Sequence Analysis in Sociology

Brendan Halpin
Department of Sociology
University of Limerick
brendan.halpin@ul.ie

Helsinki, May 19, 2010

Introduction

What this workshop is intended to cover

- Explore the use of Sequence Analysis (SA) in the social sciences, particularly Optimal Matching
- Explore the analytical use of “empirical typologies” and other measures derived from SA
- Consider alternatives to Optimal Matching (OM)
- Provide enough practical information for participants to conduct their own analyses using software
- How to think about using SA

What this workshop is intended to cover

- Explore the analytical use of “empirical typologies” and other measures derived from SA
  - use of OM to generate classifications for further analysis
  - or to generate other sorts of trajectory level information

What this workshop is intended to cover

- Consider alternatives to Optimal Matching (OM)
  - Non-aligning methods
  - Combinatorial methods
  - Duration-sensitive methods

What this workshop is intended to cover

- Provide enough practical information for participants to conduct their own analyses using software
  - Using SQ add-on for Stata
  - Using my faster but less user-friendly add-on for Stata
  - Using TraMineR for the R statistical language

How to think about using SA

- What's it good for?
- When to use other methods?
- What method to choose?
- How to adapt it for your problem

Key questions

- Is sequence analysis useful?
  - Does it go beyond exploratory and descriptive?
  - If not, is that enough?
- What does it "mean"?
  - How do the results inform us about sociological issues?
  - How can we manipulate the inputs to get better meaning?
  - How should we choose between algorithms for different substantive problems?

What is Sequence Analysis?

- “Sequences” are temporal (or at least linear) trajectories through a state space
- Sequence Analysis treats sequences “holistically”
- Alternative methodologies are more analytical (and often stochastic) and focus on factors such as the generative processes
- Usually at the cost of ignoring some aspect of the information encoded in the sequence
- In contrast, SA tends to be blind to the processes generating the sequence, thus focusing on the epiphenomenal?

What are sequences?

- Sequences are ordered “trajectories” through a state space
- Their nature depends on
  - the nature of the state space: multi-dimensional, continuous, real, categorical?
  - the nature of the time dimension: atomic, discrete, continuous, stretchable or rigid?
  - how they start and finish, what they mean substantively

What are sequences? – state spaces

- The state space is important for how we think about distances between points
  - if multi-dimensional in $\mathbb{R}^n$, point distances are naturally Euclidean or other function of the space
  - if categorical, we often define pairwise point distances *a priori* or empirically (how?)
  - if many categories?

What are sequences? – the nature of time

- The nature of time has a large bearing on the adequacy of sequences as a representation of the phenomenon
  - “Atomic” sequences consist of elements that are naturally separate and sequential, such as a series of purchases or votes, or steps or utterances, or CAGT bases in DNA – “time” is naturally discrete
  - Continuous time can be discretised
    - If the state changes very frequently (especially if continuous state) we can consider this as “sampling”, for example, digitisation of an audio stream
    - If change is relatively rare, we may wish to represent the sequence as a series of spells (start and end-times of a period in which the state is constant)

What is a whole sequence?

- How we define a "whole" sequence is also an important issue – where does it start and end?
- For some sequences, it is natural:
  - The steps of a dance, the words of a song
  - The rhetorical structure of a journal article or a folk tale
- For some purposes, fragmentary sequences can be used (e.g., searching for DNA matches)

"Conventional" methods for longitudinal data

- Many conventional approaches
  - Hazard rate modelling (event history analysis)
  - Time-series analysis
  - Loglinear or Markov modelling of transition rates
  - Start-end tables
  - Panel analysis (cross-sectional time-series analysis, multiple-response approaches)
  - Use of simple summaries of trajectory to predict future outcomes
- All have analytical strengths – allow inference with respect to clear hypotheses
- In what way does SA offer something more than they do?

SA versus conventional

- Clearly a classification based on SA will do better than one based on summaries such as start/finish state or cumulated duration in states – respects order
- Relative to hazard rate modelling, SA respects the whole trajectory, rather than looking at time to a single event (note the existence of repeated events hazard models)
- Models based on transition rates have difficulty with transition matrices which change in complex ways through time (due to life course and period effects, for instance)
- Multi-dimensional sequences are even more complex
- None of the conventional methods offer a digestible descriptive overview

Special considerations for life-course sequences

- For life-course and other sequences, the requirements of the analysis impose structure
  - Usually cannot just match random segments of employment history
  - We impose comparability criteria (e.g., t_0 is a specific event, follow until a particular outcome or for a specified duration)
  - Issues of left- and right-censoring become relevant
- The various SA methods tend to be blind to these issues

Space consequences

- For life-course and other sequences, the requirements of the analysis impose structure
- Depending on the nature of the state space and the time dimension different approaches will be required
- A $\mathbb{R}^n$ state space simplifies state distance issues compared with categorical states, where we need to find a justification for our distances

Time consequences

- Time that is naturally discrete fits a token-sequence representation naturally
- “Sampling” continuous time raises issues of distortion due to the frequency of sampling
- Whether time has a “ruler” or calendar, or is “developmental” or stretchable, has a bearing on how attractive alignment is

What do we do with SA: distances (1/2)

- SA is very simple: turn information about the state space into information about the trajectories – pairwise distances or similarities
- Permits the use of cluster analysis (CA) to generate a data-driven classification
- Permits comparing all sequences to a set of “typical” sequences
- Permits analysis of the multi-dimensional space implied by the inter-trajectory distances (multi-dimensional scaling, MDS)
- Also permits comparing grouped or paired sequences (e.g., couple’s work histories, Han and Moen, 1999)

What do we do with SA: distances (2/2)

- Well chosen ideal-typical sequences can make for very interpretable results
- CA may or may not generate a useful classification
- MDS can inform CA, may yield interpretable dimensions itself

What do we do with SA: empirical typology

- Why is data-driven classification – an empirical typology – attractive?
- Up to $\sum_{i=1}^{m} n^i$ possible sequences, with n states and spells up to m tokens long (e.g., for 4 states over 20 months, more than a trillion possibilities)
- Observed sequences represent a highly structured subset
- The structure is much more than, say, that summarised by
  - starting state distribution and
  - the transition matrix averaged over the data set (or even changing through time)
- A good classification should (?) pick up this structure

Empirical/theoretical typology

- If we can generate a typology of sequences from theory, we may not need SA
- However, can be difficult to write foolproof rules to assign sequences to groups
- SA linking the observed to the ideal typical sequences will allow us to populate a theoretical typology automatically
- Classification by inspection may be possible, but can be impractical

Empirical typology by CA

- Cluster analysis on the pairwise distance matrix creates the classification automatically for us
- But can be difficult to characterise the resulting groups cleanly
- And clustering may be unstable if there are not natural groups

Exploratory/descriptive: enough?

- Key questions: can it be more than exploratory and descriptive?
- Is exploration/description enough justification?

Distance measures

How do we define distance?

- How do we go from distances within a state space to distances between sequences in that space?
- How do we think about the sociological meaningfulness of inter-sequence distance?
- Clearly the more a pair of sequences go through the same or similar states, at the same or similar times, the more similar they are
- However, there are multiple ways to approach measuring this relationship

Measures of distance

- Counting element-wise matches is intuitive, but only picks the *same* states at the *same* time

$$d = l - \sum_{i=1}^{i=l} (x_{1i} == x_{2i})$$

- The Longest Common Prefix method (e.g., Elzinga, 2009) is even “worse”, as it only picks up common states at the beginning

Measures of distance

- Hamming distance utilises information about the state-wise differences so it weights the *same and similar* states at the same time

$$d_H = \sum_{i=1}^{i=l} (|x_{1i} - x_{2i}|)$$

or

$$d_H = \sum_{i=1}^{i=l} \delta(x_{1i}, x_{2i})$$

where $\delta(a, b)$ is the distance between state a and state b

- However, such methods do not recognise similarity of states at *similar but not identical times* – no “alignment” or sliding along

Similarity at same or similar times

- One early method (Degenne et al., 1996) to pick up similarity at similar times has re-emeged recently as "Qualitative Harmonic Analysis" (Robette and Thibauld, 2008)
- This groups the sequence into short intervals and summarises the state distribution within them
- Data reduction (correspondence analysis) is used to simplify these summaries, and the results are clustered
- The same states at the same or different times *within the short intervals* contribute to similarity
- Intuitively appealing, though how to choose interval length and the nature of the summaries may need to be theorised

sociology
UNIVERSITY OF LIMERICK

DT distance

- Dijkstra and Taris (1995) suggested a method that focused on having the same states in a similar order
  - With a pair of sequences, drop all states that do not occur in both
  - In each sequence, drop duplicate states
  - Count how many moves are needed to reorder one state to match the other
- An intuitively satisfying focus on *same states in similar order*, and a tractable algorithm, but has problems (Abbott, 1995a)
- Not least is throwing away lots of useful information

sociology
UNIVERSITY OF LIMERICK

The Optimal Matching Algorithm

- Going back to computer science work in the 1950s and 1960s by Levenshtein, and described in the 1980s in Sankoff and Kruskal (1983), the Optimal Matching Algorithm was introduced into sociology by Andrew Abbott (e.g., Abbott and Forrest, 1986; Abbott and Hrycak, 1990; Abbott, 1995b; Abbott and Tsay, 2000)
- A very general method for comparing token sequences
- Can recognise same and similar states via its *substitution cost matrix*
- Can recognise similarity that is out of alignment via *insertion and deletion*
- Using the Needleman–Wunsch algorithm to implement it, is quite efficient in computing terms

sociology
UNIVERSITY OF LIMERICK

OMA dominant

- OMA is the leading method for SA in social research
- Has received much criticism (some deserved), e.g., from Levine (2000) and Wu (2000)
- Recurrent themes:
  - that while its application in molecular biology is successful, it has poor sociological meaning
  - that there is no good basis for choosing substitution costs or insertion–deletion costs

sociology
UNIVERSITY OF LIMERICK

Combinatorial approaches

- The main competition to OM comes from Cees Elzinga (e.g., Elzinga, 2003, 2005, 2009)
- Coming from the Dijkstra–Taris tradition, at least in motivation, his methods focus on the extent to which sequences pass through
  - the same states
  - in the same order (not necessarily consecutively)
- Different logic and rather different results to OM
- Beginning to appear in the literature (e.g., Aisenbrey and Fasang, 2010)

