Understanding Substitution Costs: parameterising the Optimal Matching Algorithm

Brendan Halpin
Department of Sociology
University of Limerick
brendan.halpin@ul.ie[1]

2 Sept 2008, RC33 Naples

[1] Work in progress! See http://teaching.sociology.ul.ie/seqanal/d2D.pdf shortly

The *problem* of substitution costs

- As sequence analysis (SA) becomes more common in sociology, increasing interest in its sociological meaningfulness
- Does the Optimal Matching Algorithm (OMA) make sense for sociological data?
  - Is the algorithm suitable? (see elsewere)
  - How to parameterise it: substitution and *indel* costs

sociology
UNIVERSITY OF LIMERICK

A problem?

- 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
- This paper explores substitution costs and attempts to clarify the issue

Mapping states to sequences

- The essence of SA is mapping a view of a state space onto a view of a trajectory space: $d(s) \rightarrow D(S)$
- 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

OMA coherent?

- Can we expect OMA to provide a coherent $d(s) \rightarrow D(S)$ mapping?
- Elementary operations are intuitively appealing:
  1. $D(\texttt{ABC}, \texttt{ADC}) = f(d(\texttt{B}, \texttt{D}))$
  2. $D(\texttt{ABCD}, \texttt{ABD}) = f(indel)$
  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

sociology
UNIVERSITY OF LIMERICK

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

sociology UNIVERSITY OF LIMERICK

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)
- Even proponents of OMA are sometimes concerned about “impossible” transitions (e.g., Pollock)
- But substitutions are not transitions, **not even a little bit!**
  - substitutions happen across sequences, $D(\texttt{ABC}, \texttt{ADC}) = f(d(\texttt{B}, \texttt{D}))$ (similarity of states)
  - transitions happen within sequences (movement between state)

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 usually 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)
  - $d(\text{Con}, \text{Lab}) > d(\text{Con}, \text{LibDem}) + d(\text{LibDem}, \text{Lab})$
- Procedure confuses party state space and voter characteristics
- Voter polarisation/loyalty is trajectory information, not state information
- Another type of problem: irrelevant distinctions can cause similar states to have low transition rates

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

sociology
UNIVERSITY OF LIMERICK

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

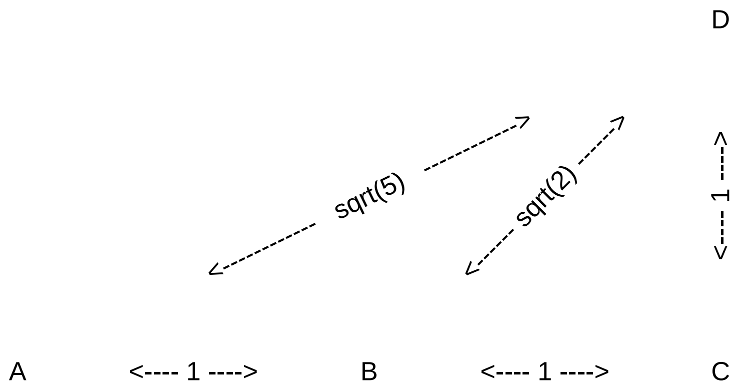
0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	0

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?

sociology
UNIVERSITY OF LIMERICK

2-D example



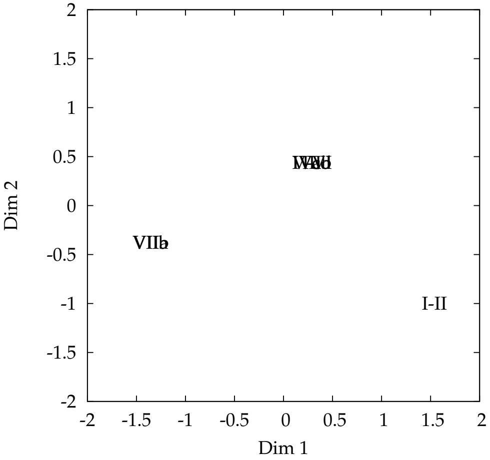
Spatial structure of theoretical spaces

- We can analyse “theoretically-informed” or *ad hoc* state spaces spatially
- Principle components analysis of substitution matrix
- Examples: Halpin and Chan, 1998; McVicar/Anyadike-Danes 2002:

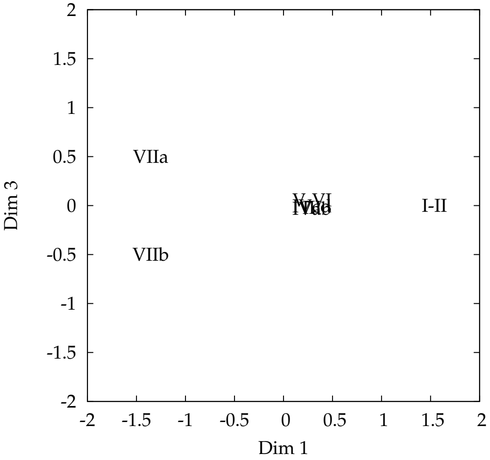
I–II	0	2	2	2	2	3	3
III	2	0	1	1	1	2	2
IV*ab*	2	1	0	1	1	2	2
IV*cd*	2	1	1	0	1	2	2
V–VI	2	1	1	1	0	2	2
VII*a*	3	2	2	2	2	0	1
VII*b*	3	2	2	2	2	1	0

E	0	1	1	2	1	3
F	1	0	1	2	1	3
H	1	1	0	2	1	2
S	2	2	2	0	1	1
T	1	1	1	1	0	2
U	3	3	2	1	2	0

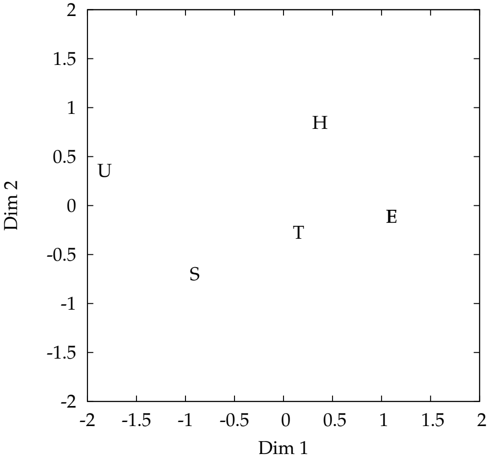
H&C, 1st two PCA dimensions



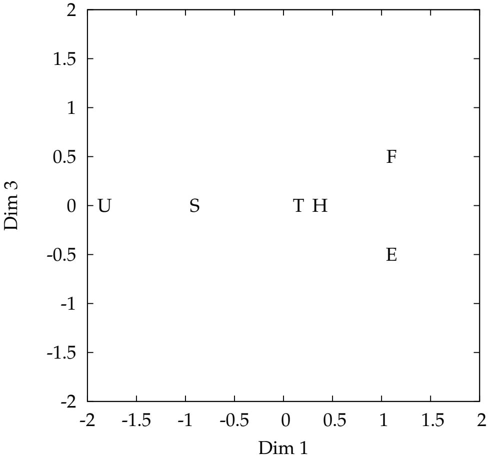
H&C, dimensions 1 & 3



MVAD, 1st two dimensions



MVAD, dimensions 1 & 3



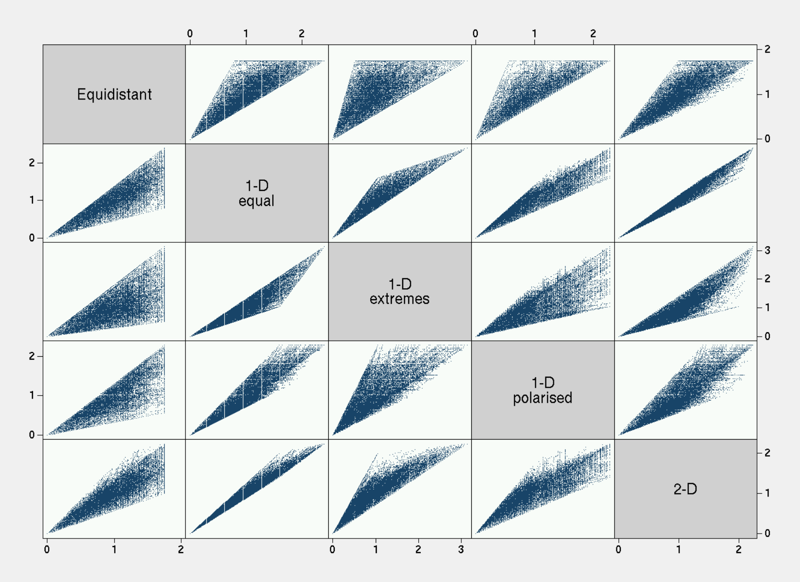
sociology
UNIVERSITY OF LIMERICK

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 $d(\texttt{"A"},\texttt{"B"}) << d(\texttt{"A"},\texttt{"C"})$ then $D(\texttt{"..AAAA.."},\texttt{"..BBB.."})$ will tend to be less than $D(\texttt{"..AAAA.."},\texttt{"..CCC.."})$
  - Differential distances promote alignment: `AADDAAA` and `AAADDAA` are more likely to be aligned to match the `DD` if $d(\texttt{"A"},\texttt{"D"})$ 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

