Multidomain/multichannel Sequence Analysis

Gilbert Ritschard

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

Sequence Analysis Association (SAA)
https://sequenceanalysis.org

Ceremade, Oct 2024

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

Outline

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

Outline

© G. Ritschard (2024), 3/38. Distributed under licence CC BY-NC-ND 3.0

Classical Sequence Analysis

- Abbott and Tsay (2000) describe typical SA as a 3-step program
  1. Coding narratives as sequences
     - state tokens (eg., {work, no work} vs {Full-time, Part-time, Unempl., No activity})
     - time granularity (e.g., yearly vs monthly)
  2. Computing pairwise dissimilarities between sequences
  3. Analyzing sequences based on their dissimilarities
     - building typologies by means of clustering
     - studying how clusters relate to covariates (sex, social origin, ...) by means of regression models

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

SA is much more (Liao et al., 2022)

- Visualization
- Multichannel (Multidimensional) sequences
- Dissimilarity-based analysis different from clustering (SOM, Anova-like, regression trees, representative sequences, ...)
- Non-dissimilarity-based clustering (LCA, Mixture of HMM, ...)
- Characteristics of individual sequences (Entropy, Complexity, Badness, ...)
- Sequence network
- Probabilistic sequence models (HMM, PST, ...)
- Combining SA with related methods (EHA, HMM, Multistate models, ...)

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

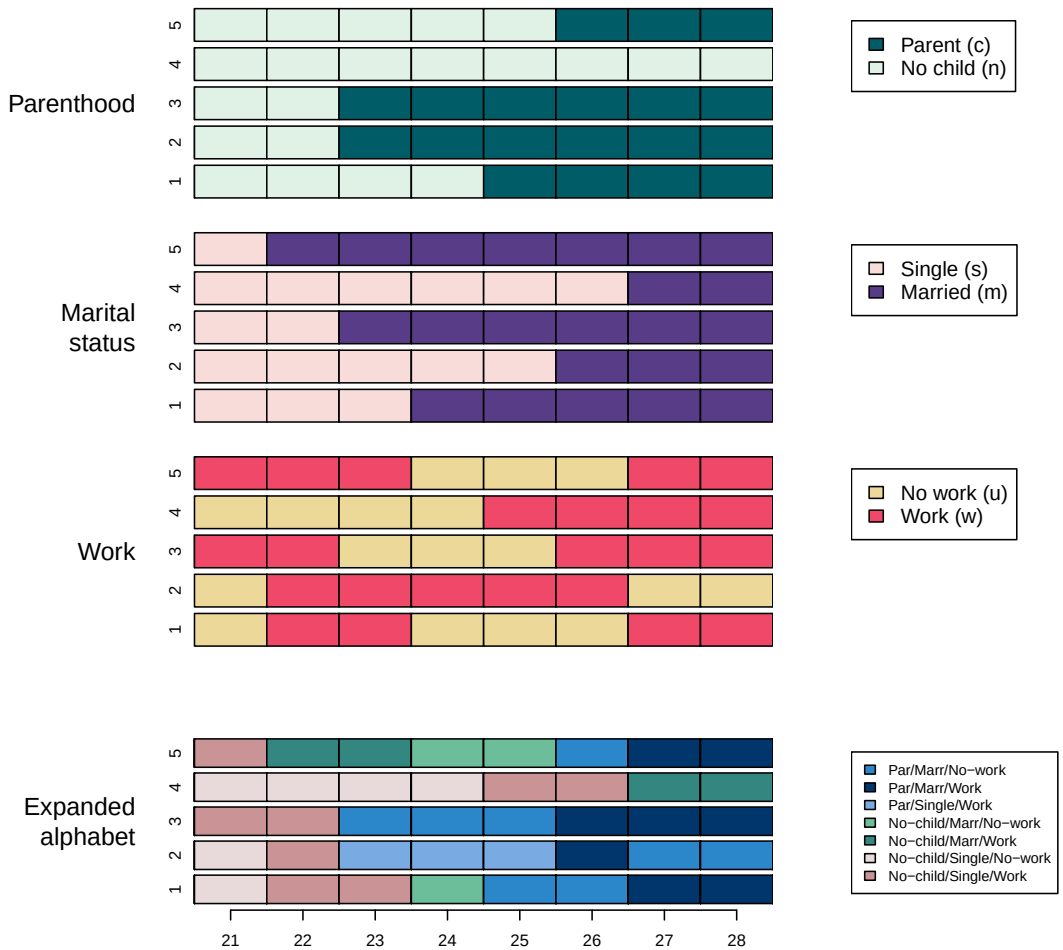
Multidomain SA: What is it?