Optimal Matching Algorithm

The Optimal Matching Algorithm

- We now look more closely at the Optimal Matching Algorithm
- How it operates
- How to carry out analysis with it

The logic of OM

- At base, OM has a very simple logic: the “cost” of “editing” one sequence to match another
- Two sequences which differ at one point have a distance proportional to the difference between the points
  - `ABC` and `ABD` have a difference related to $\delta(\texttt{C}, \texttt{D})$, the cost of substituting `C` with `D`
- Two sequences which differ in that one has an extra element differ –
  - in relation to the cost of deleting the extra element
  - which is entirely equivalent to inserting it in the sequence where it is absent

Substitution, indel, concatenation

- The former operation is called *substitution*, the latter *indel*
- Any sequence can be turned into any other by a concatenation of these operations
- The OM algorithm identifies the “cheapest” concatenation that achieves this
- “Cheap” implies cost: substitutions and *indel*s need to have weights applied

Costing substitutions

- Substitution costs may be determined in many ways
    - From *a priori* knowledge about the relations within the state space
    - Derived from data about the state space
    - Derived from observed transitions in the state space
    - Or just give up:

$$c_s = \begin{cases} 0 & (x_i = x_j) \\ 1 & (x_i \neq x_j) \end{cases}$$

Costing *indels*

- The cost of insertion and deletion affects how prone to “alignment” the algorithm will be
- If *indel*s are cheap they will often be chosen in preference to substitutions
- If they are sufficiently expensive they will only be used to equalise unequal sequence lengths
    - In the case of equal-length sequences and high *indel* costs, OM returns the same distance as the Hamming measure (no alignment – similar states at same times)

Subs and *indels*

- There are some key limits with respect to the relation between substitution and *indel* costs
- Since substitution is equivalent to an insertion plus a deletion, substitution costs greater than 2 × *indel* will have no effect
- How high *indel* costs have to be to completely suppress alignment of equal-length sequences seems to depend on the data
    - In my experience, more than about 1.5 to 2 times the largest substitution cost is often enough
- Approx. working range: $0.5 < \frac{indel}{\max(c_s)} < c. 1.5 \rightarrow 2$

Programming OM

- Determining the cheapest set of "elementary operations" is potentially complex – a large population of candidates
- However, it can be stated as a recursive problem and programmed very efficiently
- Understanding how it is programmed can help understand the principle of OM

sociology
UNIVERSITY OF LIMERICK

OM: Recursive problem

$$\Delta_{OM}(A^p, B^q) =$$

$$min \begin{cases} \Delta_{OM}(A^{p-1}, B^q) & + indel \\ \Delta_{OM}(A^{p-1}, B^{q-1}) & + \delta(a_p, b_q) \\ \Delta_{OM}(A^p, B^{q-1}) & + indel \end{cases}$$

(Δ represents distance between sequences, and δ differences within the state space)

sociology
UNIVERSITY OF LIMERICK

Implementing the recursive algorithm

Cell value: $min(c_{i-1,j-1} + \omega_{i,j}, c_{i,j-1} + \Delta, c_{i-1,j} + \Delta)$
$= min(0 + 2, 2 + 2, 2 + 2) = 2$
$= min(2 + 1, 2 + 2, 4 + 2) = 3$
$= min(4 + 0, 3 + 2, 6 + 2) = 4$
$= min(6 + 1, 4 + 2, 8 + 2) = 6$

s_2

		A	B	C	D
	C	2	1	0	1
s_1	D	3	2	1	0
	A	0	1	2	3
	A	0	1	2	3
	B	1	0	1	2

0	2	4	6	8
2	2	3	4	6
4	4	4	4	4
6	4	5	6	6
8	6	5	7	8
10	8	6	6	8

sociology
UNIVERSITY OF LIMERICK

Tracing the operations

To convert `ABCD` into `CDAAB` the following set of operations gives the cheapest path:

Operation	Intermediate state	Cost
Sequence 2	`ABCD`	= 0
insert `C`	`CABCD`	+2 = 2
insert `D`	`CDABCD`	+2 = 4
const `A = A`	`CDABCD`	+0 = 4
subs `B`→`A`	`CDAACD`	+1 = 5
subs `C`→`B`	`CDAABD`	+1 = 6
delete `D`	`CDAAB-`	+2 = 8
Sequence 1	`CDAAB`	= 8

sociology
UNIVERSITY OF LIMERICK

Standardising on length

- Where sequence may be of different lengths, the distance is usually divided by the length of the longer sequence
- In this case, 8 units thus become a pairwise distance of 1.6

Looking at alignments

- It may be interesting to look at the resulting alignments, but this is done much less in social science than in other contexts
- Note that more than one “cheapest” route may exist
- Hence there may be no single “best” alignment

OM and Hamming on example sequences

- In the next slides I present OM and Hamming distances on selected short sequences
- Using two cost configurations:
  - “Flat”, i.e. all states equally different

A	0	1	1	1
B	1	0	1	1
C	1	1	0	1
D	1	1	1	0

indel=1

  - “Linear”, i.e. all states on a single dimension

A	0	1	2	3
B	1	0	1	2
C	2	1	0	1
D	3	2	1	0

indel=2

Examples, “flat” substitution costs

```
AAAAAAAA |   0                                   OM distance
AAABBBCD |   5     0
ABCDDDDD |   7     6     0
BAAACCDD |   5     5     6     0
BBDDACCC |   7     7     6     5*    0
DDDDABCD |   7     5     6     6*    4     0
DCCCBBAA |   6     6     7     8     8     6     0
DDABCDAA |   5     6     6*    6     6**   4**   5     0

AAAAAAAA |   0                                   Hamming distance
AAABBBCD |   5     0
ABCDDDDD |   7     6     0
BAAACCDD |   5     5     6     0
BBDDACCC |   7     7     6     6     0
DDDDABCD |   7     5     6     7     4     0
DCCCBBAA |   6     6     7     8     8     6     0
DDABCDAA |   5     6     7     6     8     6     5     0
```

Examples, "linear" substitution costs

```
AAAAAAAA |    0                                   OM distance
AAABBBCD |    8     0
ABCDDDDD |   18    10     0
BAAACCDD |   11     5     9     0
BBDDACCC |   14    10     8    9*     0
DDDDABCD |   18    12   10*   13*     6     0
DCCCBBAA |   11    13    15    16    11     9     0
DDABCDAA |   12    14   14*    13  10**   8**     7     0

AAAAAAAA |    0                                   Hamming distance
AAABBBCD |    8     0
ABCDDDDD |   18    10     0
BAAACCDD |   11     5     9     0
BBDDACCC |   14    10     8    11     0
DDDDABCD |   18    12    12    15     6     0
DCCCBBAA |   11    13    15    16    11     9     0
DDABCDAA |   12    14    16    13    16    14     7     0
```

Some interesting sequence pairs

- AAAAAAAA with anything: no "order" implies no possible improvement from alignment
- DDABCDAA with ABCDDDDD: alignment reduces cost somewhat (14 vs 16)

```
D D A B C D A A - -
- - A B C D D D D D
2 2 0 0 0 0 3 3 2 2
```

Some interesting sequence pairs

- BBDDACCC with DDABCDAA: alignment strongly reduces cost (10 vs 16)

```
B B D D A C C C - -
- - D D A B C D A A
2 2 0 0 0 1 0 1 2 2
```

- DDDDABCD with ABCDDDDD: alignment reduces cost only for the more diverse substitution cost configuration (10 vs 12)

```
- - D D D D A B C D
A B C D D D - - D D
2 2 1 0 0 0 2 2 1 0
```

Applying Optimal Matching

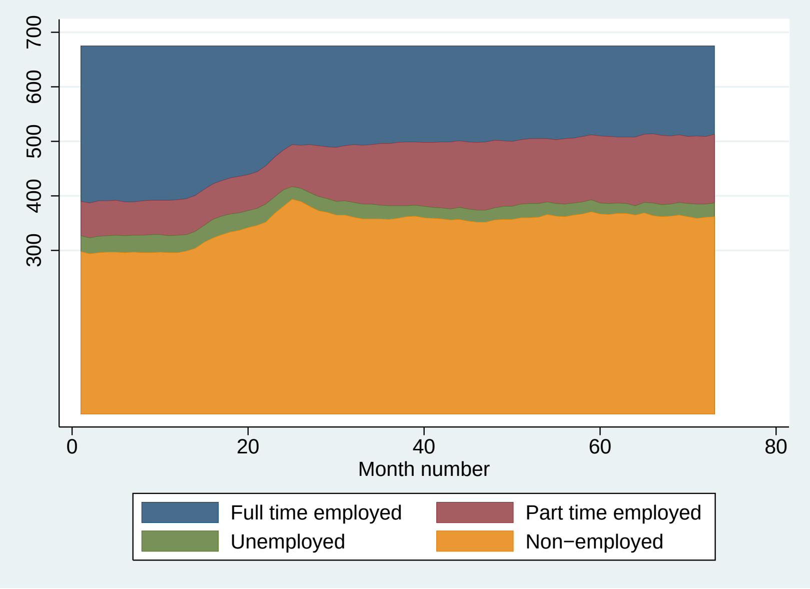
An example: birth and the labour market

- An application of OM: 5 years of labour market history of women who have a birth in month 25
- BHPS data – note that even with a quite large sample observing people for c.15 years, only 675 cases fit the criteria
- We classify labour market status thus:

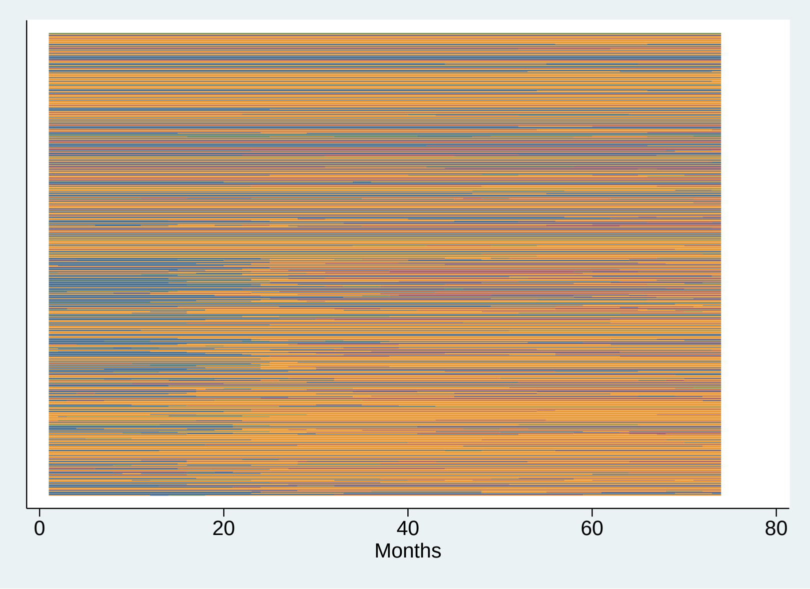
Full-time employed	🟦	0	1	2	3
Part-time employed	🟥	1	0	1	2
Unemployed	🟩	2	1	0	1
Not in labour market	🟧	3	2	1	0

