What matters in differences between life trajectories? A comparative review of sequence dissimilarity measures

Matthias Studer[1,2] & Gilbert Ritschard[1,2]

[1]LIVES NCCR

[2]Institute of Demography and Socioeconomics
University Geneva

FORS – SSP Methods and Research meetings, University of Lausanne, December 1, 2015

Outline

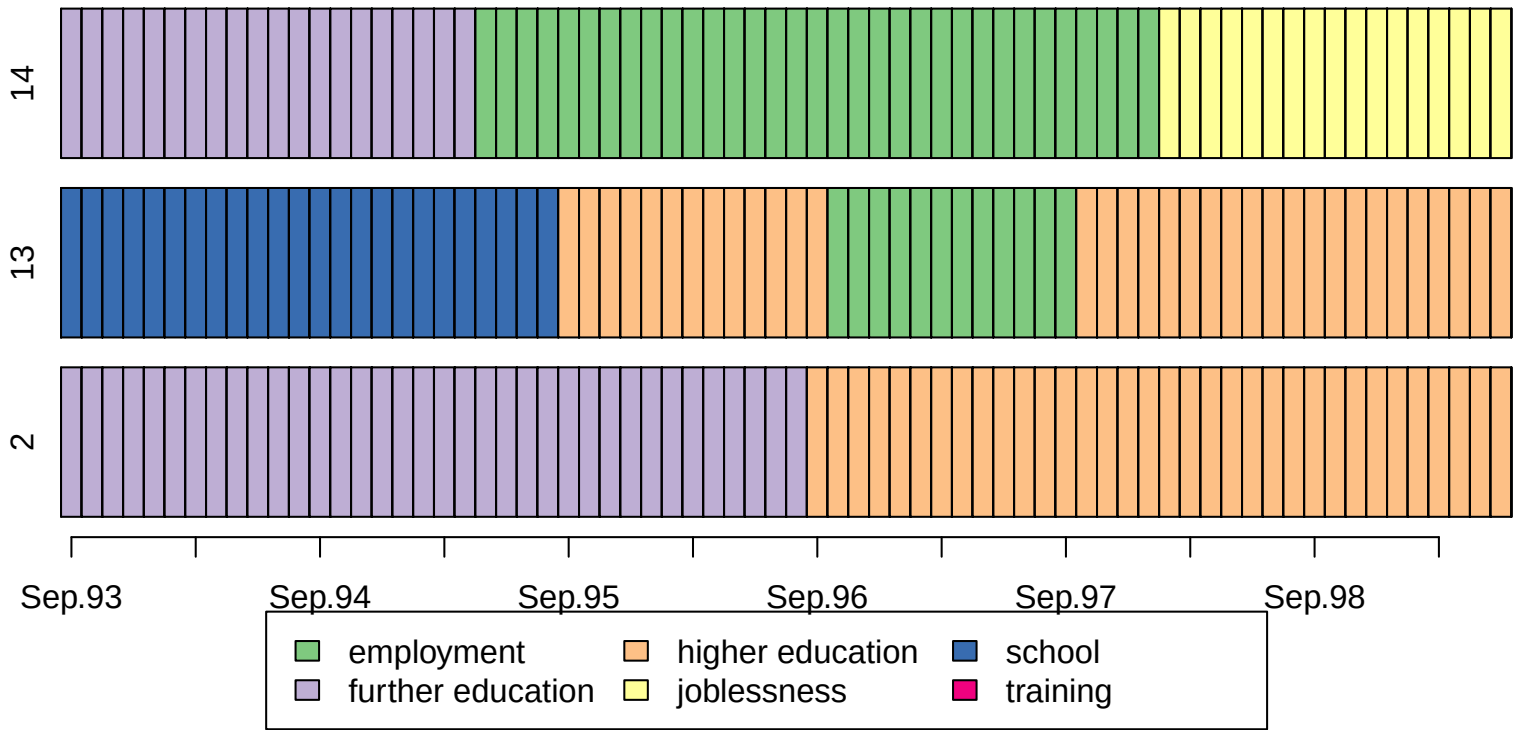
Outline

Sequence Analysis in the Social Sciences

- SA aims to describe trajectories.
  - Professional carriers.
  - Cohabitational life courses.
  - History of organizations.
- Typology of the trajectories.
- Common questions in sequence analysis.
  - What are the typical patterns of trajectories?
  - How are the trajectories related to explanatory factors?
  - How is a given outcome related to a previous trajectory?