- Analyzing how life trajectories:
  - in different domains: family, work, health, ... (Bernardi et al., 2019)
  - of linked people: partners, family members, ... (Elder et al., 2003)

  jointly unfold.
- Measuring association between domains of life (Piccarreta, 2017)
- Identifying spill over effects across domains (or linked people).

© G. Ritschard (2024), 6/38. Distributed under licence CC BY-NC-ND 3.0

Multidomain data: toy example data



© G. Ritschard (2024), 7/38. Distributed under licence CC BY-NC-ND 3.0

Outline

© G. Ritschard (2024), 8/38. Distributed under licence CC BY-NC-ND 3.0

Association between domains

- Joint analysis of domains of interest when domains are associated,
  i.e., when there are relationships between domains.
- Two types of association:
  - State association: when state tokens of one domain tend to co-occur with tokens of the other domain.
    - Measured by cross-tabulating token occurrences of the two domains
  - Trajectory association: when trajectories in one domain tend to co-occur with trajectories in the other domain.
    - Measured by correlation between pairwise distances in each of the domains.

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

Contingency tables of state occurrences

Toy MD sequences

	s	m
c	3	16
n	14	7

$v = 0.51^{***}$

	u	w
c	8	11
n	9	12

$v = 0.0076$

	u	w
s	6	11
m	11	12

$v = 0.13$

Note: Degree of significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

Association between domains: Toy MD sequences

Toy MD sequences. Association between domains. State association (Cramer's v) and trajectory association (Pearson's r correlation between pairwise INDELSLOG-based OM distances in each domain)

	Cramer's v	$p(v)$	Pearson's r	$p(r)$
Parent with Married	0.51	0.00	0.08	0.83
Parent with Work	0.01	0.96	0.05	0.88
Married with Work	0.13	0.43	0.35	0.32

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

Computing association measures between domains

Association between states

```
load(file = "data/toyMD.Rdata")
dnames <- c("Parent", "Married", "Work")
channels <- list(s.ch, s.mar, s.work)
seqdomassoc(channels, dnames = dnames)
```

```
##                       df      LRT p(LRT)      v    p(v)
## Parent_with_Married    1 11.24061 0.0008 0.5140 0.00115
## Parent_with_Work       1  0.00231 0.9617 0.0076 0.96169
## Married_with_Work      1  0.63265 0.4264 0.1253 0.42802
```

Computing association measures between domains

Association between trajectories

```
cost.meth <- "INDELSLOG"
toy.met = "OM"
d.ch <- seqdist(s.ch, method = toy.met, sm = cost.meth)
d.mar <- seqdist(s.mar, method = toy.met, sm = cost.meth)
d.work <- seqdist(s.work, method = toy.met, sm = cost.meth)
domdiss <- list(d.ch, d.mar, d.work)
dissdomassoc(domdiss, dnames = dnames)
```

```
## $Pearson
##         Parent Married   Work
## Parent  1.0000   0.080 0.0549
## Married 0.0800   1.000 0.3481
## Work    0.0549   0.348 1.0000
##
## $p.pearson
##            Parent   Married      Work
## Parent  1.69e-131  8.26e-01  8.80e-01
## Married  8.26e-01 1.99e-128  3.24e-01
## Work     8.80e-01  3.24e-01 2.97e-128
##
## $Pearson.Rsquare
##  Parent Married    Work
##  0.0047  0.0638  0.0621
##
## $dnames
## [1] "Parent"  "Married" "Work"
##
## attr(,"class")
## [1] "list"       "ddomassoc"
```

© G. Ritschard (2024), 13/38. Distributed under licence CC BY-NC-ND 3.0

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

Summary of domain association between domains

```
dass.traj <- dissdomassoc(domdiss, dnames = dnames, what = "all")
summary(dass.traj)
```

```
##                      Pearson p.Pearson Spearman p.Spearman
## Parent_with_Married   0.0800     0.826   0.1325      0.715
## Parent_with_Work      0.0549     0.880   0.0248      0.946
## Married_with_Work     0.3481     0.324   0.3676      0.296
```