- *indel* cost is 2

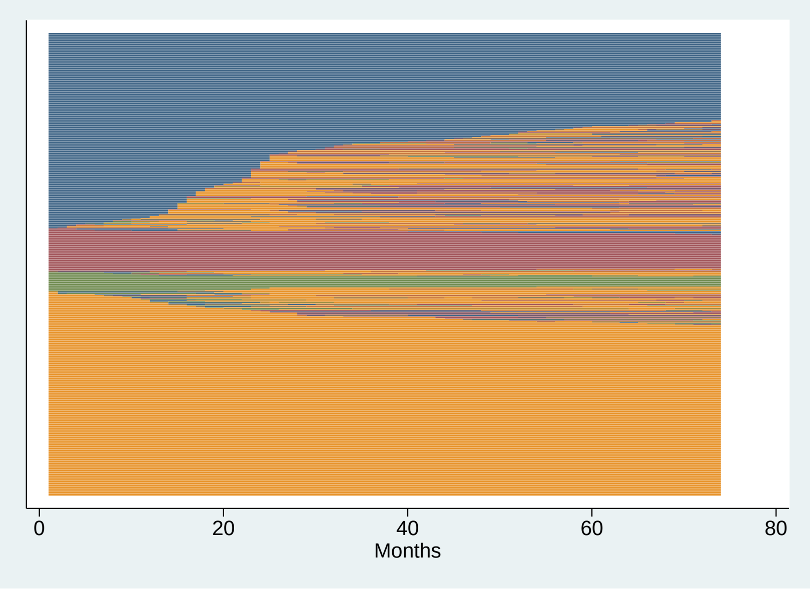
State distribution or "chronogram"



Index plot, unordered



Index plot, lexically sorted



SQ summary data

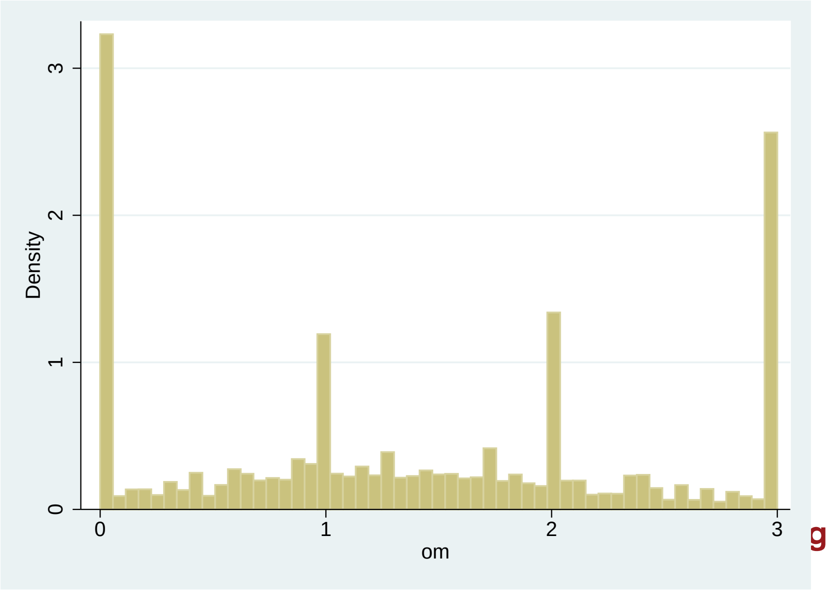
```
  # of observed sequences: 675
overall # of obs. elements: 4
       max sequence length: 73
 # of producible sequences: 8.920e+43

----------------------------------------------------------
Observations |    Sequences  % of observed           Cum.
-------------+--------------------------------------------
           1 |          211       31.25926       31.25926
           2 |            5       .7407407             32
           3 |            1       .1481481       32.14815
           5 |            2       .2962963       32.44445
          15 |            1       .1481481       32.59259
          50 |            1       .1481481       32.74074
         127 |            1       .1481481       32.88889
         249 |            1       .1481481       33.03704
             |
       Total |          223       33.03704
----------------------------------------------------------
```

SQ Tabulating sequences

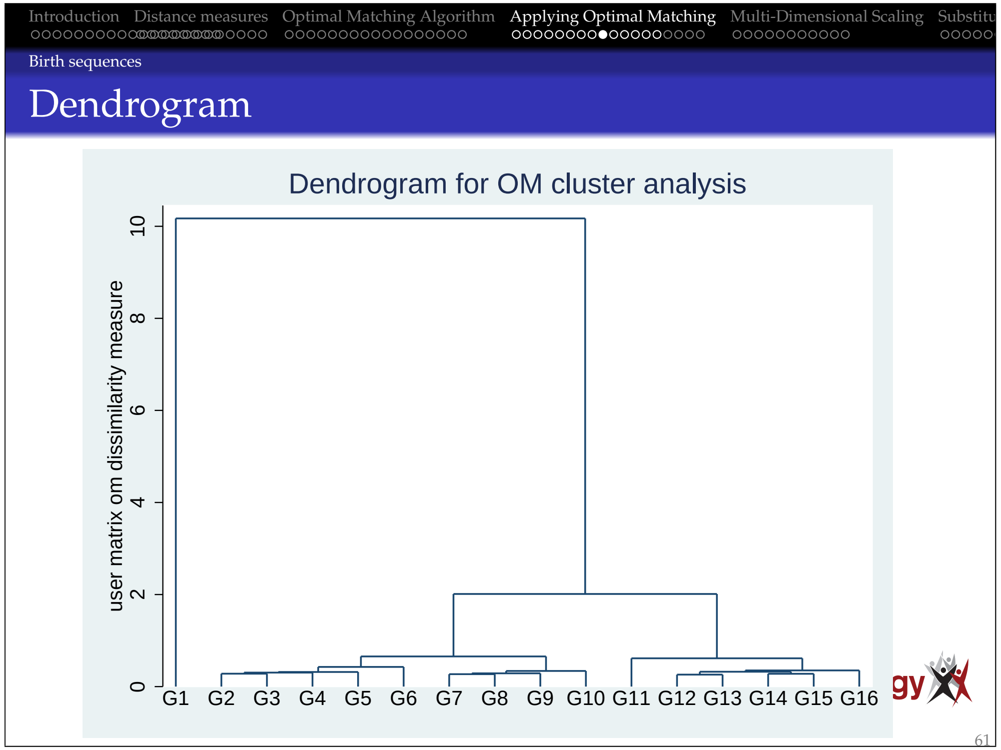
```
                     Sequence-Pattern |      Freq.     Percent        Cum.
--------------------------------------+-----------------------------------
4444444444444444444444444444444444444 |        249       36.89       36.89
1111111111111111111111111111111111111 |        127       18.81       55.70
2222222222222222222222222222222222222 |         50        7.41       63.11
3333333333333333333333333333333333333 |         15        2.22       65.33
1111111111111111111114444444444444444 |          5        0.74       66.07
1111111111111111111144444444444444444 |          5        0.74       66.81
1111111111111111111444444444444444444 |          3        0.44       67.26
1111111111111111111111111111111111111 |          2        0.30       67.56
1111111111111111111111111111111111111 |          2        0.30       67.85
1111111111111111111111111144444444444 |          2        0.30       68.15
 . . .                                        . . .       . . .       . . .
4444444444444444444444444444444444444 |          1        0.15       99.56
4444444444444444444444444444444444444 |          1        0.15       99.70
4444444444444444444444444444444444444 |          1        0.15       99.85
4444444444444444444444444444444444444 |          1        0.15      100.00
--------------------------------------+-----------------------------------
                                Total |        675      100.00
```

