Setting costs for multidomain sequence analysis

Gilbert Ritschard

Institute of Demography and Socioeconomics, University of Geneva
http://mephisto.unige.ch/traminer

SAA symposium at SLLS 2021
September 22, 2021

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

Outline

Multidomain sequences

Multidomain sequences (MDseq)

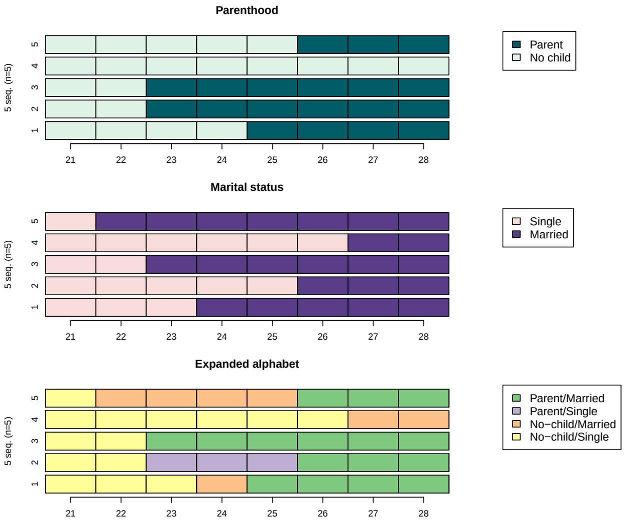
When elements in sequences reflect combination of states from different domains

- Family life sequences may combine marital status with parenthood
- Professional careers may combine education, domain of activity, rates of activity, roles.
- Life sequences may combine family, work, health, ...
- Linked lives: status of one person and status of partner

Interest of considering multidomain sequences:

- Studying relationships between domains

Toy example of multidomain sequences



Distance between MDseq

Strategies for distances between MDseq

- Extended alphabet : combine alphabets, e.g. Child * Single, Child * Married
    - Combining dimensions may generate big alphabets
    - Consequence for edit distances: dramatic increase of number of substitution (and possibly indel) costs.
    - Trick for edit distances : derive multichannel costs from costs of each channel (Pollock, 2007; Gauthier et al., 2010).
- Sum (linear combination) of distances computed on the different channels.
- Other approaches: e.g. GIMSA (Robette et al., 2015)

Costs for combined states

How to set costs for combined states?

- Proceed as for regular sequences
  - unique, theoretically-based, data-driven, ...
- Additive trick (AT): MD costs defined as sum (average) of costs of individual domains
  - Example:
    $sc(child + single, nochild + married)$
    $= sc(child, nochild) + sc(single, married)$

Additive trick and state dependent indels

- For single state independent indel per channel, additive indel:

$$\text{indel}_{MD} = \text{indel}_{parenthood} + \text{indel}_{marital}$$

- OM and related methods work as well with a vector of state dependent indels:
- Straightforward extension: additive state dependent indels

$$\text{indel}(child, married) = \text{indel}(child) + \text{indel}(married)$$

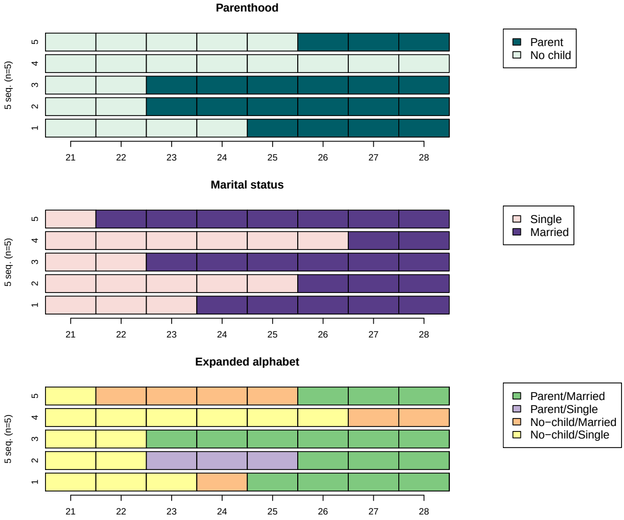
Pro and cons of AT costs

- AT reduces number of independent costs to set.
  Computationally, this is not an issue.
- Interpretation: AT-based MD costs get an nice interpretation:
  - When SC is same (constant) in all channel, the AT cost is proportional to the number of channels on which the two MD states differ.
- However, deriving MD costs by summing (averaging) costs of each channel has a severe flaw.
  - Supposes independence of domains, while MD analysis is of interest to study the dependence between domains.