© G. Ritschard (2024), 14/38. Distributed under licence CC BY-NC-ND 3.0

Outline

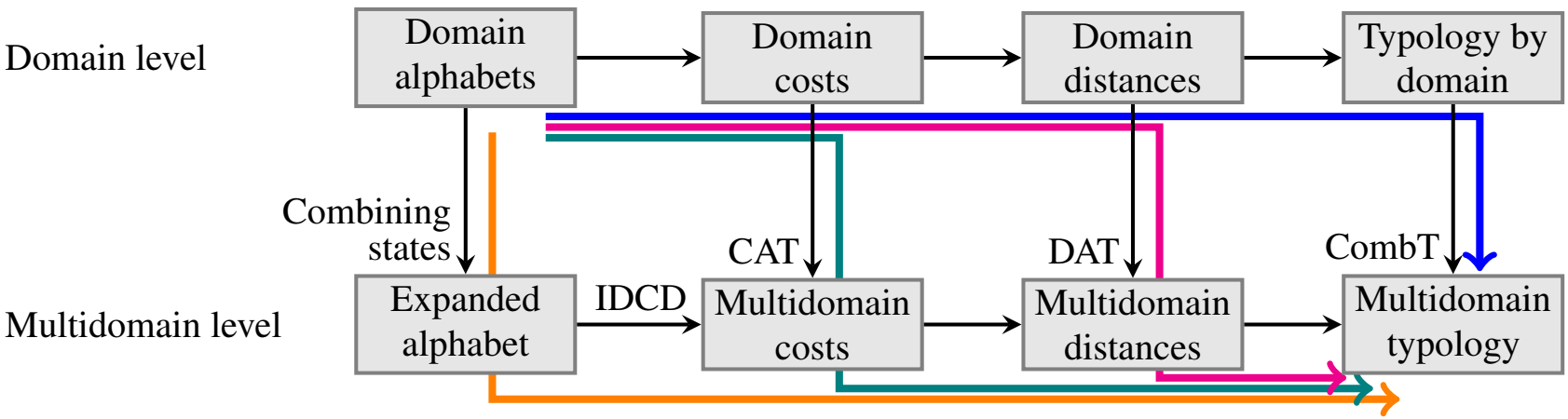
© G. Ritschard (2024), 15/38. Distributed under licence CC BY-NC-ND 3.0

Methods of multidomain SA

- As in classical SA, multidomain SA proceeds by building a typology
- In multidomain SA, we need a joint typology of the different domains
- Multidomain SA consists then in studying how the trajectories in the different domains are related within the different types (clusters).
  - For example: insecure work trajectories may typically co-occur with unstable family life trajectories and bad health pathways.
- Comparing the trajectories of the different domains within the clusters is generally done graphically: need for specific visualization tools of multidomain sequences.

© G. Ritschard (2024), 16/38. Distributed under licence CC BY-NC-ND 3.0

Strategies for building a multidomain typology



Using distances independent from domain costs and distances (IDCD)
Using distances based on costs derived from domain costs (CAT)
Using distances derived from domain distances (DAT)
Deriving multidomain typologies from domain types (CombT)

- Other approaches based on domain distances: MD distances derived from quantitative representations (principal coordinates, MDS) of sequences for each domain. (Example: GIMSA, Robette, Bry, and Lelièvre, 2015)

Independence constraints of additive tricks (CAT, DAT)

- CAT (Pollock, 2007; Gauthier et al., 2010)
  assumes independence across state tokens of the different domains:
  - Substitution cost between "`single + no child`" and "`married + child`" same as between "`single + child`" and "`married + no child`"
  - When domain costs are constant (e.g., all equal to 2), resulting multidomain substitution costs are proportional to number of non-matching domains.
    "`single + no child`" and "`married + no child`" differ on one domain.
- DAT (Han and Moen, 1999)
  assumes independence between trajectories:
  - Distance between `(s+n, s+n, ..., s+n)` and `(m+c, m+c, ..., m+c)` same as between `(s+c, s+c, ..., s+c)` and `(m+n, m+n, ..., m+n)`

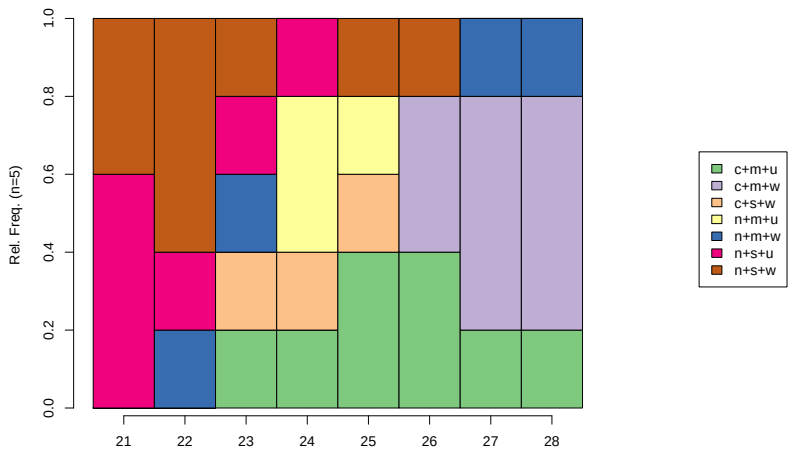
© G. Ritschard (2024), 18/38. Distributed under licence CC BY-NC-ND 3.0