OM distances

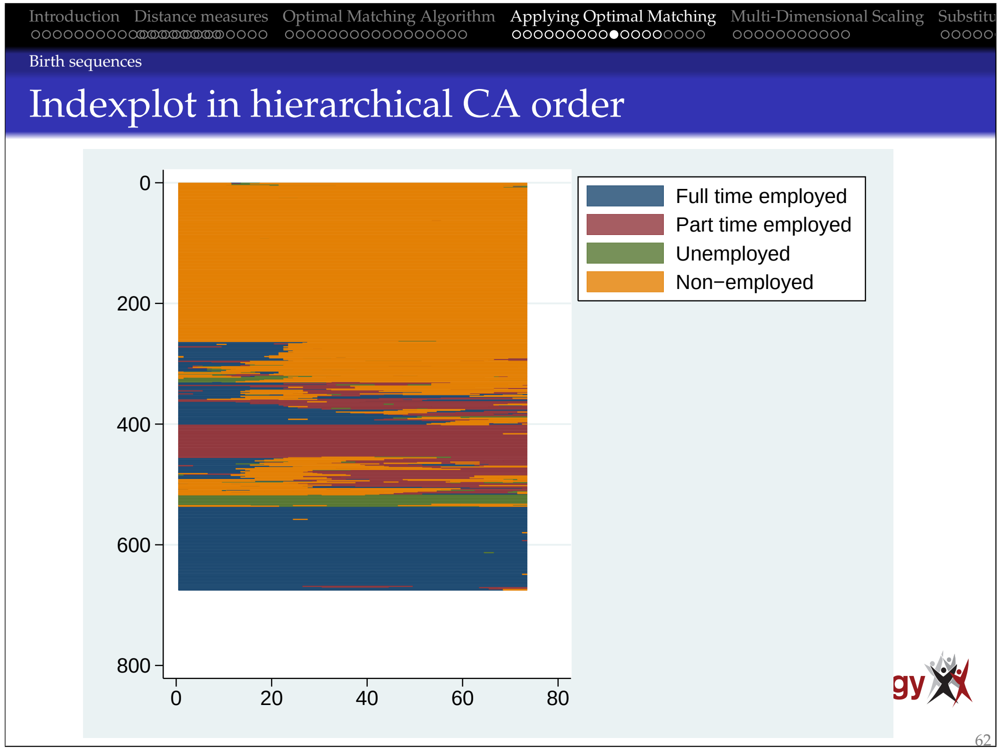


Cluster analysis

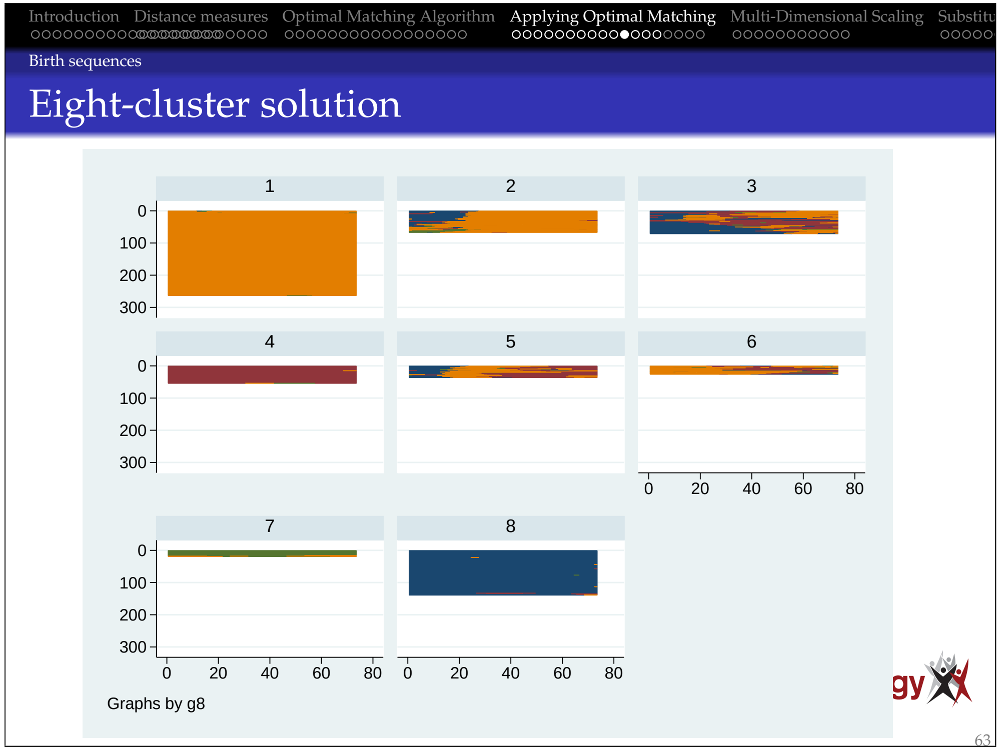
- We proceed by using cluster analysis
- Wards' method – reasonably stable, widely available, tends to produce relatively balanced groups
- Hierarchical – nested classifications sometimes an advantage
- Hierarchical – how to define the appropriate stopping level?



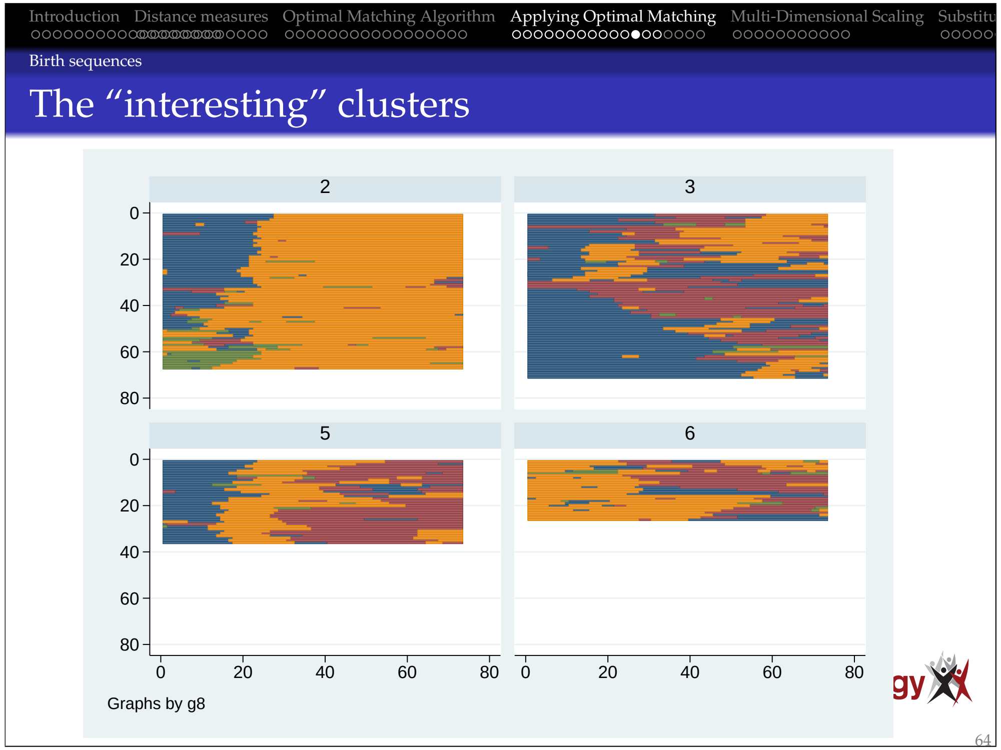
Dendrogram



Indexplot in hierarchical CA order

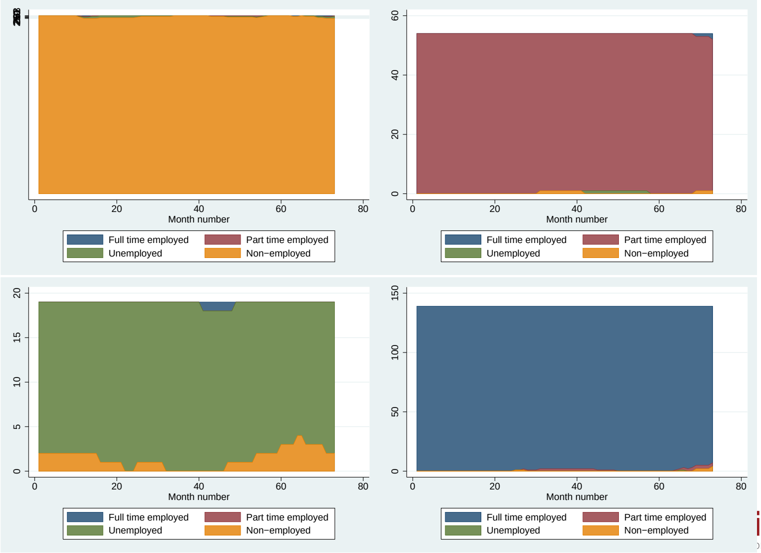


Eight-cluster solution

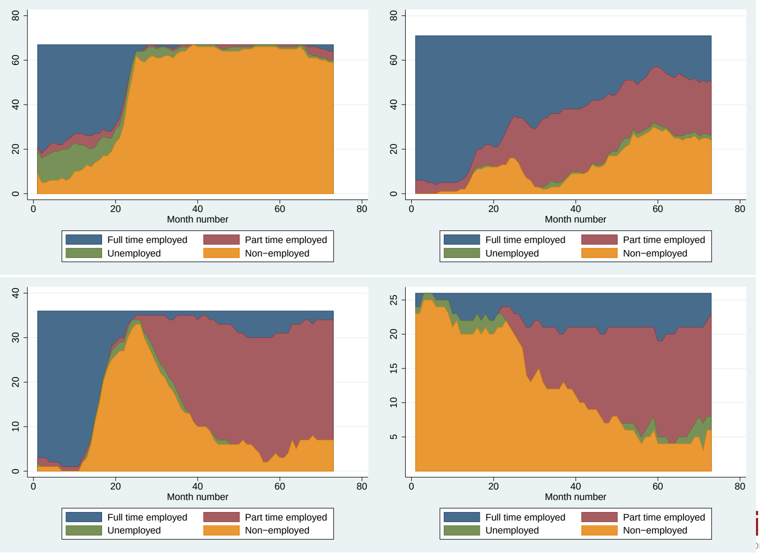


The “interesting” clusters

Chronogram, “boring” clusters (1, 4, 7 & 8)



Chronogram, interesting clusters (2, 3, 5 & 6)



Typical sequences

- We may also wish to characterise the cluster by a “typical” sequence
- Some cluster approaches supply this readily
  - The centroid
  - or medoid
  - or otherwise-defined sequence closest to the cluster centre
- Wards’ method with Stata doesn’t make this easy
- An alternative is the “modal” sequence

sociology
UNIVERSITY OF LIMERICK

Modal sequences

- The modal sequence is composed of the most common token at each time point
- Note this is a synthetic sequence, not drawn from the observed sequences
- While it summarises the time-ordered distribution it does not necessarily reproduce real transistions
- In fact, it can have improbable or impossible features