Costs and interaction between domains

- A priori, no reasons to have
  - $sc(nochild + single, child + married)$
    $= sc(child + single, nochild + married)$
  - which automatically follows when applying the additive trick
- In particular, likelihood of having child may depend on marital status such that
  - $nochild + single$ more common than $child + single$
  - $child + married$ more common than $nochild + married$
- and, because of this interaction, we would expect:
  - $sc(nochild + single, child + married)$
    $< sc(child + single, nochild + married)$

Toy example of 5 MD sequences



INDELSLOG, extended alphabet vs additive trick

True multichannel INDELSLOG costs

	c.m	c.s	n.m	n.s
c.m	0.00	0.98	0.89	0.75
c.s	0.98	0.00	1.15	1.01
n.m	0.89	1.15	0.00	0.92
n.s	0.75	1.01	0.92	0.00
Indel	0.36	0.62	0.53	0.39

AT costs based on additive trick (symmetry around both diagonals)

	c.m	c.s	n.m	n.s
c.m	0.00	0.58	0.58	1.15
c.s	0.58	0.00	1.15	0.58
n.m	0.58	1.15	0.00	0.58
n.s	1.15	0.58	0.58	0.00
Indel	0.54	0.64	0.51	0.61

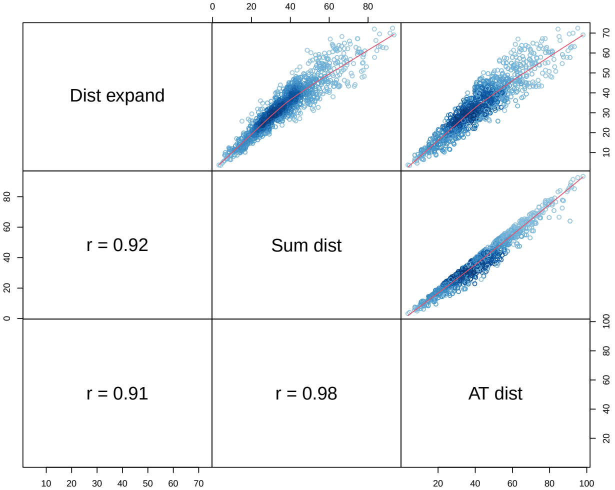
Illustration: Cohabitation and working status, Switzerland

- Biographic data from Swiss Household Panel (SHP)
- 2 domains: Living arrangement (8 states + NA), Working status (8 states + NA)
- Expanded alphabet: 71 states (2485 sc costs, 71 indel costs)
- 1990 life sequences of length between 41 and 60 years ($\frac{n(n-1)}{2} = 1,979,055$ dissimilarities)

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

Plot of distances (random sample of 1000 pairs)

```
[1] 1979055       3
```



UNIVERSITÉ DE GENÈVE

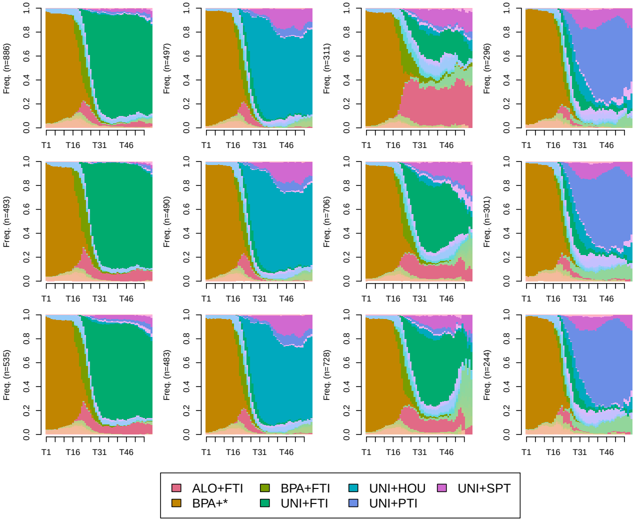
Clusters (chronograms of combined states)