Sequences analysis: common strategy

- Code processes/trajectories as state sequences.



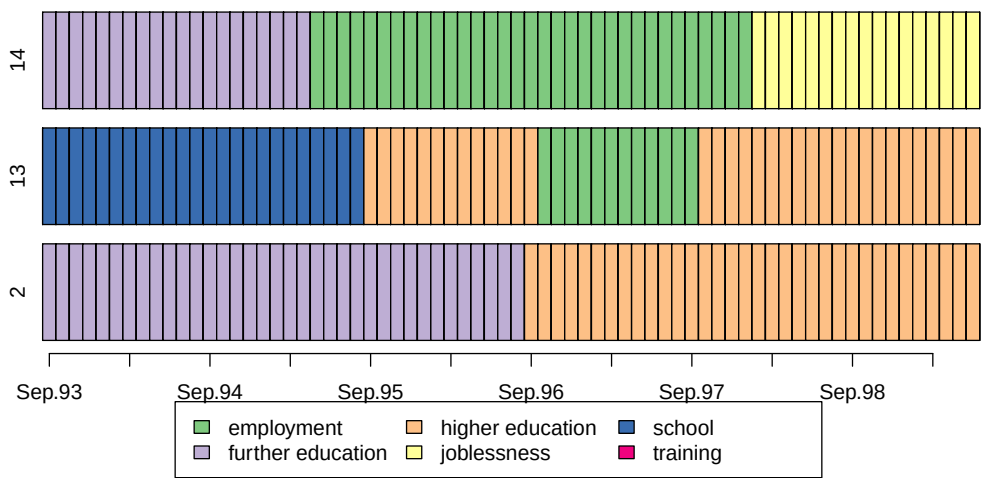
- Compute distances between sequences, i.e. Optimal matching.

Typology of processes

- Reveals main patterns.

Optimal Matching

- "Optimal Matching": distance measure between sequences.
  - Definition: number of operation needed to transform one sequence into another one.
    - Substitution.
    - Insertion–deletion.
  - Operation cost can be weighted.



Criticism

- Many critics (Levine, 2000; Wu, 2000; Elzinga, 2003).
- Lack a sociological interpretation.
- High number of parameters.
- Parameters values set by the user.
- Timing and sequencing of sequences are not sufficiently taken into account.

New developments

- New developments as answers to criticisms (Aisenbrey and Fasang, 2010):
  - New distances measures.
  - New methods to automatically compute parameters values.
- Result in many distances measures.
  - Seven article in *Sociological Method and Research*.
  - Each having at least one parameter.
- Scattered development.
  - Answer to one critic at a time.
  - Only compare to classic OM.

New developments

- New developments as answers to criticisms (Aisenbrey and Fasang, 2010):
  - New distances measures.
  - New methods to automatically compute parameters values.
- Result in many distances measures.
  - Seven article in *Sociological Method and Research*.
  - Each having at least one parameter.
- Scattered development.
  - Answer to one critic at a time.
  - Only compare to classic OM.

Choosing a distance

- SA users common questions:
  - How to choose distance measure?
  - How to set the parameters?
- Aim: Help SA users to choose a distance and set the parameters.
  - Review all distances measures.
  - Provide guidelines.

Choosing a distance

- SA users common questions:
  - How to choose distance measure?
  - How to set the parameters?
- Aim: Help SA users to choose a distance and set the parameters.
  - Review all distances measures.
  - Provide guidelines.

