Simulating Sequences

Brendan Halpin
Department of Sociology
University of Limerick

Journée Trajectoires ’11, October 14 2011
Université Paris I

1Work in progress – brendan.halpin@ul.ie
Sequence analysis and lifecourse data

- What does SA do for us with lifecourse data?
- Life course data: long spells, few states, important individual characteristics and an “interesting” time dimension
- What does clustering pairwise inter-sequence do for us?
  - Descriptive overview, visualisation – enough?
  - Can it pick up things other techniques miss?
- Today I discuss using simulations, both “artificial” and data-based, to address this question
Sequences are messy

- Lifecourse sequences are epiphenomena of more fundamental underlying processes
- The processes are potentially complex: difficult to predict distribution of sequences
- Other techniques (hazard rate models, models of late outcome using history, models of the pattern of transition rates) give a powerful but partial view
- SA clearly allows us visualise complex data; possibly allows us observe features that will otherwise be missed
Processes

- The generating processes are complex:
  - individuals bring different characteristics from the beginning
  - history matters, including via duration dependence (individuals accumulate characteristics)
  - time matters:
    - calendar time (e.g. economic cycle), state distribution may change dramatically
    - developmental time (maturation)
    - processes in other lifecourse domains
- Too many parameters to model, hard to visualise distribution of life courses, also the possibility of emergent features
Outline of the presentation

- Use of “ideal” simulations to test how well SA can recover information about the generative processes
- Looking at an alternative visualisation of sequence structure: a time-dependent average transition matrix
- Using this structure to create data-based simulations, to ask more precise questions of the sequence analysis, including some simple hypothesis testing
Simple simulation

- The purpose of the first exercise is to simulate sets of sequences with a very simple, known structure, and examine how our usual practice (cluster analysis of pairwise OM distances) can recover this structure.
- A 3-state space, sequences 40 units long.
- Four simple scenarios:
  - Two distinct transition matrices, constant over time.
  - One transition matrix, but two different rates of transition.
  - One matrix, subgroup is initially faster, then slower.
  - One matrix, subgroup is forced to state 3 at a random point.
- Unrealistic, little structure, no history, time almost absent.
Simulation process

- The test is whether the cluster solution is associated with the generating type
- Association is a much weaker requirement than actually recovering the type information
- 1000 sequences generated at a time, different cluster solutions considered, underlying rate of transition varied
Summary result

Mean p-value of $\chi^2$ test, 16-cluster solution

<table>
<thead>
<tr>
<th>Monthly transition rate</th>
<th>Different matrices</th>
<th>Forced state at random point</th>
<th>Different rates of transition</th>
<th>Changing rates of transition</th>
</tr>
</thead>
<tbody>
<tr>
<td>2%</td>
<td>0.037</td>
<td>0.000</td>
<td>0.048</td>
<td>0.041</td>
</tr>
<tr>
<td>5%</td>
<td>0.055</td>
<td>0.000</td>
<td>0.026</td>
<td>0.011</td>
</tr>
<tr>
<td>10%</td>
<td>0.090</td>
<td></td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td>15%</td>
<td>0.107</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18%</td>
<td>0.133</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These distributions are skewed, so the proportion significant is higher than the average would suggest. The 0.037 in the first simulation corresponds with 60% of cases with $p < 0.01$. 
More detail: scenario 1

- Sequences are assigned to a random starting point
- Two transition regimes:

<table>
<thead>
<tr>
<th></th>
<th>Type 1</th>
<th></th>
<th>Type 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>0.80</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>B</td>
<td>0.10</td>
<td>0.80</td>
<td>0.10</td>
</tr>
<tr>
<td>C</td>
<td>0.10</td>
<td>0.10</td>
<td>0.80</td>
</tr>
</tbody>
</table>

- State distribution won’t change, is the same across type, mean number of spells is the same across type
- “Threshold” parameter to vary monthly rate of change
- Type 2 will simply generate more A, B, C sequences etc., than type 1
Distribution of $\chi^2$ p-values, by rate of transition and cluster size

