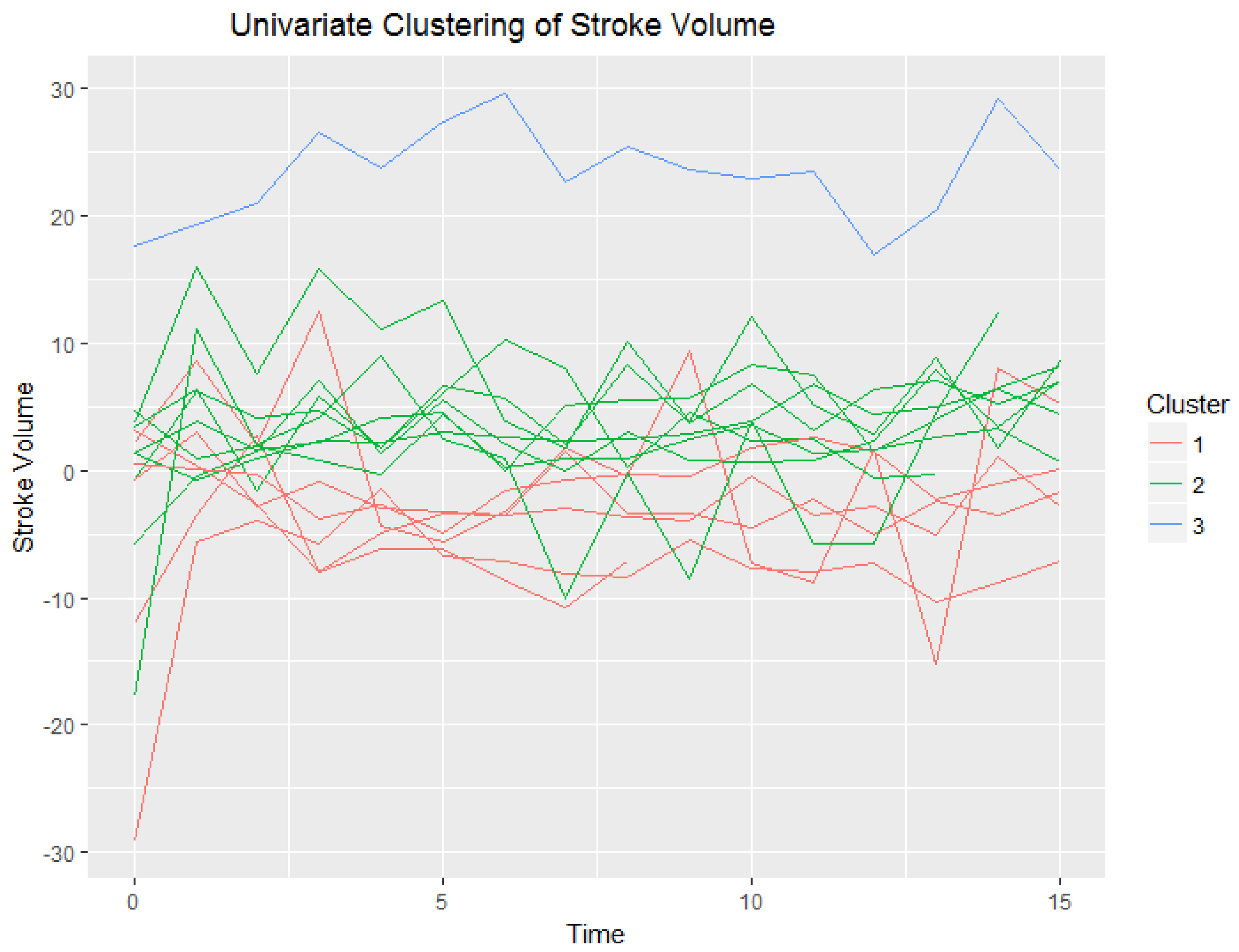
A network-based algorithm for clustering multivariate repeated measures data