Outline

Review of distance measures properties

	Type		Description	Properties					Parameters		
Measure	Dis	AttEdt		Metric	Eucl	T.warp	S.dep	Ctxt	Subst.	Indels	Others
CHI2, EUCLID	x		Distance between per period state distributions	x	x	x					Number of periods K
CHI2fut (Rousset)	x		Position-wise state distances based on shared future	x	x			x			Time-lag weighting function
NMS (Elzinga)		x	Based on number of matching subsequences	x	x	x		x			
SVRspell (Elzinga & Studer)		x	Based on number of matching spell subsequences with spell-length weights	x	x	x	x	x	User		Subsequence length weight a, spell duration weight b
HAM (Hamming)	x	x	Number of mismatches	x	x[b]						
generalized		x	Sum of mismatches with state-dependent weights	x[a]	x[b,c]		x		User		
DHD (Lesnard)		x	Sum of mismatches with position-wise state-dependent weights				x	x	Data		
OM		x	Minimum cost for turning x into y using theoretically defined costs	x[a]		x	x		User	Mult	
LCS / OM(1,2) / Levenshtein-II	x	x	Based on length of LCS / Number of indels	x		x					
feature		x	Costs based on state features	x		x	x		Features	Single	State features
future (new)		x	Costs based on similarity between conditional state distributions q periods ahead	x		x	x		Data	Single	Forward lag q
trate		x	Costs based on transition rates			x	x		Data	Single	Transition lag q
opt[na](Gauthier)		x	Costs adjusted to increase similarity between similar sequences	[n]		x	x		Data	Single	Similarity rate
indels, indelslog (new)		x	State dependent indels based on inverse or log inverse state frequencies.	x		x	x			Auto	
OMloc (Holister)		x	Context dependent indel costs			x	x	x	User	Auto	Expansion cost e, Context g
OMslen (Halpin)		x	Costs weighted by spell length	x		x	x	x	User	Mult[na]	Spell length weight h
OMspell (new)		x	OM between sequences of spells	x[a]		x	x	x	User	Mult[na]	Expansion cost e
OMstran (new)		x	OM between sequences of transitions	x[a]		x	x	x	User	Mult	Origin-transition trade-off w, Transition indel cost function

[a] If costs fulfil the triangle inequality. [b] Squared Euclidean distance. [c] If costs are squared Euclidean distances. [na] Not available in TraMineR. [n] Can generate negative dissimilarities.

Review

- Theoretical review.
- Many distance measures.
- Highlight mathematical distances properties.
- Many non-metric dissimilarities.
  - 5 out of 7 distance published in SMR do not satisfy triangle inequality.
  - 2 with serious issues (Wrong algorithm or negative distances).
- Overlooked mathematical properties?

Reviewing distances

- How to choose a distance measure?
- How to evaluate a distance measure?
- A distance measure defines how two sequences are compared.
- Which aspects should we use to compare trajectories?
  - Sociological issue.
  - Five aspects based on Settersten and Mayer (1997) and Billari et al. (2006).

Reviewing distances

- How to choose a distance measure?
- How to evaluate a distance measure?
- A distance measure defines how two sequences are compared.
- Which aspects should we use to compare trajectories?
  - Sociological issue.
  - Five aspects based on Settersten and Mayer (1997) and Billari et al. (2006).

Reviewing distances

- How to choose a distance measure?
- How to evaluate a distance measure?
- A distance measure defines how two sequences are compared.
- Which aspects should we use to compare trajectories?
  - Sociological issue.
  - Five aspects based on Settersten and Mayer (1997) and Billari et al. (2006).

Sequence comparison aspects

- Experienced states.
  - Similar sequence should have some states/events in common.
- Distribution.
  - Total exposure time.
- Timing.
  - Age in a state/time an event occurs.
- Spell duration.
  - Consecutive time spent.
- Sequencing.
  - Order of the states/events in the sequence.

Sequence comparison aspects

- Experienced states.
  - Similar sequence should have some states/events in common.
- Distribution.
  - Total exposure time.
- Timing.
  - Age in a state/time an event occurs.
- Spell duration.
  - Consecutive time spent.
- Sequencing.
  - Order of the states/events in the sequence.

Sequence comparison aspects

- Experienced states.
  - Similar sequence should have some states/events in common.
- Distribution.
  - Total exposure time.
- Timing.
  - Age in a state/time an event occurs.
- Spell duration.
  - Consecutive time spent.
- Sequencing.
  - Order of the states/events in the sequence.