Graphs by n and Threshold
Association, not recovery of pattern

- We that for the 16-cluster solution we have a very good chance of detecting the structure
- However, we can’t “recover” the structure so easily
- Inspection of the cluster solution is interesting
- But even enumerating sequences according to their structure (e.g., counts of AB, BC and CA transitions) will only correctly identify about 60% of them
Cluster solution

<table>
<thead>
<tr>
<th>1,1</th>
<th>1,2</th>
<th>2,1</th>
<th>2,2</th>
<th>3,1</th>
<th>3,2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,1</td>
<td>4,2</td>
<td>5,1</td>
<td>5,2</td>
<td>6,1</td>
<td>6,2</td>
</tr>
<tr>
<td>7,1</td>
<td>7,2</td>
<td>8,1</td>
<td>8,2</td>
<td>9,1</td>
<td>9,2</td>
</tr>
<tr>
<td>10,1</td>
<td>10,2</td>
<td>11,1</td>
<td>11,2</td>
<td>12,1</td>
<td>12,2</td>
</tr>
<tr>
<td>13,1</td>
<td>13,2</td>
<td>14,1</td>
<td>14,2</td>
<td>15,1</td>
<td>15,2</td>
</tr>
<tr>
<td>16,1</td>
<td>16,2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Graphs by g16 and type
Cluster solution: some key groups

Graphs by g16 and type
Parameterising OM

- We can use this framework to explore the effect of indel and substitution costs.
- Indel costs can vary from half the max substitution cost up – low values make it easy to detect similarity at different times, higher values reduce OM to Hamming distance.
- The base simulation uses a substitution cost based on the three states being equally different ("flat" or 2-D solution).
- What if $A \rightarrow C$ is twice $A \rightarrow B$ ("linear" or 1-D solution).
Parameterising OM: Varying indel costs

(800 sequences, 50 replications)
Parameterising OM: Varying substitution costs

(1000 sequences, 50 replications)
Time-warping

- An alternative distance measure to OM
- Local expansion and compression of the time dimension
- Subtly different logic, similar to implement, somewhat different results
- Described (as TWED) by Marteau (2007, 2008), implemented as a Stata plugin
- See http://teaching.sociology.ul.ie/seqanal/naplestw.pdf for more info
- What difference does it make here?
TWED and OM in 2-matrix scenario

600 sequences, 200 replications
TWED and two-speed scenario

Graphs by Threshold

OM, indel=0.5

TWED, l=0.15

TWED, l=0.25

TWED, l=0.50

800 seqs, 50 reps
Simulating Sequences

Brendan Halpin
Department of Sociology
University of Limerick

Exploring lifecourse data
Potentially complex processes
Pure simulations
Scenario 1
Parameterising OM
Time-warping
Time-dependent transition regime
Visualising transition rates
Data-based simulation

TWED and forced event scenario

Graphs by n and Threshold

800 seqs, 50 reps
TWED and fast-then-slow

Graphs by n and Threshold

8, 2%

8, 5%

op ap
bp cp
TWED vs OM

- In the two-matrix scenario, TWED performs a little better than OM at low numbers of spells, but in contrast to OM improves as the number of spells increases.
- TWED is also competitive with OM in the two-speed and fast-then-slow scenarios.
- The forced event scenario shows OM doing well: OM may be better at dealing with states at approximate times, TWED better at recognising sequence.
- Conclusion: The distance measure matters and OM is not the last word.
Taking transition rates seriously

- The foregoing simulations hinge on the structure of transition rates
- Let’s apply this to real data:
  - First as a visualisation of the temporal structure of transitions
  - Second using this structure as a base for simulations against which we compare the reality.
Sequences are generated by a complex, messy set of processes

Individual difference, state dependence and temporal change are all likely to be important

However, we can readily account for time by calculating the $M \times M \times (T - 1)$ transition structure
Example: Mothers’ labour market sequence data

- Five years labour market history of women who have a birth at end of year 2
- Simple chronogram is informative but incomplete
- Cluster solution gives a digestible but messy overview
Mothers’ labour market sequence 8-cluster solution