Matthew Koslovsky[1,*], John Arellano[2], Caroline Schaefer[2,*], Alan Feiveson[3], Millennia Young[3], and Stuart Lee[1]

[1]KBRwyle, Houston, TX; [2]MEI Technologies, Houston, TX;
[3]NASA Johnson Space Center, Houston, TX

Study Background

- Parent study evaluated the efficacy of gradient compression garments (GCG) to prevent orthostatic intolerance after a 14-day 6° head-down tilt, bed rest[12]
- Eight cardiovascular measures were repeatedly collected during 15-min head-up tilt tests on BR-5, BR+0, BR+1, and BR+3
  - Heart rate (bpm), systolic blood pressure (mmHg), diastolic blood pressure (mmHg), plasma volume index (l/m2), stroke volume (ml), cardiac output (l/min), total peripheral resistance, and mean arterial pressure
- Treatment group wore GCG and thigh-high compression garments incrementally through BR+2
- Control group wore GCG from 6 am to ≈ 11 am on BR+0
- There was no discernible effect of the garments on responses to orthostatic testing on BR+3 without garments
- GCGs were beneficial when subjects were tilted head-up, helping maintain orthostatic tolerance and preventing tilt-induced increase in heart rate and decrease in stroke volume
- The aim was to assess the algorithm's ability to recover the treatment assignment, using only the subjects' repeated measures, cardiovascular data.



Contact Information

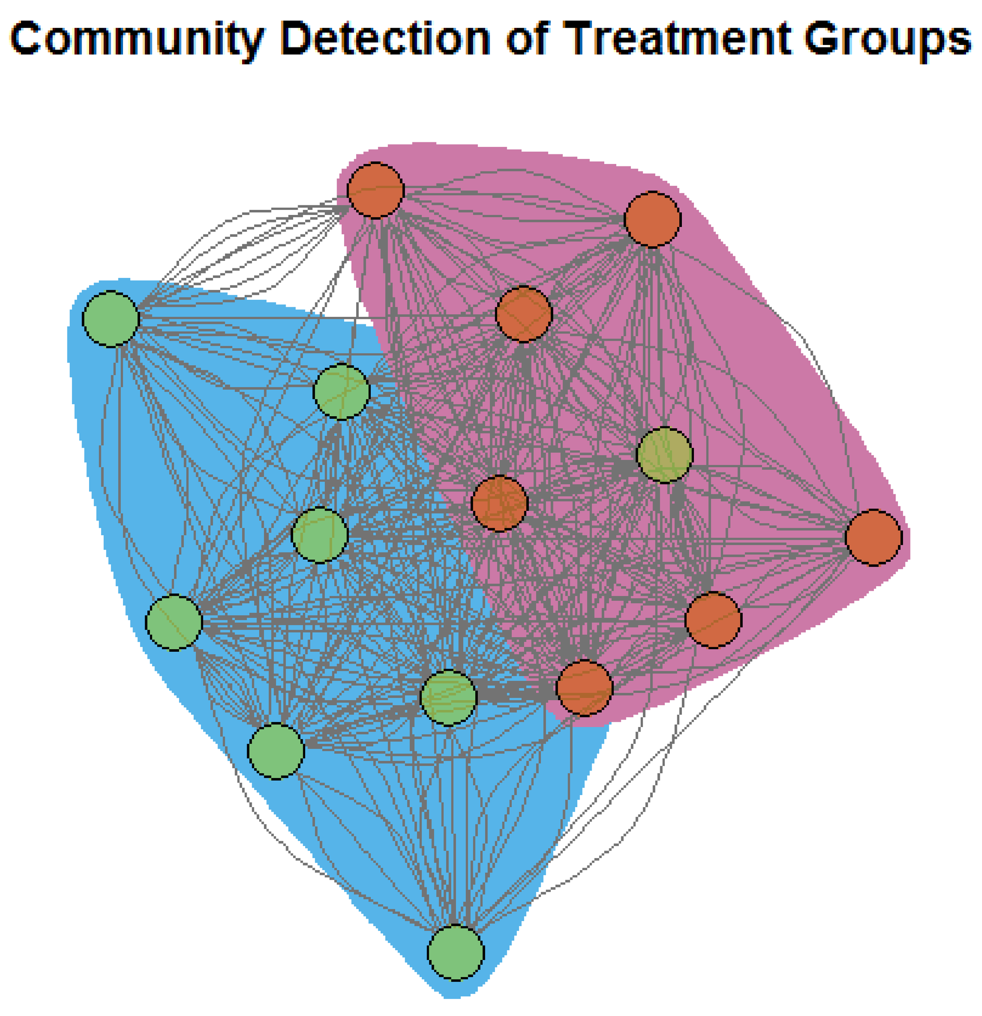
* matthew.d.koslovsky@nasa.gov
* caroline.m.schaefer@nasa.gov