Sequence comparison aspects

- Experienced states.
  - Similar sequence should have some states/events in common.
- Distribution.
  - Total exposure time.
- Timing.
  - Age in a state/time an event occurs.
- Spell duration.
  - Consecutive time spent.
- Sequencing.
  - Order of the states/events in the sequence.

Sequence comparison aspects

- Experienced states.
  - Similar sequence should have some states/events in common.
- Distribution.
  - Total exposure time.
- Timing.
  - Age in a state/time an event occurs.
- Spell duration.
  - Consecutive time spent.
- Sequencing.
  - Order of the states/events in the sequence.

Outline

Simulations

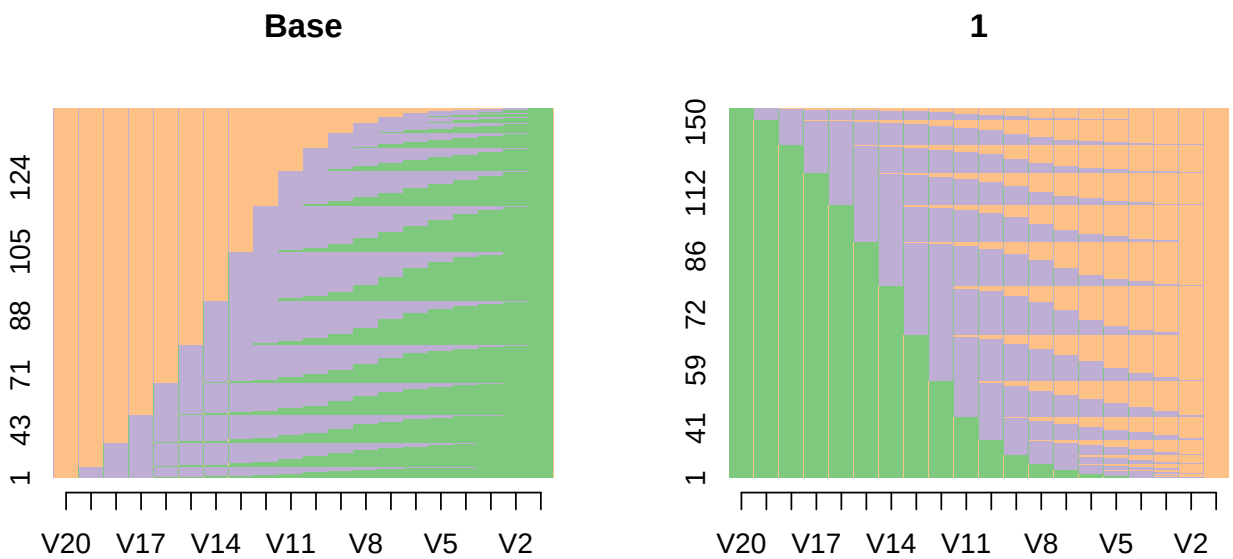
- The sensitivity to each aspect vary between distance measures.
- Use simulation to measure this sensitivity.
- Aim: describe the behaviour of each distance/configuration of parameter.
- Generate two groups of sequences.
  - Groups differ on one aspect.
  - Measure ability of each distance to discriminate between groups.
  - Based on discrepancy analysis (Pseudo-R^2).
- Randomize untested aspects: groups should *only* differ on one aspect.

Simulations

- The sensitivity to each aspect vary between distance measures.
- Use simulation to measure this sensitivity.
- Aim: describe the behaviour of each distance/configuration of parameter.
- Generate two groups of sequences.
  - Groups differ on one aspect.
  - Measure ability of each distance to discriminate between groups.
  - Based on discrepancy analysis (Pseudo-R^2).
- Randomize untested aspects: groups should *only* differ on one aspect.

Sequencing Simulation

- Generate two groups of sequences.
  - Group 1: $\mathbf{x} = (A, B, C)$
  - Group 2: $\mathbf{x} = (C, B, A)$
  - Durations and timings random in both groups.
- 2'000'000 sequences.



