Clustering Discrete State Trajectories of Varying Lengths: Health Care Utilization Patterns

Laura Hatfield, PhD
Associate Professor, Harvard Medical School
ISPOR May 22, 2018
Baltimore, MD

Thank you to my collaborators

Brianna Heggeseth
• Williams College -> Macalester College

Megan Schuler
• RAND Boston

Nina Joyce
• Brown University

Haiden Huskamp
• Harvard Medical School

Elizabeth Lamont
• Harvard Medical School
Health care utilization trajectories

Medicare claims
14,257 people
Advanced lung cancer

Individual trajectories: sequences of days in each care setting

Diagnosis
Death

Can we use clustering to **discover** and **illustrate** variation in experiences?

Latent class analysis

For response pattern $y$ and class $c_k$

$$\Pr(Y = y) = \sum_{k=1}^{K} \Pr(C = c_k) \prod_{j=1}^{J} \Pr(Y_j = y_j | C = c_k)$$

The class indicators are missing data.
Four distinct classes

66% of people
Mainly at home
10 months survival

17% of people
Mainly in hospital and post-acute
1 month survival

11% of people
Mainly in hospice
4 months survival

6% of people
Mainly in ICU
1 month survival


Classes have distinct trajectories

Remaining methods gaps

<table>
<thead>
<tr>
<th>Limitations of feature extraction + LCA</th>
<th>A new distance measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discards ordering information</td>
<td>Uses sequence information directly</td>
</tr>
<tr>
<td>Requires good feature selection</td>
<td>Does not require feature selection by investigator</td>
</tr>
<tr>
<td>Sensitive to choice of features</td>
<td>Facilitates standard clustering methods</td>
</tr>
</tbody>
</table>

Distance is a weighted combination of

1. moving average of discordant days and
2. length difference

\[
d(a, b) = w \frac{1}{K} \sum_{k=1}^{K} \frac{|s(a_t | t \in (k, k + \tau)) - s(b_t | t \in (k, k + \tau))|}{2\tau} \\
+ (1 - w) \frac{|l(a) - l(b)|}{\max\{l(a), l(b)\}}
\]

\(s(a_t | t \in (k, k + \tau))\) Vector number of “days” in each state during time window of width \(\tau\)

\(\tau, w\) Bandwidth and weight tuning parameters
Conclusions

- Clustering can show variation in longitudinal data
- Feature extraction enables use of LCA clustering
- Custom distance measure enables other clustering methods

Thanks!

hatfield@hcp.med.harvard.edu

@laura_tastic
@HPDRLab

healthpolicydatascience.org