sociology
UNIVERSITY OF LIMERICK

Modal sequences for BS data

```
Cluster                                    Modal sequence
1     nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
2     FFFFFFFFFFFFFFFFFFFFFnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
3     FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF  ppF pppnnnnnnnnnnppp nn nnp
4     ppppppppppppppppppppppppppppppppppppppppppppppppppppppppppppppppppppppppp
5     FFFFFFFFFFFFFFFFnnnnnnnnnnnnnnnnnnppppppppppppppppppppppppppppppppppppppp
6     nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnppppppppppppppppppppppppppppppppp
7     UUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUU
8     FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

- On the whole they pick up the main features of the clusters
- Note gaps for cluster 3: no single mode for some months
- However, lots of information discarded

Other summaries

Cluster	Average months in				Average		N
	FT	PT	UE	NonE	N-spells	Turbulence	
1	0.0	0.1	0.3	72.6	1.11	1.17	263
4	0.1	72.3	0.3	0.3	1.11	1.20	54
7	0.4	0.0	67.2	5.4	1.63	2.17	19
8	72.4	0.4	0.0	0.1	1.12	1.24	139
2	14.2	2.0	4.7	52.0	3.21	5.14	67
3	38.1	20.4	1.1	13.4	3.69	6.40	71
5	19.8	30.3	1.2	21.7	4.17	7.73	36
6	11.7	22.5	2.5	36.3	4.31	7.23	26

Multi-Dimensional Scaling

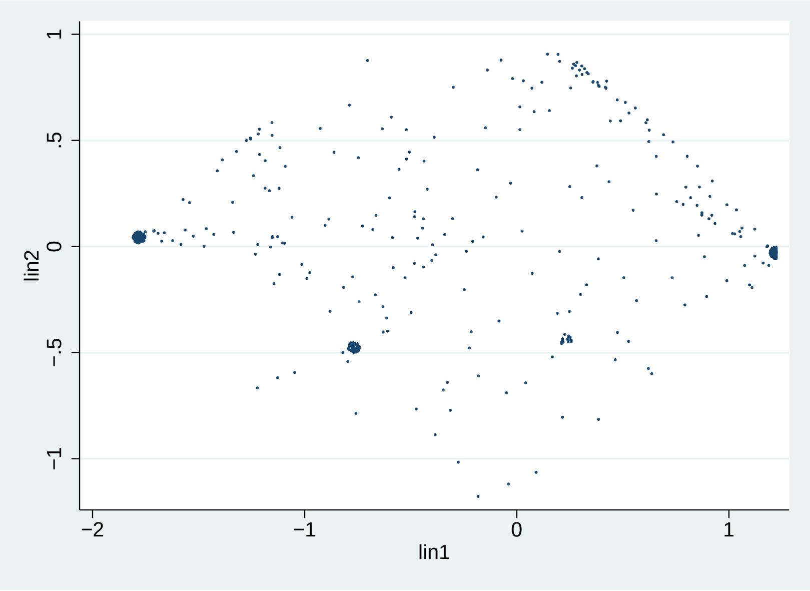
Alternative to CA: MDS

- Cluster analysis is not the only option for analysis of distance matrices
- Multi-dimensional scaling (MDS) also works on the space implied by the distances
  - Rather than group cases based on closeness,
  - attempt to extract the structure of the space
  - Relatively few “dimensions” rather than many pairwise distances
- The extracted dimensions may be meaningful and useful for further analysis
- Throws light on the cluster analysis

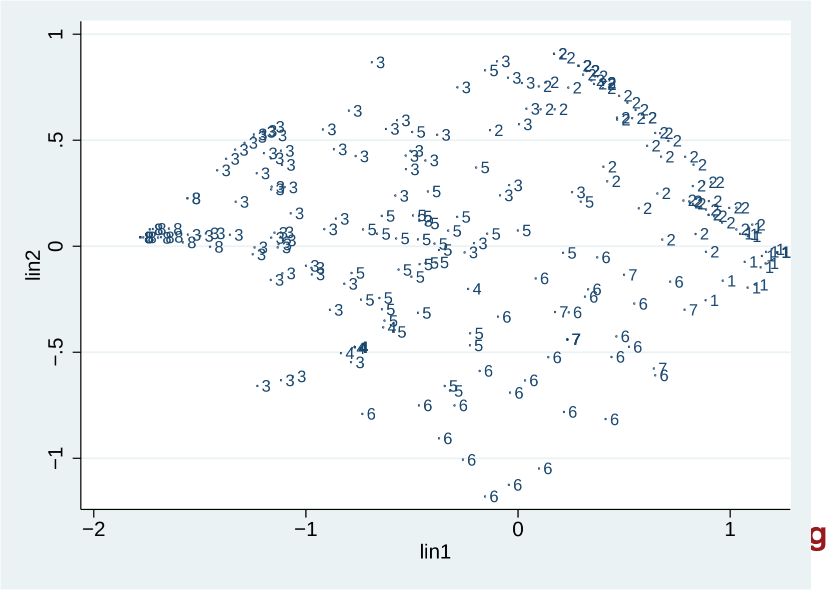
MDS in Stata