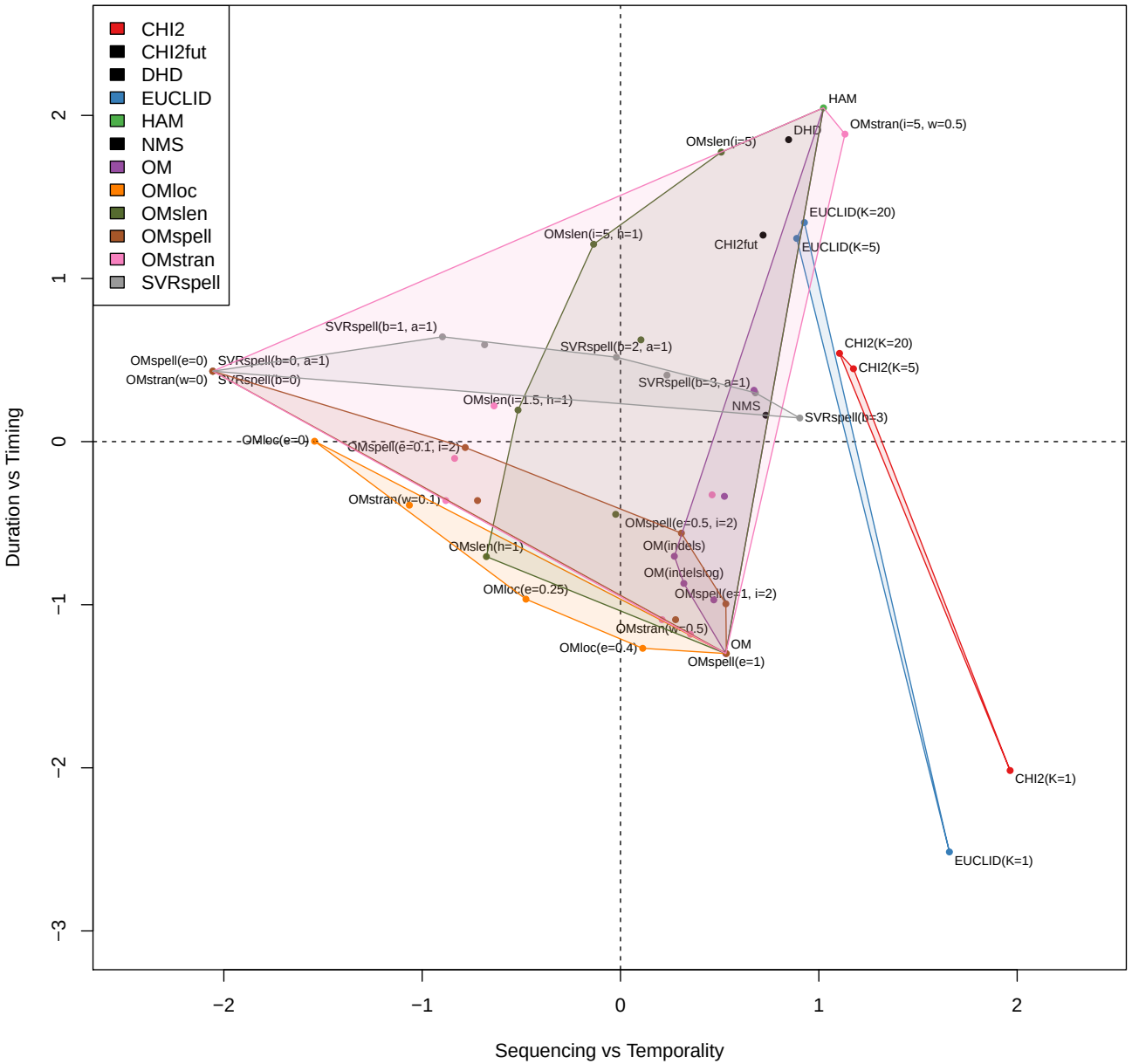
Sets of simulations

- State based:
  - Sequencing:
    - Difference of patterns.
    - Random small perturbations.
  - Timing: age at the beginning of a spell.
  - Duration: duration of a spell.
- Event based (based on three events e_1, e_2, e_3)
  - Sequencing: order of underlying events.
  - Timing: age at a given event.
  - Duration: "spacing" between events.
- Simulations chosen among those considered in Studer (2012).