Building MD sequences with expanded alphabet

```
(s.toyMD <- seqMD(channels = channels))

##   Sequence
## 1 n+s+u-n+s+w-n+s+w-n+m+u-c+m+u-c+m+u-c+m+w-c+m+w
## 2 n+s+u-n+s+w-c+s+w-c+s+w-c+s+w-c+m+w-c+m+u-c+m+u
## 3 n+s+w-n+s+w-c+m+u-c+m+u-c+m+u-c+m+w-c+m+w-c+m+w
## 4 n+s+u-n+s+u-n+s+u-n+s+u-n+s+w-n+s+w-n+m+w-n+m+w
## 5 n+s+w-n+m+w-n+m+w-n+m+u-n+m+u-c+m+u-c+m+w-c+m+w

seqdplot(s.toyMD, with.legend = "right")
```



© G. Ritschard (2024), 19/38. Distributed under licence CC BY-NC-ND 3.0

Retrieving CAT costs

```
(CATcost <- seqMD(channels, what = "cost", sm = "INDELSLOG"))
```

```
##       c+m+u c+m+w c+s+w n+m+u n+m+w n+s+u n+s+w
## c+m+u 0.000 0.578 1.156 0.576 1.154 1.154 1.731
## c+m+w 0.578 0.000 0.578 1.154 0.576 1.731 1.154
## c+s+w 1.156 0.578 0.000 1.731 1.154 1.154 0.576
## n+m+u 0.576 1.154 1.731 0.000 0.578 0.578 1.156
## n+m+w 1.154 0.576 1.154 0.578 0.000 1.156 0.578
## n+s+u 1.154 1.731 1.154 0.578 1.156 0.000 0.578
## n+s+w 1.731 1.154 0.576 1.156 0.578 0.578 0.000
## attr(,"indel")
## [1] 0.882 0.782 0.882 0.849 0.749 0.949 0.849
## attr(,"alphabet")
## [1] "c+m+u" "c+m+w" "c+s+w" "n+m+u" "n+m+w" "n+s+u" "n+s+w"
## attr(,"cweight")
## [1] 1 1 1
```

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

IDCD, CAT, and DAT distances

```
(dIDCD <- seqdist(s.toyMD, method = "OM", sm = "INDELSLOG"))
```

```
##      1    2    3    4    5
## 1 0.00 4.53 2.20 5.52 3.39
## 2 4.53 0.00 5.48 6.76 6.90
## 3 2.20 5.48 0.00 6.47 4.48
## 4 5.52 6.76 6.47 0.00 5.50
## 5 3.39 6.90 4.48 5.50 0.00
```

```
(dCAT <- seqdist(s.toyMD, method = "OM", sm = CATcost, indel = attr(CATcost,
    "indel")))
```

```
##      1    2    3    4    5
## 1 0.00 5.06 2.31 6.35 2.31
## 2 5.06 0.00 5.20 6.44 7.37
## 3 2.31 5.20 0.00 7.50 3.26
## 4 6.35 6.44 7.50 0.00 7.23
## 5 2.31 7.37 3.26 7.23 0.00
```

```
(dDAT <- d.ch + d.mar + d.work)
```

```
##      1    2    3    4    5
## 1 0.00 3.84 2.31 4.99 2.31
## 2 3.84 0.00 3.77 5.97 5.67
## 3 2.31 3.77 0.00 6.82 2.78
## 4 4.99 5.97 6.82 0.00 6.15
## 5 2.31 5.67 2.78 6.15 0.00
```

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

CombT: Deriving MD types from domain types

- Combining domain types $\Rightarrow$ (too) many combined types, many scarce types
- Optimally merging combined types (so as to minimize quality loss of the partitioning)
  - Function `dissmergegroups()`

Outline

© G. Ritschard (2024), 23/38. Distributed under licence CC BY-NC-ND 3.0

Rendering group of MD sequences

- Two ways of plotting MD sequences:
  - Plot of sequences of combined states
    - Large MD alphabet: Select contrasting colors for 10-12 most frequent tokens and display color legend for those tokens only.
  - For each group, separate plots by domains (groups by rows and domains by columns or conversely).

    TraMineR function `seqplotMD` since version 2.2-7
    - For plots depending on characteristics such as sequence order or representative sequences, characteristics must be determined at the MD level and applied to each domain level.