1

2

3

4

5

6

7

8

FT
PT
UE
NE

Exploring lifecourse data
Potentially complex processes
Pure simulations
Scenario 1
Parameterising OM
Time-warping
Time-dependent transition regime
Visualising transition rates
Data-based simulation
Exploring lifecourse data
Potentially complex processes
Pure simulations
Scenario 1
Parameterising OM
Time-warping
Time-dependent transition regime
Visualising transition rates
Data-based simulation

Temporal structure of transitions

Simulating Sequences
Brendan Halpin
Department of Sociology
University of Limerick

FT

PT

UE

NW
Cluster solution, MVAD data

Exploring lifecourse data
Potentially complex processes
Pure simulations
Scenario 1
Parameterising OM
Time-warping
Time-dependent transition regime
Visualising transition rates
Data-based simulation
Simulating Sequences
Brendan Halpin
Department of Sociology
University of Limerick

Exploring lifecourse data
Potentially complex processes
Pure simulations
Scenario 1
Parameterising OM
Time-warping
Time-dependent transition regime
Visualising transition rates
Data-based simulation

Temporal structure, MVAD data
### Simulating Sequences

Brendan Halpin  
Department of Sociology  
University of Limerick

Exploring lifecourse data  
Potentially complex processes  

Pure simulations  
Scenario 1  
Parameterising OM  
Time-warping  

Time-dependent transition regime  
Visualising transition rates  

Data-based simulation

### Comparing real and simulated, $\chi^2$ test $p$-values

<table>
<thead>
<tr>
<th>Groups</th>
<th>Births</th>
<th>MVAD</th>
<th>IMS</th>
<th>BHPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>mean</td>
<td>0.388</td>
<td>0.559</td>
<td>0.502</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td>0.384</td>
<td>0.617</td>
<td>0.509</td>
</tr>
<tr>
<td>4</td>
<td>mean</td>
<td>0.003</td>
<td>0.415</td>
<td>0.506</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td>0.000</td>
<td>0.377</td>
<td>0.502</td>
</tr>
<tr>
<td>8</td>
<td>mean</td>
<td>0.000</td>
<td>0.262</td>
<td>0.291</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td>0.000</td>
<td>0.188</td>
<td>0.153</td>
</tr>
<tr>
<td>16</td>
<td>mean</td>
<td>0.000</td>
<td>0.024</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>32</td>
<td>mean</td>
<td>0.000</td>
<td>0.000</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
## Single run of birth sequences

<table>
<thead>
<tr>
<th>G2</th>
<th>G4</th>
<th>G8</th>
<th>G16</th>
<th>Simulated</th>
<th>Real</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.74</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>χ² p-values (this run)</td>
<td></td>
</tr>
<tr>
<td>0.39</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
<td>χ² p-values (mean over 50 runs)</td>
<td></td>
</tr>
</tbody>
</table>

### Table:

<table>
<thead>
<tr>
<th>G2</th>
<th>G4</th>
<th>G8</th>
<th>G16</th>
<th>Simulated</th>
<th>Real</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.74</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>χ² p-values (this run)</td>
<td></td>
</tr>
<tr>
<td>0.39</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
<td>χ² p-values (mean over 50 runs)</td>
<td></td>
</tr>
</tbody>
</table>

### Diagram:

- **Key:**
  - Full-time work
  - Part-time work
  - Unemployed
  - Not available

### Data:

1. 1 1 1 1
2. 1 1 1 2
3. 1 2 2 3
4. 1 2 2 4
5. 2 3 3 5
6. 2 3 3 6
7. 2 3 4 7
8. 2 3 5 8
9. 2 3 5 9
10. 2 3 5 10
11. 2 3 6 11
12. 2 4 7 12
13. 2 4 7 13
14. 2 4 7 14
15. 2 4 7 15
16. 2 4 8 16
Simulated vs Real

KEY: Full-time work  Part-time work  Unemployed  Not available

1 1 1
1 1 2
2 2 3
2 2 4
3 3 5
3 3 6
3 4 7
3 5 8
3 5 9
3 5 10
3 6 11
4 7 12
4 7 13
4 7 14
4 7 15
4 8 16