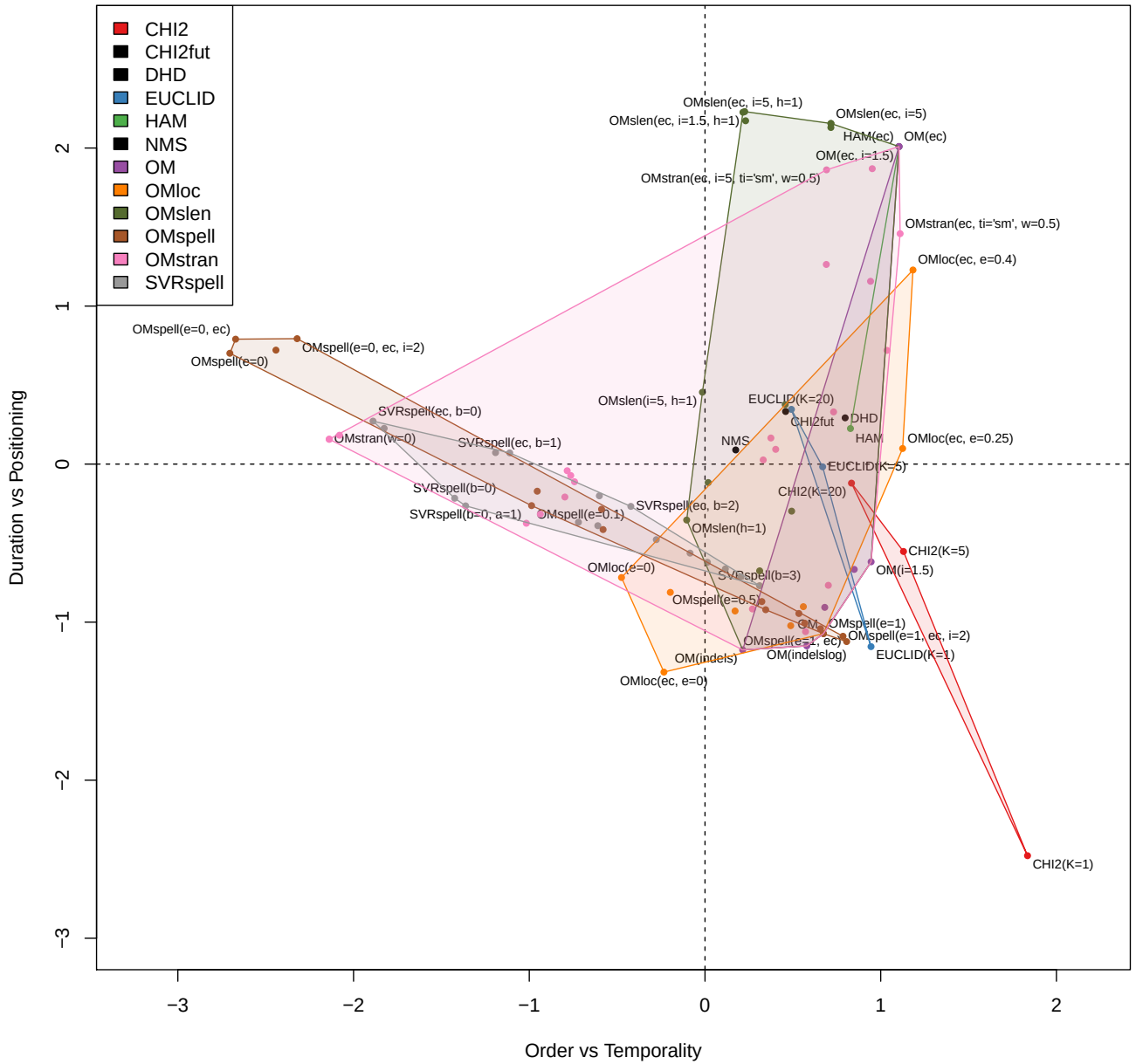
Distance included in analysis

Distance	Configurations
Distribution-based	EUCLID(K=1, 5, 20) (Euclidean), CHI2(K=1, 5, 20), (χ^2-distance between distributions within K periods), CHI2fut (metric based on distributions of subsequent states)
Hamming	HAM (simple and generalized Hamming), DHD (Dynamic Hamming)
Optimal Matching (OM)	OM, OM(i=1.5), OM(trate), OM(indelslog), OM(indels), OM(future)
Localized Optimal Matching (OMloc)	OMloc(e=0, 0.1, 0.25, 0.4)
Spell-Length-Sensitive Optimal Matching (OMslen)	OMslen(h=1, i=1, 1.5, 5), OMslen(i=1, 1.5, 5)
Optimal Matching of Spell Sequences (OMspell)	OMspell(e=0, 0.1, 0.5, 1), OMspell(e=0, 0.1, 0.5, 1, i=2)
Optimal Matching of Transition Sequences (OMstran)	OMstran(w=0, 0.1, 0.5), OMstran(i=1.5, w=0.1, 0.5), OMstran(i=5, w=0.1, 0.5), OMstran(tm=raw)
Number of Matching Subsequences (NMS)	NMS
Subsequence Vectorial Representation (SVRspell)	SVRspell(b=0, 1, 2, 3), SVRspell(b=0, 1, 2, 3, a=1)