© G. Ritschard (2024), 24/38. Distributed under licence CC BY-NC-ND 3.0

Outline

© G. Ritschard (2024), 25/38. Distributed under licence CC BY-NC-ND 3.0

SHP biographic data (2000)

Swiss Household Panel

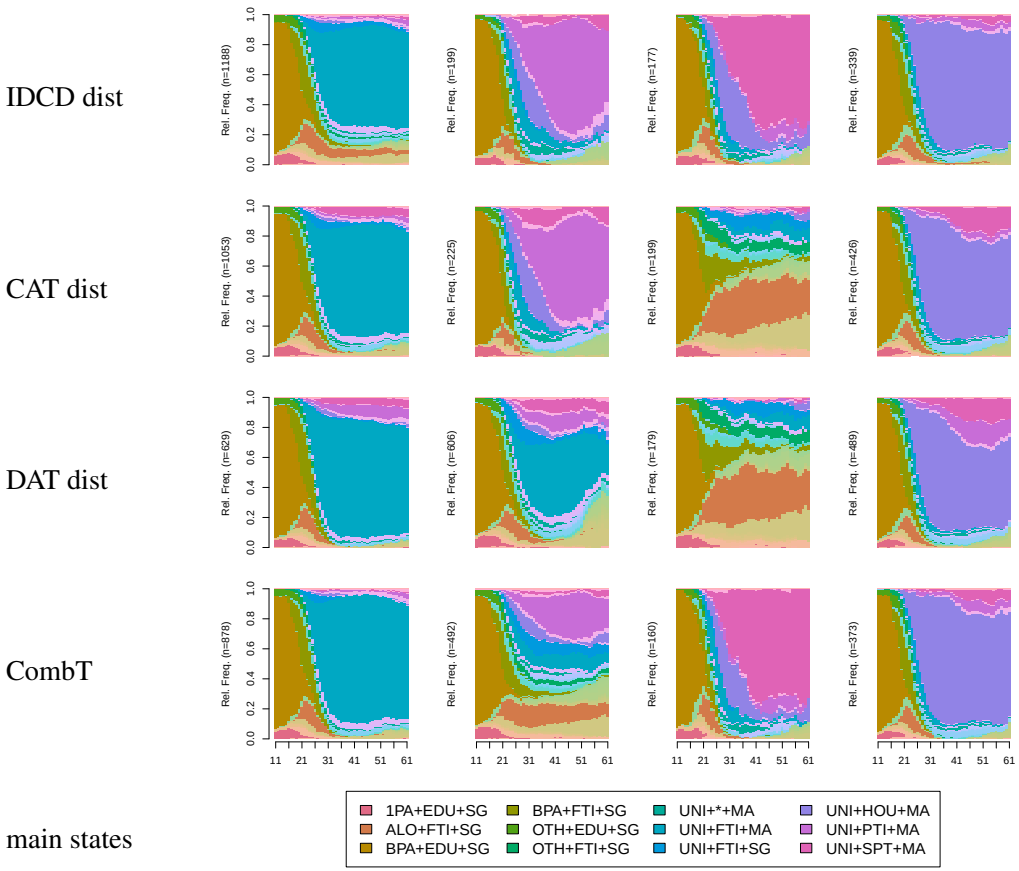
- Three domains: living arrangement (LA, 9 tokens), work life (WL, 10 tokens), civil status (CS, 5 tokens)
- 1903 MD life sequences, ages 11 to 61
- 229 observed combined tokens
- Pairwise domain association

	Cramer v	$p(v)$	Pearson r	$p(r)$
LA with WL	0.29	0.0000	-0.002	0.0375
LA with CS	0.46	0.0000	0.509	0.0000
WL with CS	0.34	0.0000	-0.001	0.2577

