A network-based algorithm for clustering multivariate repeated measures data
Matthew Koslovsky¹∗, John Arellano², Caroline Schaefer²∗, Alan Feiveson³, Millennia Young³, and Stuart Lee¹
¹KBRwyle, Houston, TX; ²MEI Technologies, Houston, TX; ³NASA 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

https://ntrs.nasa.gov/search.jsp?R=20170006954 2019-06-05T02:46:20+00:00Z
https://ntrs.nasa.gov/search.jsp?R=20170006954 2019-06-05T02:46:20+00:00Z
A network-based algorithm for clustering multivariate repeated measures data
Matthew Koslovsky1,*, John Arellano2, Caroline Schaefer2,*, Alan Feiveson3, Millenia Young3, and Stuart Lee1
1KBRwyle, Houston, TX; 2MEI Technologies, Houston, TX; 3NASA 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 vast amount of repeated measures data collected on the National Aeronautics and Space Administration (NASA) Astronaut Corps

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>
<thead>
<tr>
<th>Method</th>
<th>Univariate Clustering Algorithm</th>
<th># of Clusters</th>
<th>Community Detection</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KML Uncivariate [2]</td>
<td>Leading eigenvector (9)</td>
<td>KML-lec</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>LongclustEM [6]</td>
<td>Leading eigenvector</td>
<td>Longclust-lec</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Non-parametric model &amp; PAM [4]</td>
<td>Leading eigenvector</td>
<td>Non-lec</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Walktrap</td>
<td>Walltrap</td>
<td>Non-walk</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Scagnostics and PAM [13]</td>
<td>Leading eigenvector</td>
<td>Scalgo-legen</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Walktrap</td>
<td>Walltrap</td>
<td>Scalgo-legen</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Scagnostics and PAM [13]</td>
<td>Leading eigenvector</td>
<td>Scalgo-legen</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Walktrap</td>
<td>Walltrap</td>
<td>Scalgo-legen</td>
<td></td>
</tr>
</tbody>
</table>

Simulations

- Bivariate, repeated measures data, Yij = (Yij(1), Yij(2)), were simulated in 4 different scenarios, similar to [2]
- Scenarios included random individual i = 1, 2, ..., N, and measurement error, had various numbers of true clusters (g = 1, ..., G), and varied in the patterns of trajectories (e.g., linear, curvilinear)
- In each scenario, we assessed N = 10 and 50 individuals in each cluster n = 11 equally spaced assessment times tij (j = 1, 2, ..., n) in [0, 10]
- Scenario 3 was additionally analyzed with randomly spaced assessment times (3rd)
- Algorithm performance for Scenario 4 was evaluated with 20% and 50% missing data (4th and 5th)

Example Scenario 1:

\[
\begin{align*}
Y_{ij}^{(1)} &= 0 + b_1 + \epsilon_{ij} \\
Y_{ij}^{(2)} &= 0 + b_2 + \epsilon_{ij}
\end{align*}
\]

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

Simulation Results

Table 2: Best performing* model for each scenario

<table>
<thead>
<tr>
<th>Scenario N</th>
<th>Model</th>
<th>Jaccard Median</th>
<th>Time, s</th>
<th># of Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0.73 (0.23-1.00)</td>
<td>0.27 (0.01)</td>
<td>3.49 (1.62)</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>1.44 (0.06-1.00)</td>
<td>4.65 (0.11)</td>
<td>3.63 (0.92)</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>0.27 (0.14-0.60)</td>
<td>0.88 (0.09)</td>
<td>3.89 (0.92)</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>0.29 (0.13-0.62)</td>
<td>1.09 (0.17)</td>
<td>3.73 (1.09)</td>
</tr>
</tbody>
</table>

*Performance determined by highest Jaccard index [3] 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\textsuperscript{1,*}, John Arellano\textsuperscript{2}, Caroline Schaefer\textsuperscript{2,*}, Alan Feiveson\textsuperscript{3}, Millennia Young\textsuperscript{3}, and Stuart Lee\textsuperscript{1}

\textsuperscript{1}KBRwyle, Houston, TX; \textsuperscript{2}MEI Technologies, Houston, TX; \textsuperscript{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\degree head-down tilt, bed rest\textsuperscript{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/m\textsuperscript{2}), 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 \approx 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.

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)

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\textsuperscript{1,*}, John Arellano\textsuperscript{2}, Caroline Schaefer\textsuperscript{2,*}, Alan Feiveson\textsuperscript{3}, Millennia Young\textsuperscript{3}, and Stuart Lee\textsuperscript{1}

\textsuperscript{1}KBRwyle, Houston, TX; \textsuperscript{2}MEI Technologies, Houston, TX; \textsuperscript{3}NASA Johnson Space Center, Houston, TX

References


Contact Information

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