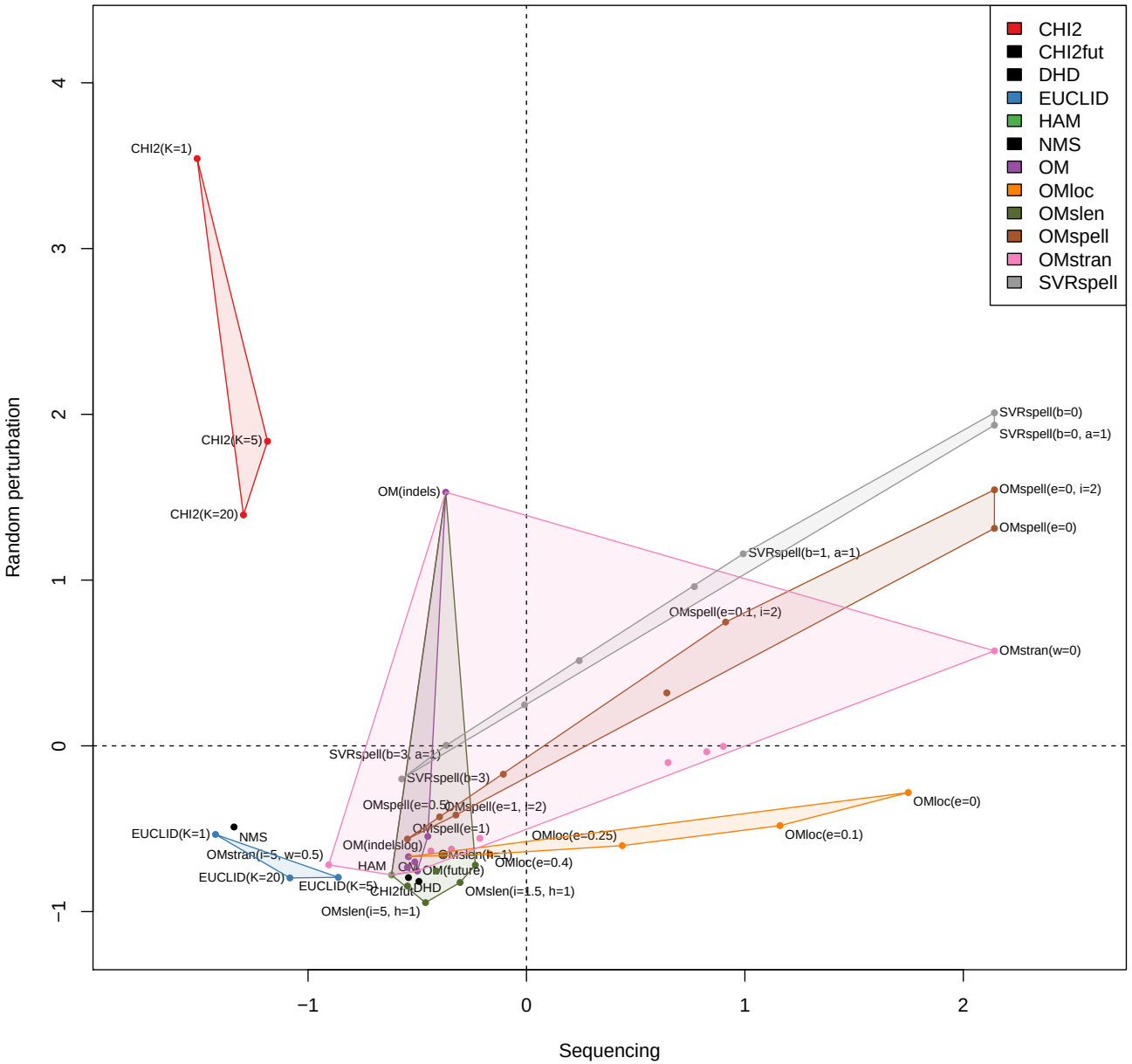
Scores for state-based simulations



Scores for event-based simulations



Random perturbation vs sequencing



Outline

Conclusions

- Similar overall scores for all distances, except NMS.
- Strange results for non-metric distances:
  - Localized OM.
  - Duration-sensitive OM.
  - “Optimized costs”.
- Advice: avoid non-metric distances.
- Limited effect of data-driven substitution costs.
  - does the added complexity worth it?
- Alternatives are available.

Guidelines

- Similar overall scores implies that a choice is needed.
- Which aspects to focus on?
- Family destandardisation:
  - Pattern change (rise of unmarried cohabitation).
  - Changes in age norms (age at marriage)
  - Changes in spacing (marriage–first child).
- Definition of the research question.

Guidelines

- Timing:
  - Hamming distances.
- Duration:
  - Optimal matching.
  - Optimal matching of spells.
  - Distribution-based distances.
- Sequencing (depending of sensitivity to small perturbation).
  - SVRspell (Very sensitive).
  - Optimal matching of spells (in between).
  - Optimal matching of transitions (less sensitive).
- Intermediary position:
  - SVRspell.
  - Optimal matching of transitions.
  - Optimal matching of spells.

Other uses

- By using one distance measure sensitive to each aspect.
  - Distinction stemming from each aspect.
  - Structuration of the data according to each aspect.
- In practice, aspects may be correlated.