Methods

- The best performing method overall from the simulations: CommClust with distance matrix based on non-parametric model parameters at three equally-spaced knots, using the leading eigenvector community detection algorithm
- CommClust was compared to KML3D with Jaccard indices and their overall correct categorization percentage

Results

- CommClust correctly assigned treatment groups to 14 out of 15 subjects (Jaccard index 0.75) (see below)
- KML3D assigned all but one subject to the same group (Jaccard index 0.43)



Node colors represent true grouping, shaded regions show how CommClust detected communities

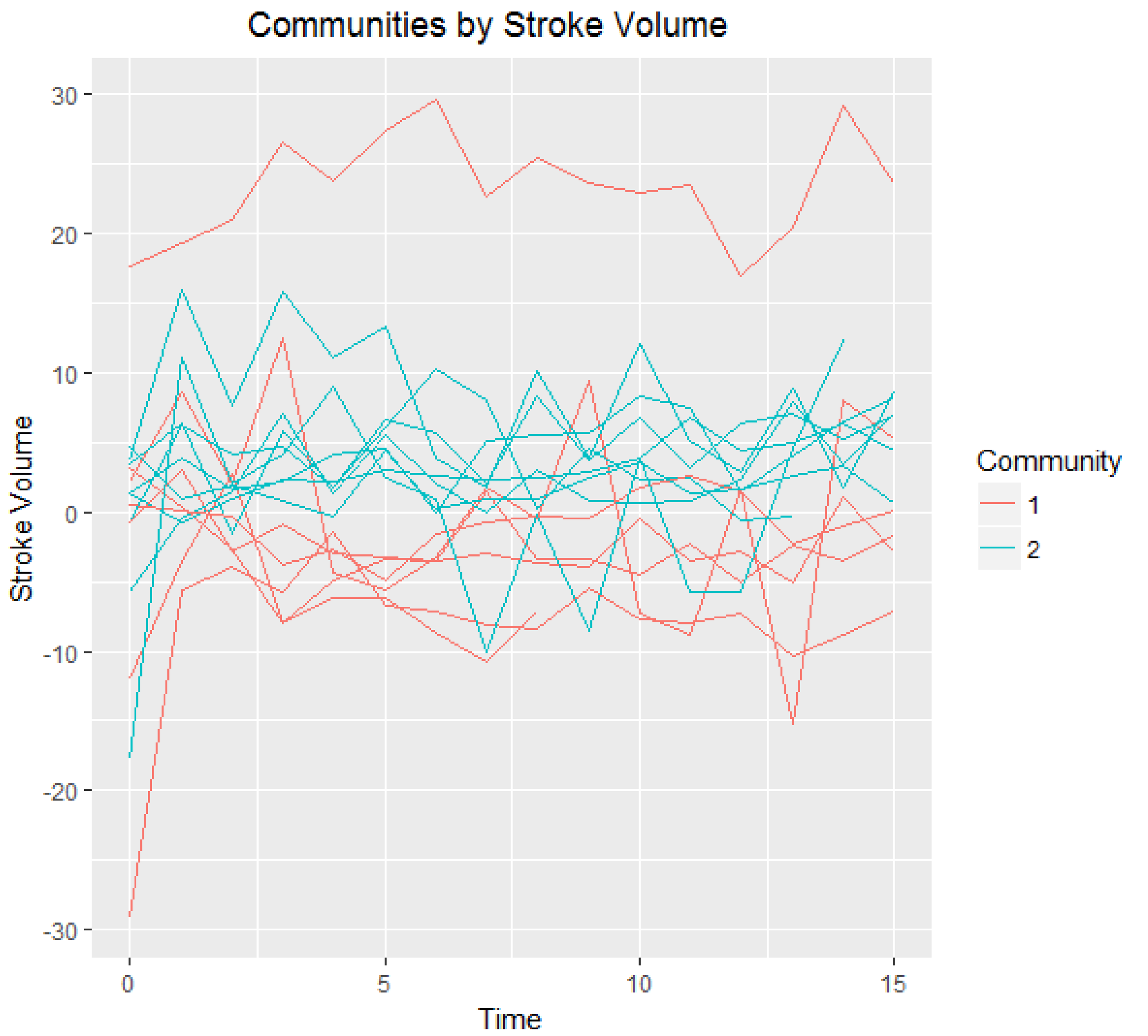
Results, continued

Table 3: Univariate clustering comparison with treatment group and CommClust output

	Heart Rate	SBP	DBP	MAP	PP	Stroke Volume	CO	TPR
Jaccard with Truth	0.32	0.27	0.17	0.20	0.21	0.64	0.32	0.20
Jaccard with Algorithm	0.44	0.24	0.17	0.18	0.24	0.88	0.37	0.20

Discussion

- CommClust was able to identify known treatment groups using a set of cardiovascular measures
- By synthesizing the univariate data collectively, CommClust was able to discriminate between the treatment and control group, whereas the univariate data alone could not
- CommClust can be used as a dimension reduction technique to identify groups of individuals who are at higher risk for a particular outcome of physiological change



A network-based algorithm for clustering multivariate repeated measures data

Matthew Koslovsky[1,*], John Arellano[2], Caroline Schaefer[2,*], Alan Feiveson[3], Millennia Young[3], and Stuart Lee[1]

[1]KBRwyle, Houston, TX; [2]MEI Technologies, Houston, TX;
[3]NASA Johnson Space Center, Houston, TX

References