```
Classical metric multidimensional scaling
 dissimilarity matrix: omlin
                                           Number of obs        =        675
 Eigenvalues > 0       =         95        Mardia fit measure 1 =     0.7953
 Retained dimensions   =          3        Mardia fit measure 2 =     0.9937
 --------------------------------------------------------------------------
             |                 abs(eigenvalue)          (eigenvalue)^2
   Dimension |  Eigenvalue    Percent     Cumul.      Percent     Cumul.
 -------------+------------------------------------------------------------
           1 |   964.93863      72.25      72.25        98.79      98.79
           2 |   67.007783       5.02      77.27         0.48      99.27
           3 |   30.236075       2.26      79.53         0.10      99.37
 -------------+------------------------------------------------------------
           4 |   21.574791       1.62      81.15         0.05      99.41
           5 |   15.707688       1.18      82.33         0.03      99.44
           6 |   9.2522026       0.69      83.02         0.01      99.45
           7 |   7.5743302       0.57      83.59         0.01      99.46
           8 |   5.9310838       0.44      84.03         0.00      99.46
           9 |   4.5038763       0.34      84.37         0.00      99.46
          10 |   4.1936765       0.31      84.68         0.00      99.46
 --------------------------------------------------------------------------
```

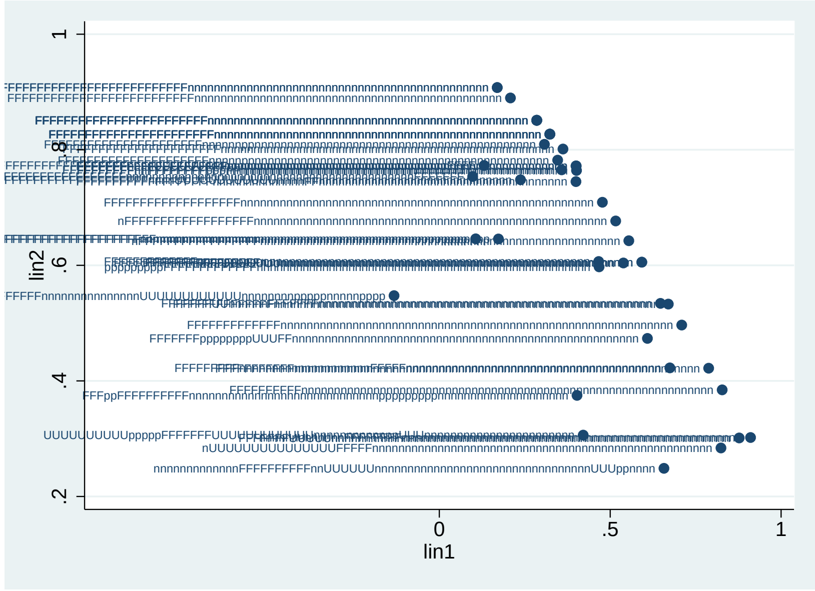
Linear state, first 2 dimensions



First 2 dimensions, with cluster membership



A “string” in cluster 2



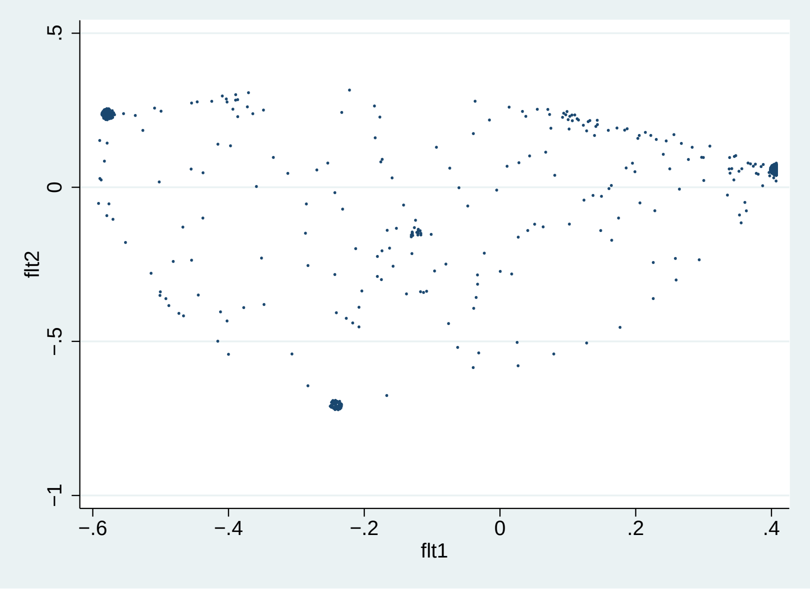
Main features

- The homogenous clusters (1, 4, 7 and 8) are distinctly located
- Clusters with substantial amounts of more than one state are located between these "vertices" and are quite diffuse
- Some evidence of "strings" – adjacent trajectories that differ slightly
- Quite clear structure, but what does it tell us about the meaningfulness of grouping? – simple sequences are distinct but complex sequences are more evenly distributed

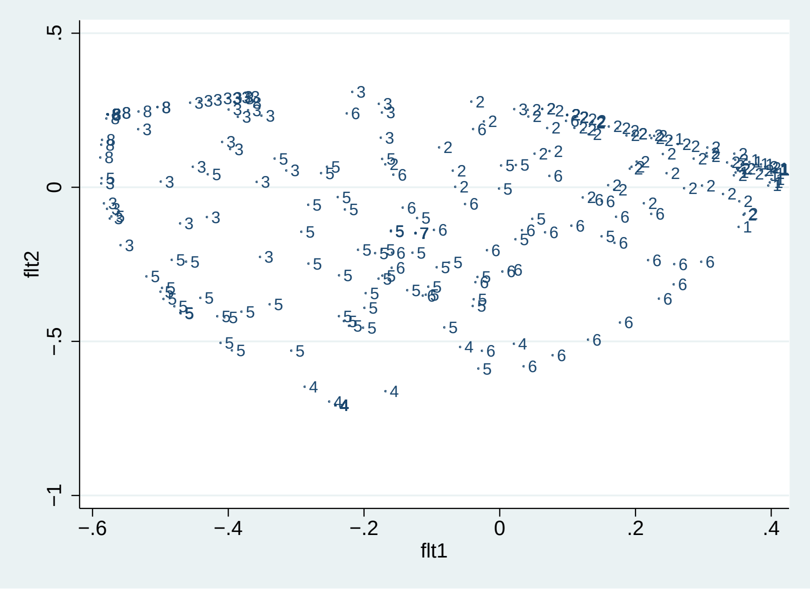
Main features: dimensions

- Dimension 1 is strongly related to the dimension of the state space
  - Set FT 0, PT 1, UE 2, NonE 3
  - Sum over the time-span to give a weighted cumulative duration
  - Correlation with first dimension is 0.9999
- Dimension 2 has non-employed who become part-timers at the low end, and the transition to non-employment at the top end

"Flat" state, first 2 dimensions



First 2 dimensions, with cluster membership



Main features in the “flat” state space

- Like before, the homogenous clusters are distinct and the others spread between them, with strings etc.
- Overall shape like a distortion of first one
- Dimension 1 is nearly as strongly related to the dimension of the state space
    - Correlation 0.9947
- Dimension 2 runs from 100% part-time to full-timers who exit late to non-employment (100% FT also high on D2)
- Affected by the state space *and* the nature of the trajectories
- Trajectory state space is structured by the states, each forming a pole

MDS useful

- As we see MDS throws light on some of the clustering processes
- Distinct tight clusters of simple sequences
- Systematically located intermediate sequences – but not in “obvious” groups
- Cluster analysis discriminates but cluster membership is sensitive to small changes
- The main dimensions are often interpretable, and may in some circumstances be useful as variables in further analysis

Substitution and indel costs

The *problem* of substitution costs

- Does Optimal Matching make sense for sociological data?
    - Is the algorithm itself suitable?
    - How to parameterise it: substitution and *indel* costs
- Repeated claims in the literature:
    - that sociologists don’t know how to set substitution costs,
    - that we can’t match the effectiveness of molecular biology
- Yes, our analytical goals are often much less well defined than those of the biologists
- No, substitution costs are not an intractable problem

A problem?

Criticism and puzzlement

- Wu (2000) treats the choice of substitutions costs as an insurmountable – while he has some misunderstandings that made it particularly difficult for him, it is an important stage of the OM process
- Many other writers agonise over the problem
  - Many opt for using transition rates to get data-driven cost
  - Or set all substitution costs equal to 1
- Neither option is really neutral – we need to understand substitution cost setting better
- Similarly, how to set *indel* costs?

Mapping s to S

Mapping states to sequences

- The essence of SA is mapping a view of a state space onto a view of a trajectory space: $\delta(s) \rightarrow \Delta(S)$
- We start with *knowledge* or a *view* of how states relate to each other (what states are like each other, what states are dissimilar)
- With a suitable algorithm we map this perspective onto trajectories through the state space: what trajectories are more or less similar
- The nature of the algorithm determines
  - Whether the mapping makes sense
  - Exactly how the structure of the state space affects the structure of the trajectory space

Mapping s to S

OMA coherent?

- Can we expect OMA to provide a coherent $\delta(s) \rightarrow \Delta(S)$ mapping?
- Elementary operations are intuitively appealing:
  1. $\Delta(\mathtt{ABC}, \mathtt{ADC}) = f(\delta(\mathtt{B}, \mathtt{D}))$
  2. $\Delta(\mathtt{ABCD}, \mathtt{ABD}) = f(\mathit{indel})$
  3. minimising concatenation of these two operations to link any pair of trajectories
- If 3 is reasonable, 1 and 2 determine how state space affects trajectory space

Mapping s to S

Thinking about state spaces and distances

- Costs can be thought of as distances between states
- If state space is $\mathbb{R}^n$, distance is intuitive
- If state space is categorical, how define distance?
  1. State space as efficient summary of clustered distribution in $\mathbb{R}^n$: distances are between cluster centroids
  2. State space can be mapped onto specific set of quantitative dimensions; each state located at the vector of its mean values; Euclidean or other distances between vectors
  3. States can be located relative to each other on theoretical grounds

Transitions and substitutions

- Transition rates frequently proposed as basis for substitution costs
- Critics of OMA complain of substitution operations implying impossible transitions (e.g., Wu, 2000)
- Even proponents of OMA are sometimes concerned about "impossible" transitions (e.g., Pollock, 2007),
- But substitutions are not transitions, **not even a little bit!**
    - substitutions happen across sequences, $\Delta(\texttt{ABC}, \texttt{ADC}) = f(\delta(\texttt{B}, \texttt{D}))$ (similarity of states)
    - transitions happen within sequences (movement between states)

sociology
UNIVERSITY OF LIMERICK

Informative transition rates

- No logical connection between substitutions and transition rates
- but under certain circumstances transition rates can inform us about state distances
- If state space is a partitioning of an unknown $\mathbb{R}^n$, movement is random (unstructured), and the probability of a move is inversely related to its length, then
- Distance between states will vary inversely with the transition rates
- However, these conditions are often not met

sociology
UNIVERSITY OF LIMERICK

Deceptive transiton rates

- Example: using voting intentions as a way of defining inter party distances
- UK: relatively high Con–LibDem two-way flows; ditto Lab–LibDem
- But Con–Lab transitions much lower: implies a potentially incoherent space (non-metric, more below)
    - $\delta(\text{Con}, \text{Lab}) > \delta(\text{Con}, \text{LibDem}) + \delta(\text{LibDem}, \text{Lab})$

sociology
UNIVERSITY OF LIMERICK

Confusing state and trajectory information

- This procedure confuses party state space and voter characteristics
- Voter polarisation/loyalty is trajectory information, not state information
- There is a strong analytical argument for trying to keep the two concepts as separate as possible
- Another type of problem: irrelevant distinctions can cause similar states to have low transition rates

sociology
UNIVERSITY OF LIMERICK

Take “space” seriously

- Very useful to think in spatial terms
  1. State space as efficient summary of clustered distribution in $\mathbb{R}^n$
  2. State space mapped onto specific set of quantitative dimensions
  3. State space defined on theoretical grounds
- For 1 and 2, explicitly multidimensional, in case 2 dimensions are explicit
- For 1 and 3, we can attempt to recover the implicit dimensions

Looking at state spaces

- Two very simple state spaces:
  - Single dimension, equally spaced:

0	1	2	3
1	0	1	2
2	1	0	1
3	2	1	0

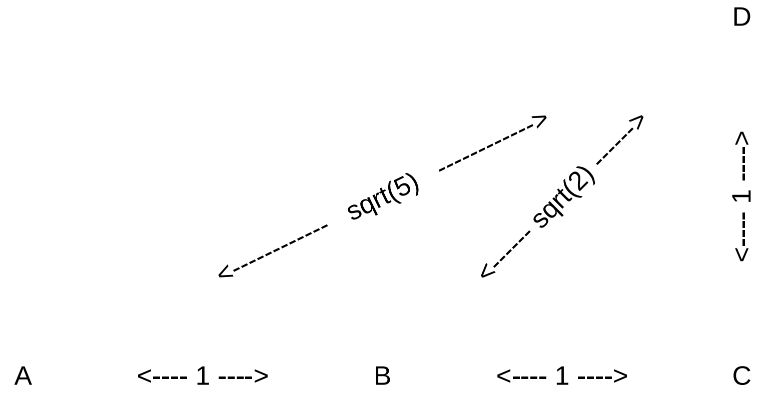
  - All states equidistant – $n - 1$ dimensions

0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	0

More dimensions

- E.g., 2D picture of inter-party distances: location on left–right scale, plus on pro-/anti-EU scale
- Distances are Euclidean or other metric (e.g., L1)
  - Euclidean: $\sqrt{\sum_i (r_i - s_i)^2}$
  - L1 (city block): $\sum_i |r_i - s_i|$
- Generalises easily to many dimensions
- Problem: how to weight different dimensions?
  - Scale by standard deviation? Substantive importance?

2-D example


D
sqrt(5)
sqrt(2)
<---- 1 ---->
A
<---- 1 ---->
B
<---- 1 ---->
C

Structure passes through

- State space structure passes through to trajectory space structure
  - Distances between states clearly affect distances between trajectories containing high proportions of those states
    - If $\delta(\texttt{"A"},\texttt{"B"}) << \delta(\texttt{"A"},\texttt{"C"})$ then $\Delta(\texttt{"..AAAA.."},\texttt{"..BBB.."})$ will tend to be less than $\Delta(\texttt{"..AAAA.."},\texttt{"..CCC.."})$
  - Differential distances promote alignment: `AADDAAA` and `AAADDAA` are more likely to be aligned to match the `DD` if $\delta(\texttt{"A"},\texttt{"D"})$ is large
  - If the state distances are non-metric, the trajectory distances may also be non-metric (at least between trajectories consisting of near 100% one state)
  - Unidimensional states spaces will tend to be reflected strongly in 1st principle component of trajectory space

Empirical comparison

- For an empirical test of the effect of different cost regimes, see `http://teaching.sociology.ul.ie/ofpr/` (similar notes from a longer course)

Designing state spaces

- Be explicit about state spaces and what distances mean
- Think spatially
  - Choose high or low dimensions, but have your reasons
- Simplify state space as far as possible
  - Drop irrelevant distinctions
  - Drop longitudinal information: let the sequence encode the temporal information, make state space cross-sectional

Dropping temporal information

- e.g., Simplify marital status:

	Living alone	Living with partner
Legally married	Separated	Married
Not legally married	Single, never married, post-cohabitation, divorced	Cohabiting

- The sequence will distinguish adequately between the various “single” states
- Parity sequences: Women’s annual fertility history
  - in parity terms: `000112333344444`
  - in birth event terms: `000101100010000`

Indel costs

- Finally, to adress *indel* costs
- As previously described, there is a specific lower limit and an empirical upper limit:

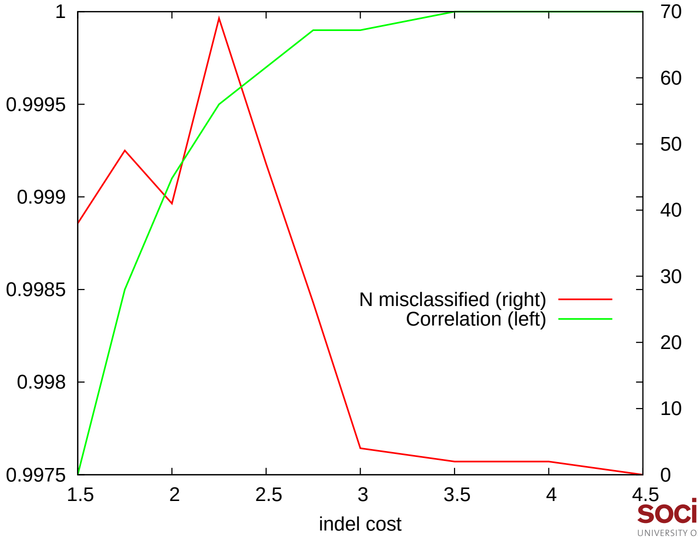
$$0.5 < \frac{\textit{indel}}{\max(c_s)} < c.\,1.5 \rightarrow 2$$

- Substitution costs greater than twice *indel* are ignored
- Raising *indel* costs to as little as twice the highest substitution cost will tend to prevent alignment

Varying indel and closeness to Hamming result

- The lower limit of the indel cost is 0.5 max substitution cost
- What of the top limit? How high does it need to get to reduce OM to Hamming distance?
- With the same data set, I present the effect of varying the *indel* cost from 1.5 to 4.5 with the linear substitution cost

Varying indel and closeness to Hamming result



Varying indel results

- The number of cases “misclassified” by CA relative to Hamming is somewhat chaotic up to about 2.5, and then falls sharply
- The correlation between OM and Hamming distances moves more steadily, hitting 1.000 at 3.5
- Note, though, the misclassification of 10% of the cases even though the correlation is 0.9995! CA can be funny.

Conclusions

- Substitution costs make a big difference
  - but largely understandable in operation
  - and an asset – more meaningful state space, more meaningful trajectory space
- Think spatially! Use data and geometric models
- Simplify
- Let the sequence do the temporal work

Alternatives

Elzinga’s combinatorial approach

- The main competitor to OM is from Cees H. Elzinga
- In a series of papers (Elzinga, 2003, 2005, 2009; Elzinga and Liefbroer, 2007; Elzinga et al., 2008), he proposes a number of related approaches with a different logical and mathematical underpinning
- In the tradition of Dijkstra and Taris (1995), he focuses on “the same states, in the same order”
- His novelty and power is to bring to bear set theory and combinatorics

Substrings and subsequences

- The intuition is that two sequences are more alike, the more they have the same states in the same order
- This explicitly brings the focus on sub-sequences (as distinct from substrings)
  - A subset of elements from the string
  - Not necessarily contiguous
  - But retaining the same order
- `AB` is both a substring and a subsequence of `ABC`
- `AC` is a subsequence but not a substring of `ABC`

Enumerating common subsequences

- Enumerating subsequences is the key to this approach
- A number of measures are proposed including
  - The Longest Common Subsequence
  - The Number of Common Subsequences (count how many distinct subsequences both sequences have at least once)
  - The Count of Common Subsequences (for each shared subsequence, the sum of the product on the number in sequence 1 and the number in sequence 2)

Longest Common Subsequence

- Unlike the Longest Common Prefix (the substring starting at position 1), the LCS can span the whole of the two sequences
- The measure of similarity is the length of the longest subsequence present in both
- Elzinga (2009) shows that under certain cost configurations, OM is equivalent to LCS

Enumerating subsequences

- Indentifying the longest common subsequence is intuitively clear, the other measures involve enumerating sequences – combinatorics
- For a sequence of length l there are 2^l combinations of its elements, from length 0 (the "null" sequence) to l (the sequence itself)

- Thus `ABC` has as subsequences
  - . (the null sequence)
  - `A`, `B` and `C`
  - `AB`, `AC` and `BC`, and
  - `ABC`

- Sequences with repetitions, like `AAC`, have repetitions in the subsequences:
  - . (the null sequence)
  - `A`, `A` and `C`
  - `AA`, `AC` and `AC`, and
  - `AAC`

Number/Count of common subsequences

- The "Number of Common Subsequences" measure counts the number of distinct sequences which are subsequences of both sequences
- Elzinga (2009) points out that while it correlates highly with the LCS measure, they are distinctly non-monotone for sequences with moderate similarity
- NCS is not affected by how often matching subsequences occur, and if elements are repeated, subsequences can occur many times
- The "Number of Matching Subsequences" measure takes account of repetition: $\phi(x, y)$ is the sum, for each matching subsequence of $n_{sx} \times n_{sy}$ (i.e., the count of the subsequence in each sequence)

A distance from similarity

- As mentioned before, Elzinga (2009) asserts that where $\phi(x, y)$ measures the amount of a characteristic shared by x and y, and the following holds:

  $$\phi(x, y) = \phi(y, x)$$

  $$0 \leq \phi(x, y) \leq \min\{\phi(x, x), \phi(y, y)\}$$

  then

  $$d(x, y) = \phi(x, x) + \phi(y, y) - 2\phi(x, y)$$

  is metric

Algorithms

- The problem with subsequences is that there are 2^l of them – enumerating them is $O(2^N)$
- Elzinga 2005 outlines an algorithm to enumerate common subsequences in a pair of sequences that is $O(l_1 \times l_2)$
- This is implemented in CHESA, downloadable from `http://home.fsw.vu.nl/ch.elzinga/`
- I have implemented a "brute-force" algorithm for Stata, which enumerates subsequences in a first pass, and then compares them in a rapid second pass – good for up to about 20 tokens in its current version

Comparison with OM

- The algorithm is the main difference from OM, but
- The absence of substitution costs is another important difference – states are either the same (1) or different (0)
- Elzinga (personal communication) has outlined a measure which takes account of partial as well as complete similarity
- But what does the NMS measure look like in practice?

Example with birth/labour market sequences

- First problem: 73 periods is far too many (for my algorithm at least)
- Solution: sample every 4th month to yield 19 tokens per 5 year sequence
- Number of matching sequence measure calculated and distance as

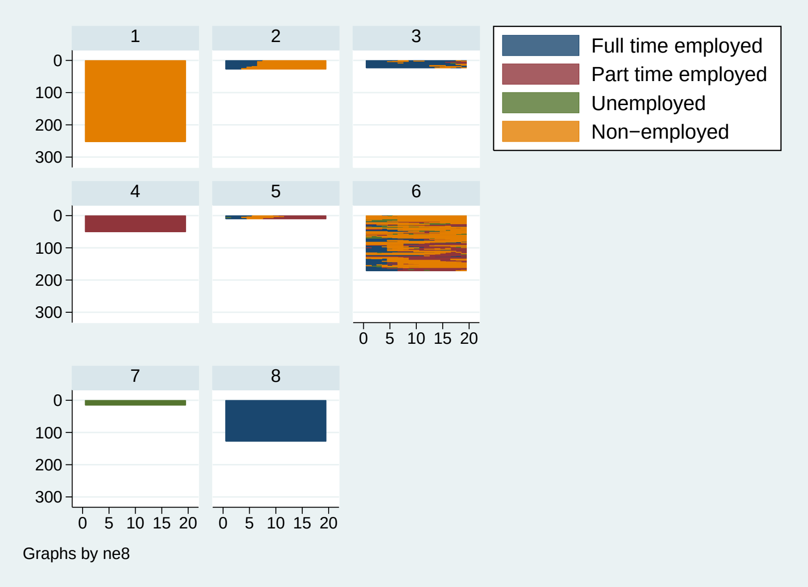
  $$d(x, y) = \phi(x, x) + \phi(y, y) - 2\phi(x, y)$$

Cluster analysis

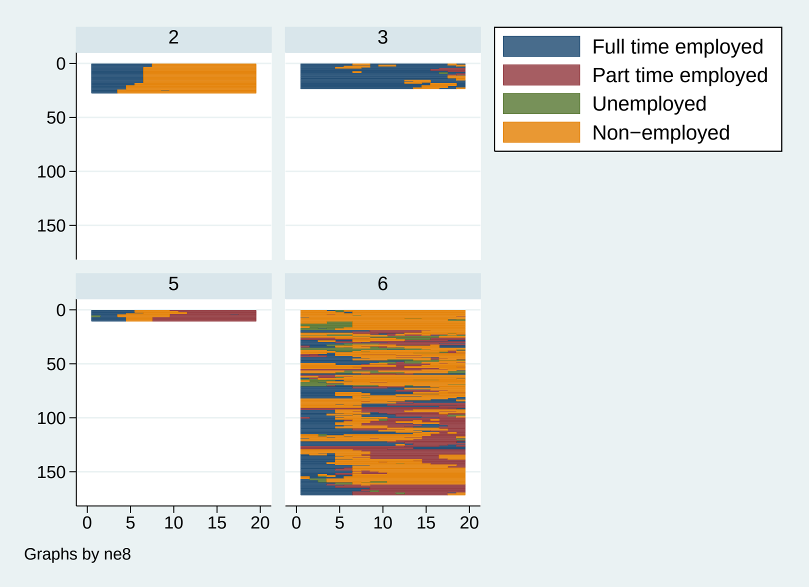
- Cluster analysis on the pairwise distances yields the following comparison with OM on the same sequences (8-cluster solutions)