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

4-cluster solution using different strategies

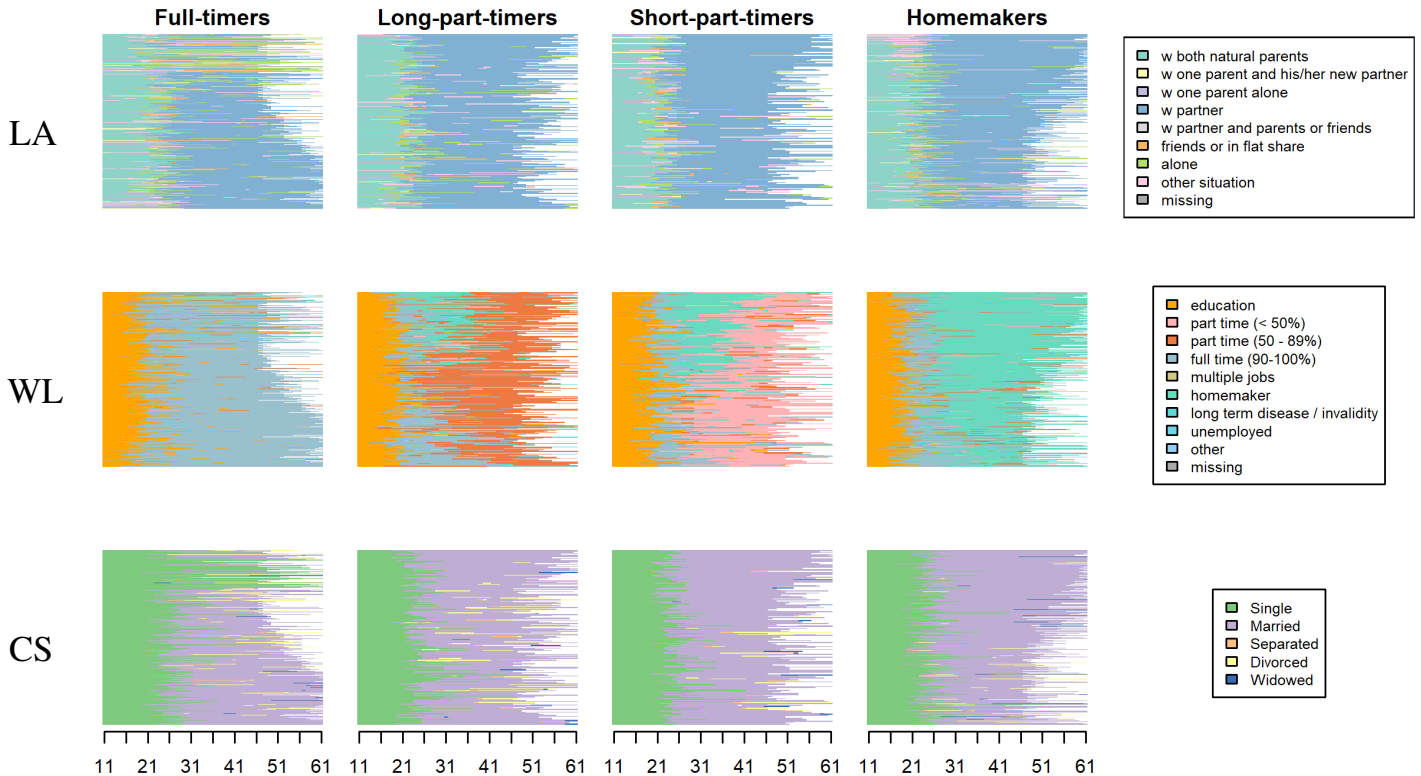
SHP. Three domains: living arrangement (LA), work life (WL), civil status (CS).
1903 MD life sequences, ages 11 to 61, 229 combined tokens



(Figure from Ritschard, Liao, and Struffolino, 2023)

IDCD 4-cluster solution by type (col) and domain (row)

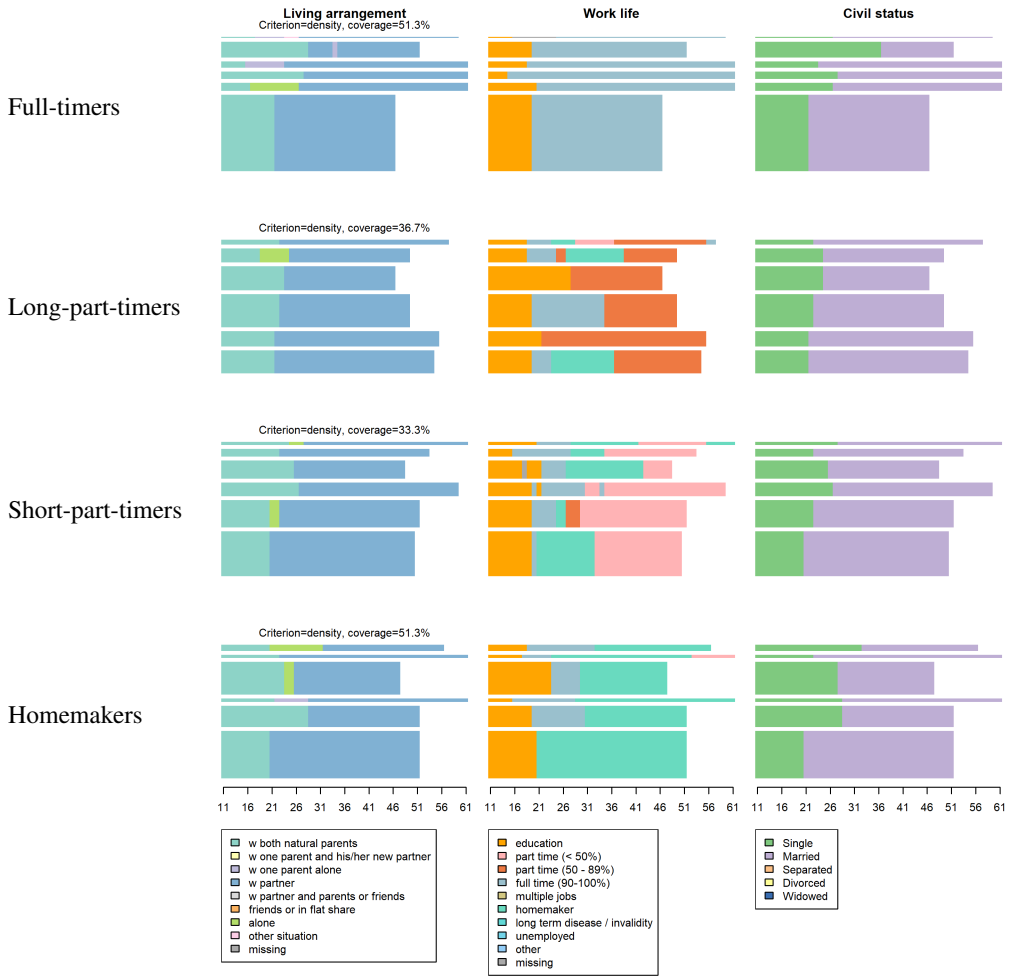
SHP. Three domains: living arrangement (LA), work life (WL), civil status (CS).
1903 MD life sequences, ages 11 to 61, 229 combined states