MD dist

Sum dist

AT dist

main states



Conclusion

- Multidomain sequence analysis (MDSA) primarily concerns the study of relationships between domains
- MDSA of interest for linked domains only
- Additive trick (AT) for setting MD costs assumes independence between domains. Therefore, not recommended.
- Sum of distances computed independently on each channel assumes independence of domains too. Also not recommended.

Thank you!

References I

Gauthier, J.-A., E. D. Widmer, P. Bucher, and C. Notredame (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology 40*(1), 1–38.

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 A 170*(1), 167–183.

Robette, N., X. Bry, and E. Lelièvre (2015). A 'global interdependence' approach to multidimensional sequence analysis. *Sociological Methodology 45*(1), 1–44.

Toy example of 5 MD sequences

	21	22	23	24	25	26	27	28
1	n	n	n	n	c	c	c	c
2	n	n	c	c	c	c	c	c
3	n	n	c	c	c	c	c	c
4	n	n	n	n	n	n	n	n
5	n	n	n	n	n	c	c	c

n no child
c child

	21	22	23	24	25	26	27	28
1	s	s	s	m	m	m	m	m
2	s	s	s	s	s	m	m	m
3	s	s	m	m	m	m	m	m
4	s	s	s	s	s	s	m	m
5	s	m	m	m	m	m	m	m

s single
m married

INDELSLOG costs

We illustrate with INDELSLOG data-driven costs

INDELSLOG method

- Indel and substitution costs are state dependent
- $\text{indel}(x_i) = \log(2/(1 + f_i))$
  where f_i is relative frequency of state x_i in data set.
- $sc(x_i, x_j) = \text{indel}(x_i) + \text{indel}(x_j)$

$\text{indel}(x_i)$ decreases with frequency f_i of state x_i

INDELSLOG costs: main differences

We observe:

- Using true MD INDELSLOG costs
  - Low cost (.75) for substituting $c.m$ with $n.s$
  - High cost (1.15) for substituting $c.s$ with $n.m$
  - Cost (1.01) for substituting $c.m$ with $n.s$ (diff. on 2 domains) is lower than for substituting $c.s$ with $n.s$ (diff. on 1 domain)
  - State dependent indel costs: less frequent combined states ($c.s$ and $n.m$) get higher indel costs.
- Using AT, i.e. summing single channel INDELSLOG costs
  - Same cost (1.15) for substituting $c.m$ with $n.s$ and for substituting $c.s$ with $n.m$
  - Cost for substituting $c.m$ with $n.s$ is approximatively twice the cost for substituting $c.s$ with $n.s$ (diff. on 1 domain only)
  - Similar indel costs for unfrequent and frequent combined states.

Using TRATE costs

TRATE method

- Substitution costs are state dependent
- $sc(i, j) = 2 - p(x_{i,t}|x_{j,t-1}) - p(x_{j,t}|x_{i,t-1})$

Using TRATE costs, we observe differences between true MD costs (computed on expanded alphabet) and additive MD TRATE costs similar to those observed using INDELSLOG.

TRATE, extended alphabet vs additive trick

True multichannel TRATE costs

	c.m	c.s	n.m	n.s
c.m	0.00	1.67	1.67	1.93
c.s	1.67	0.00	2.00	1.93
n.m	1.67	2.00	0.00	1.79
n.s	1.93	1.93	1.79	0.00
Indel	1.00	1.00	1.00	1.00

AT costs based on additive trick (observe double symmetry)

	c.m	c.s	n.m	n.s
c.m	0.00	1.71	1.80	3.51
c.s	1.71	0.00	3.51	1.80
n.m	1.80	3.51	0.00	1.71
n.s	3.51	1.80	1.71	0.00
Indel	1.75	1.75	1.75	1.75