[1] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, 3(1):1–27, 1974.

[2] Christophe Genolini and Bruno Falissard. KML: k-means for longitudinal data. *Computational Statistics*, 25(2):317–328, 2010.

[3] Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

[4] Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Finding Groups in Data: An Introduction to Cluster Analysis*, pages 68–125, 1990.

[5] Karen Leffondré, Michal Abrahamowicz, Armelle Regeasse, Gillian A Hawker, Elizabeth M Badley, Jane McCusker, and Eric Belzile. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. *Journal of Clinical Epidemiology*, 57(10):1049–1062, 2004.

[6] Paul D McNicholas and T Brendan Murphy. Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, 38(1):153–168, 2010.

[7] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

[8] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2):191–218, 2006.

[9] Peter J Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[10] WS Sarle. SAS Technical Report A-108. *The Cubic Clustering Criterion. Cary, NC: SAS Institute*, 1983.

[11] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[12] Michael B Stenger, Stuart MC Lee, L Christine Ribeiro, Tiffany R Phillips, Robert J Ploutz-Snyder, Michael C Willig, Christian M Westby, and Steven H Platts. Gradient compression garments protect against orthostatic intolerance during recovery from bed rest. *European Journal of Applied Physiology*, 114(3):597–608, 2014.

[13] Leland Wilkinson, Anushka Anand, and Robert Grossman. Graph-theoretic scagnostics. *In IEEE Symposium on Information Visualization*, 2005, pages 157-164, Washington, DC, USA, 2005. *IEEE C*omputer *S*ociety.

Contact Information

* matthew.d.koslovsky@nasa.gov
* caroline.m.schaefer@nasa.gov