(Figure from Ritschard, Liao, and Struffolino, 2023)

IDCD 4-cluster solution: by type (row) and domain (col)

Representative MD sequences (Gabadinho and Ritschard, 2013)



(Figure from Ritschard, Liao, and Struffolino, 2023)

Aspects under development

- Tools to help identifying combinations of domain characteristics unique to each MD types.
- How does association between domains depend on covariates?
- Visualization:
  - Automatize efficient color selection for multidomain state tokens.
    - Degree of darkness proportional to mean domain rank of combined tokens
    - Contrasting colors for a subset of main tokens (`seqmaintokens`)

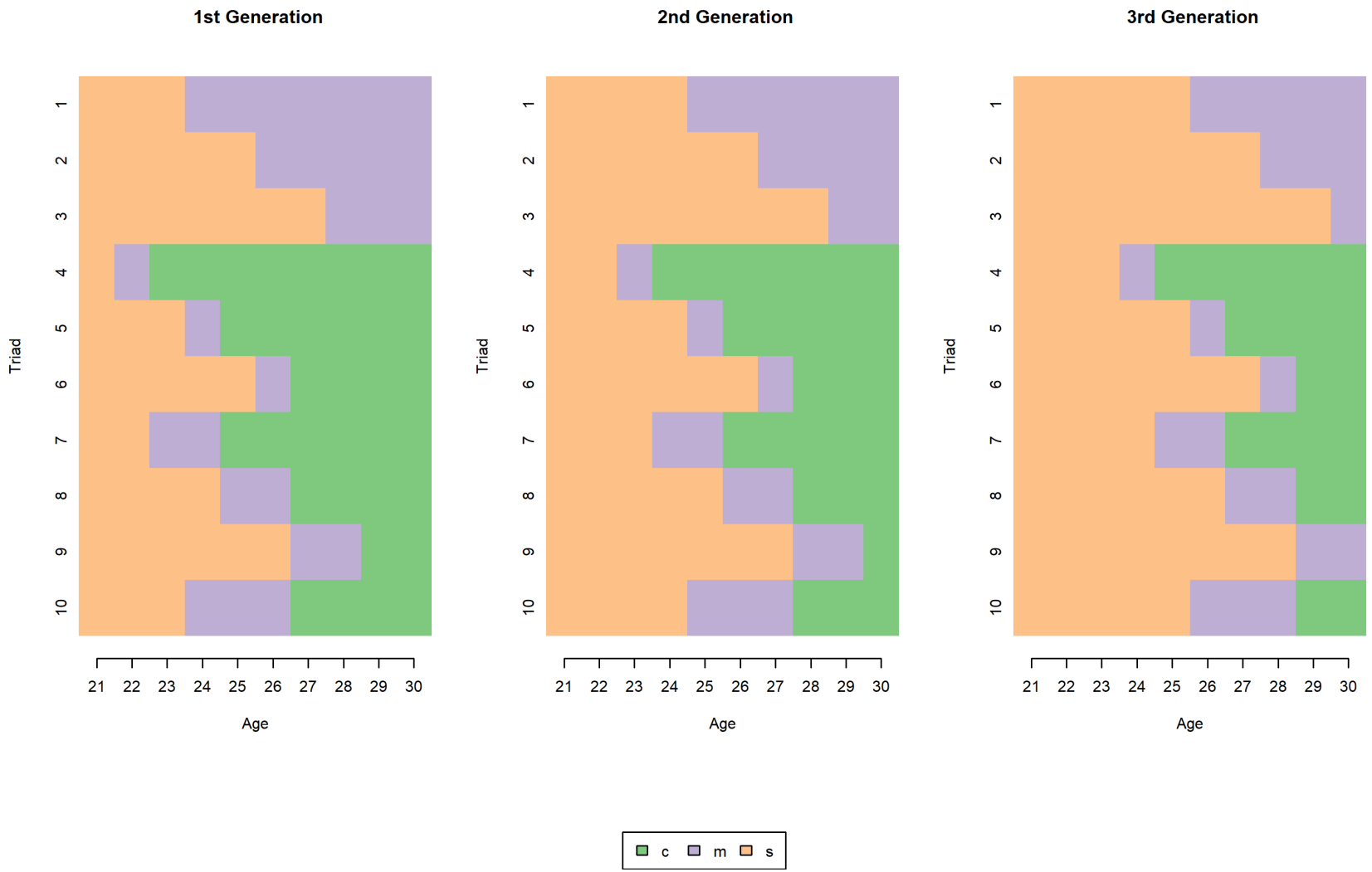
© G. Ritschard (2024), 30/38. Distributed under licence CC BY-NC-ND 3.0