sociology
UNIVERSITY OF LIMERICK

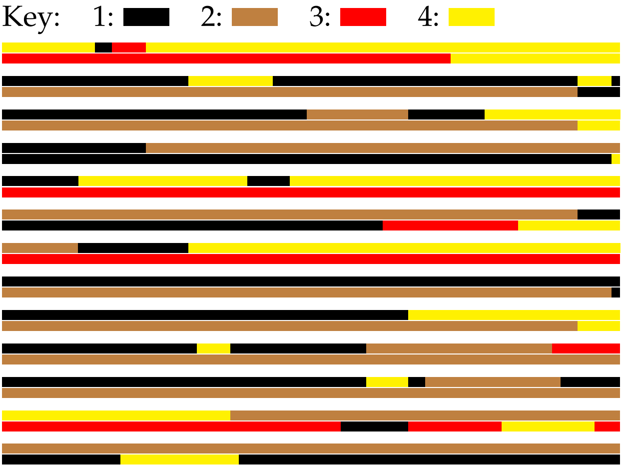
Comparing effects



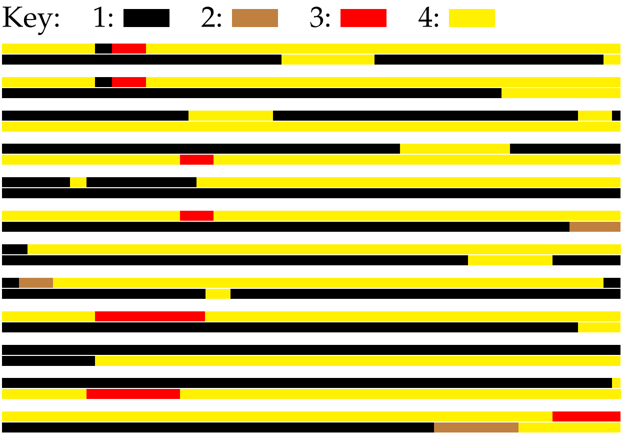
Correlations

Equidistant	1.00				
1-D equal	0.85	1.00			
1-D extremes	0.66	0.93	1.00		
1-D polarised	0.83	0.94	0.81	1.00	
2-D	0.87	0.98	0.91	0.90	1.00

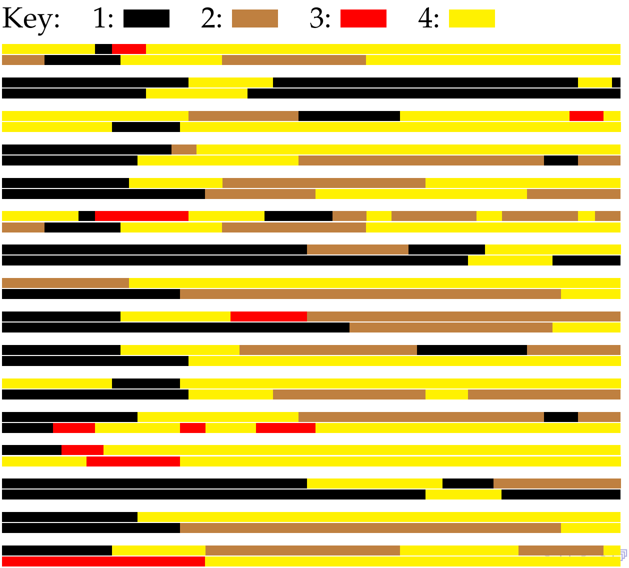
Equidistant relatively greater than 1-D



Equidistant relatively less than 1-D



Equidistant close to 1-D



sociology
UNIVERSITY OF LIMERICK

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

sociology
UNIVERSITY OF LIMERICK

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

sociology
UNIVERSITY OF LIMERICK