Contributions

- Review of sequence dissimilarities.
- Guidelines.
- Methodology to evaluate sequence dissimilarities.
- New contribution:
  - Two new distances measures (OMspell and OM of transition)
  - Mostly sensitive to sequencing.
  - New strategies to set costs.
- All distances measures will be included in TraMineR software.
- Currently in the development R package “seqdist2”.

Studer, M. and G. Ritschard (2015). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. DOI: 10.1111/rssa.12125.

References I

Aisenbrey, S. and A. E. Fasang (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), 430–462.

Billari, F. C., J. Fürnkranz, and A. Prskawetz (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population 22*(1), 37–65.

Elzinga, C. H. (2003). Sequence similarity: A non-aligning technique. *Sociological Methods and Research 31*, 214–231.

Levine, J. (2000). But what have you done for us lately. *Sociological Methods & Research 29 (1)*, pp. 35–40. English

References II

Settersten, Richard A., J. and K. U. Mayer (1997). The measurement of age, age structuring, and the life course. *Annual Review of Sociology 23*, 233–261.

Studer, M. (2012). *Étude des inégalités de genre en début de carrière académique à l'aide de méthodes innovatrices d'analyse de données séquentielles*. Thèse de doctorat n° 777, Faculté des sciences économiquese et sociales, Université de Genève.

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