Outline

© G. Ritschard (2024), 31/38. Distributed under licence CC BY-NC-ND 3.0

Dyadic/Polyadic Data

Example from help page of `TraMinerExtras::seqpolyads`



© G. Ritschard (2024), 32/38. Distributed under licence CC BY-NC-ND 3.0

Polyadic analysis

(Liao, 2021)

- Polyadic data is special case of multichannel data where all channels have same alphabet.
- This allows to compute pairwise dissimilarities between channels for each polyad.
- for measuring how strongly sequences are linked within polyads.
- Liao (2021) proposed two statistics:
  - U: difference between observed polyadic distance and mean distance of T randomized polyads
  - V: proportion of the T randomized polyads with greater polyadic distance than observed polyad

Polyadic Analysis

Example from help page of `TraMinerExtras::seqpolyads`

```
seqL <- list(seqGrandP, seqChild)
(seqG3.Tim <- seqpolyads(seqL, method = "HAM", a = 1, T = 100))
```

```
## $mean.dist
##  Obs Rand
## 3.10 4.42
##
## $U
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,] 2.42 2.42 2.42 1.42 1.42 1.42 0.42 0.42 0.42  0.42
##
## $U.tp
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]   [,7]   [,8]
## [1,] 4.02e-18 4.02e-18 4.02e-18 1.02e-08 1.02e-08 1.02e-08 0.0673 0.0673
##        [,9]  [,10]
## [1,] 0.0673 0.0673
##
## $V
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,] 0.84 0.84 0.84 0.66 0.66 0.66 0.52 0.52 0.52  0.52
##
## $V.95
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    0    0    0    0    0    0    0    0    0     0
##
## $observed.dist
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    2    2    2    3    3    3    4    4    4     4
##
```

© G. Ritschard (2024), 34/38. Distributed under licence CC BY-NC-ND 3.0

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

Thank you!
Questions?

© G. Ritschard (2024), 35/38. Distributed under licence CC BY-NC-ND 3.0

References I

Abbott, A. and A. Tsay (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research 29*(1), 3–33. (With discussion, pp 34-76).

Bernardi, L., J. Huinink, and R. A. Settersten, Jr (2019). The life course cube: A tool for studying lives. *Advances in Life Course Research 41*, 100258.

Elder, Jr, G. H., M. K. Johnson, and R. Crosnoe (2003). The emergence and development of life course theory. In J. T. Mortimer and M. J. Shanahan (Eds.), *Handbook of the Life Course*, pp. 3–19. Boston, MA: Springer.

Gabadinho, A. and G. Ritschard (2013). Searching for typical life trajectories applied to childbirth histories. In R. Levy and E. Widmer (Eds.), *Gendered life courses - Between individualization and standardization. A European approach applied to Switzerland*, pp. 287–312. Vienna: LIT-Verlag.

Gabadinho, A., G. Ritschard, N. S. Müller, and M. Studer (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software 40*(4), 1–37.

References II

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.

Han, S.-K. and P. Moen (1999). Clocking out: Temporal patterning of retirement. *American Journal of Sociology 105*(1), 191–236.

Liao, T. F. (2021). Using sequence analysis to quantify how strongly life courses are linked. *Sociological Science 8*(3), 48–72.

Liao, T. F., D. Bolano, B. Cornwell, C. Brzinsky-Fay, A. Fasang, S. Helske, R. Piccarreta, M. Raab, G. Ritschard, E. Struffolino, and M. Studer (2022). Sequence analysis: Its past, present, and future. *Social Science Research 107*, 102772.

Piccarreta, R. (2017). Joint sequence analysis: Association and clustering. *Sociological Methods & Research 46*(2), 252–287.

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.

References III

Ritschard, G., T. F. Liao, and E. Struffolino (2023). Strategies for multidomain sequence analysis in social research. *Sociological Methodology 53*(2), 288–322.

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