```
OM                           NMS
         1        2        3        4        5        6        7        8
   +-----------------------------------------------------------------------+
 1 |   252        0        0        0        0        9        0        0 |
 2 |     0       27        0        0        0       51        0        0 |
 3 |     0        0       19        0        0       23        0        0 |
 4 |     0        0        0       50        0       19        0        0 |
 5 |     0        0        3        0       10       38        0        0 |
 6 |     0        0        0        0        0       27        0        0 |
 7 |     0        0        0        0        0        2       15        0 |
 8 |     0        0        1        0        0        2        0      127 |
   +-----------------------------------------------------------------------+
```

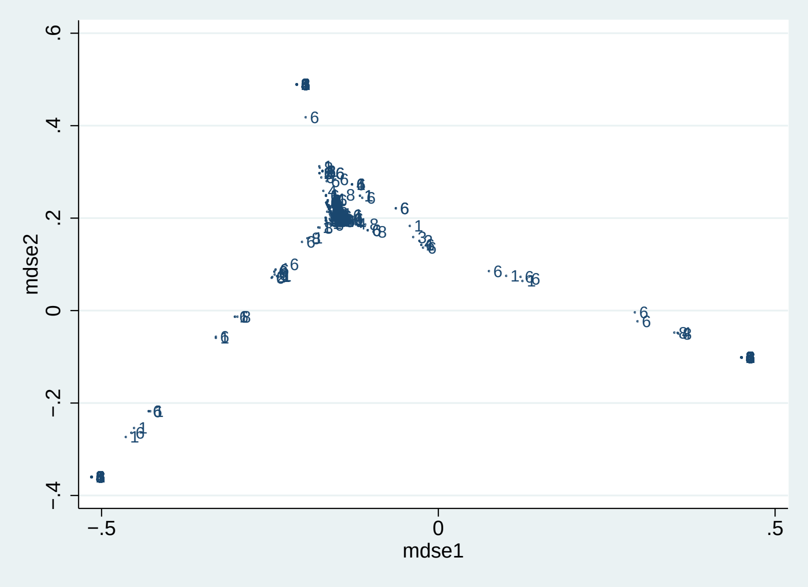
Eight-cluster solution



The “interesting” clusters



MDS solution: first two dimensions



Is OM optimal for life course data?

- Lifecourse data is usually spell structured – a sequence of periods in a single state, with a given duration
- How to deal with in OM, which works with sequences of tokens?
- Treat spells as tokens, ignore duration?
- Represent time by multiplying tokens by spell duration?

Spells as sequences of tokens

- The latter approach is usual, but it this sociologically optimal?
- For instance, OM says `AAAB` is as distant from `AACB` as from `AABB` (given $\delta(A, B) = \delta(A, C)$)
- Substantively, the first and third are very similar while the second introduces a completely new spell
- Do we need an algorithm that is aware of spells?

Ignore duration?

- A simple but extreme strategy is to ignore duration
- Makes sequences of spells, ignoring length
- It works – the sequences are still sequences, but now are unequal in length
- But duration is important: `FFnnnnnnnn` and `FFFnnnnnnn` are substantively closer to each other than to `FFFFFFFFnn` but all reduce identically to `Fn` as spell sequences

Cost deletion differentially?

- Another approach is to represent spells non-linearly
- Represent spell duration as e.g., $\sqrt{l}$
- Thus the “cost” of deleting a unit is bigger in a short spell (the fixed cost of deleting a unit represents more time in a long spell)
- One way of implementing this is OMv, a relatively simple adaptation of the OM algorithm (Halpin, 2010)

OMv: duration adjusted OM

- This approach means that the cost of edits to the token strings are sensitive to the length of the spell the token is in
- It produces distances not very different from OM unless there is a very high level of variation in spell length
- However, as currently implemented it is not a stable solution
  - Sequences composed of few, long spells are judged closer to all other sequences
  - Making for non-metric distances
- Potential solutions in scaling distances according to the most-dissimilar possible sequence

sociology
UNIVERSITY OF LIMERICK

Hollister's "Localized OM"

- Matissa Hollister proposes a similar approach (Hollister, 2009)
- She costs *indel*s differentially according to the element's neighbours
- To insert z between i and j:

$$C_{izj} = x \cdot w_{max} + y \cdot \frac{w_{zi} + w_{zj}}{2}$$

  where w is substitution cost
- This likely has similar consequences to OMv

sociology
UNIVERSITY OF LIMERICK

Dynamic Hamming distance

- (Lesnard, 2010) has proposed a "dynamic Hamming" distance, suitable for where time has a clear scale (daily, weekly, annual)
- Distances are inversely related to the moment-by-moment transition rate
- Thus states are closer at moments when many changes happen
- Further apart in times of low change
- Successfully used with daily time-use data

sociology
UNIVERSITY OF LIMERICK

Warping time

- OMv "warps time" by weighting it differently in different spells
- Harks back to Abbott's use of the term to suggest non-linear time scales (Abbott and Hrycak, 1990)
- In turn informed by Sankoff and Kruskal (1983), *Time Warps, String Edits and Macromolecules*

sociology
UNIVERSITY OF LIMERICK

Time warping algorithms

- Formally, time warping is a family of algorithms that do "continuous time-series to time-series correction" while OM *et al* do "string to string correction" (Marteau, 2007)
- Focus on comparing pairs of continuous-time high-dimensional time-series in $\mathbb{R}^n$
- Operates by locally compressing or expanding the time scale of one trajectory to minimise the distance to the other
- Distance is usually Euclidean in $\mathbb{R}^n$ or other simple distance

Multiple domains

- Handling trajectories through multiple domains simultaneously is very attractive
- Quite a few examples in the literature, often combining
  - Labour market
  - Housing
  - Partnership and family formation
- Dijkstra and Taris (1995) use as an example residential, educational and job status
- Pollock (2007) uses a similar trio

Multiple state spaces

- The immediate difficulty is how to deal with multiple state spaces
- One solution is to create a combined space, crosstabulating the others, e.g.,:
  - employed–single
  - employed–partnered
  - not employed–single
  - not employed–partnered
- However, in practice this usually generates a high number of cells
- Practical problem of determining substitutions costs

Substitution costs in multiple spaces

- It may be possible to set the costs on the cross-tabulated spaces *a priori*, using intuition or theory
- Sometimes it may also be possible to simplify the structure of this space: e.g., collapse certain regions into a single category
- If clear state-space structures exist for the sub-spaces it may be possible to combine them systematically:
  - Euclidean: $\sqrt{\sum_i (x_{ij} - x_{ik})^2}$
  - Sum the different distances: $\sum_i |x_{ij} - x_{ik}|$
- Weighting subdomains differentially is also possible

Parallel analyses

- An other option is to conduct parallel analyses in each domain
- This yields multiple distance measures
- Allows analysis of how the different domains cluster independently
- Perhaps less sensitive to coordination issues within lifecourses but should still be interesting

Multi-channel SA

- A team in Switzerland are drawing on newer bioformatics technology: Multi-channel SA
- Explicitly deals with multiple parallel trajectories
- Gauthier et al. (2008) describe their “MCSA” method and claim it is superior to parallel OM analyses, and simpler than handling multiple-state distances
- Software available: `http://www.tcoffee.org/saltt`
- Not entirely clear from the description how the method works
- Bühlmann (2008) uses their methodology to examine careers of Swiss economists and engineers, using a number of categorical measures of their status

Turbulence

- Sequence complexity is extremely important, and how to classify complex sequences is a real problem
- Simpler sequences cluster well, but not complex ones
- How to measure this complexity?
- Elzinga (2008) proposes a spell-based, combinatorial definition, which he refers to as *turbulence* (see also Elzinga and Liefbroer, 2007)

Turbulence defined

- Turbulence is higher the more transitions there are
- And the more different states are entered,
- but lower when the variance of spell durations is higher
- Elzinga (2008) offers an efficient algorithm
- Implemented in TraMineR

Stata and OM

- Kohler et al provide the SQ add-on for Stata:

```
. ssc install sq
```

- User-friendly, facilitates treatment of sequences, good graphs
- Fits OM but slow on large data sets
- My software is faster but less user-friendly: `http://teaching.sociology.ul.ie/seqanal`

TraMineR

- TraMiner (`http://mephisto.unige.ch/traminer`) is an R package for sequence analysis
- R is a free/open-source implementation of the S-Plus language (`http://www.r-project.org`)
- Available for most platforms (including Windows and Unix)
- Fast, powerful, good graphics, but R is not beginner-friendly

Abbott, A. (1995a). A comment on “Measuring the agreement between sequences”. *Sociological Methods and Research*, 24(2):232–243.

Abbott, A. (1995b). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21:93–113.

Abbott, A. and Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, XVI(3):471–494.

Abbott, A. and Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians’ careers. *American Journal of Sociology*, 96(1):144–85.

Abbott, A. and Tsay, A. (2000). Sequence analysis and optional matching methods in sociology. *Sociological Methods and Research*, 29(1):3–33.

Aisenbrey, S. and Fasang, A. E. (2010). New life for old ideas: The ”second wave” of sequence analysis bringing the

”course” back into the life course. *Sociological Methods and Research*, 38(3):420–462.

Bühlmann, F. (2008). The corrosion of career? – occupational trajectories of business economists and engineers in switzerland. *European Sociological Review*, 24(5):601–616.

Degenne, A., Lebeaux, M.-O., and Mounier, L. (1996). Typologies d’itinéraires comme instrument d’analyse du marché du travail. Troisièmes journées d’études Céreq-Cérétim-Lasmas IdL, Rennes, 23–24 May 1996.

Dijkstra, W. and Taris, T. (1995). Measuring the agreement between sequences. *Sociological Methods and Research*, 24(2):214–231.

Elzinga, C. H. (2003). Sequence similarity: A non-aligning technique. *Sociological Methods and Research*, 32(1):3–29.

Elzinga, C. H. (2005). Combinatorial representations of token sequences. *Journal of Classification*, 22(1):87–118.

Elzinga, C. H. (2008). Complexity of categorical time series. under review.

Elzinga, C. H. (2009). Sequence analysis: Metric representations of categorical time series. *Sociological Methods and Research*. (forthcoming).

Elzinga, C. H. and Liefbroer, A. C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population*, 23:225–250.

Elzinga, C. H., Rahmann, S., and Wang, H. (2008). Algorithms for subsequence combinatorics. *Theoretical Computer Science*, 409(3):394–404.

Gauthier, J.-A., Widmer, E., Bucher, P., and Notredame, C. (2008). Multi-channel sequence analysis applied to social science data. available at http://papers.ssrn.com/.

Halpin, B. (2010). Optimal matching analysis and life course

data: The importance of duration. *Sociological Methods and Research*, 38(3):365–388.

Han, S.-K. and Moen, P. (1999). Work and family over time: A life course approach. *Annals of the American Academy of Political and Social Science*, 562:98–110.

Hollister, M. (2009). Is optimal matching suboptimal? *Sociological Methods and Research*, 38(2):235–264.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods and Research*, 38(3):389–419.

Levine, J. H. (2000). But what have you done for us lately? Commentary on Abbott and Tsay. *Sociological Methods and Research*, 29(1):34–40.

Marteau, P.-F. (2007). Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching. *ArXiv Computer Science e-prints*.

Pollock, G. (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A*, 170(1):167–183.

Robette, N. and Thibauld, N. (2008). Comparing Qualitative Harmonic Analysis and Optimal Matching. *Population*, 63(4):533–556.

Sankoff, D. and Kruskal, J. B., editors (1983). *Time Warps, String Edits and Macromolecules*. Addison-Wesley, Reading, MA.

Wu, L. L. (2000). Some comments on “Sequence analysis and optimal matching methods in sociology: Review and prospect”. *Sociological Methods and Research*, 29(1):41–64.