Exploring Sequential Data
A Tutorial

Gilbert Ritschard
Institute for Demographic and Life Course Studies, University of Geneva
and NCCR LIVES: Overcoming vulnerability, life course perspectives
http://mephisto.unige.ch/traminer

Discovery Science, Lyon, October 29-31, 2012

Outline

1. Introduction
2. Overview of what sequence analysis can do
3. About TraMineR

Objectives of the course

- Methods for extracting knowledge from sequence data
- Principles of sequence analysis
  - exploratory approaches
  - more causal and predictive approaches
- Practice of sequence analysis (TraMineR)

About longitudinal data: Sequence data

- Multiple cases ($n$ cases)
- For each case a sorted list of (categorical) values

Example:

1: a a d d c
2: a b b c c d
3: b c c
. . . . .
What is longitudinal data?

Longitudinal data
- Repeated observations on units observed over time (Beck and Katz, 1995).
- "A dataset is longitudinal if it tracks the same type of information on the same subjects at multiple points in time". (http://www.caldercenter.org/whatis.cfm)
- "The defining feature of longitudinal data is that the multiple observations within subject can be ordered" (Singer and Willett, 2003)

Successive transversal data vs longitudinal data

Successive transversal observations (same units)

<table>
<thead>
<tr>
<th>id</th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td>B</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>B</td>
<td>A</td>
</tr>
</tbody>
</table>

Longitudinal observations

<table>
<thead>
<tr>
<th>id</th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td>B</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>B</td>
<td>A</td>
</tr>
</tbody>
</table>

Repeated independent cross sectional observations

Successive independent transversal observations

<table>
<thead>
<tr>
<th>id</th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This is not longitudinal ...

Individual follow-ups: Each important event is recorded as soon as it occurs (medical card, cellular phone, weblogs, ...).
Panels: Periodic observation of same units
Retrospective data (biography): Depends on interviewees' memory
Matching data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)


Rotating panels: partial follow up
E.g.; Swiss Labor Force Survey, SLFS, 5 year-rotating panel (Wernli, 2010)
State sequences: an example

- Transition from school to work, (McVicar and Anyadike-Danes, 2002)
  Monthly states: EM = employment, TR = training, FE = further education, HE = higher education, SC = school, JL = joblessness
  Sequence

Compact representation

- Sequence 1 (EM, 4) → (TR, 2) → (EM, 64)
- Sequence 2 (FE, 36) → (HE, 34)
- Sequence 3 (TR, 24) → (FE, 34) → (EM, 10) → (JL, 2)
- Sequence 4 (TR, 47) → (EM, 14) → (JL, 9)

What is sequence analysis (SA)?

- Sequence analysis (SA)
  - concerned by categorical sequences,
  - holistic: interest is in the whole sequence, not just one element in the sequence (unlike survival analysis for example)
  - Aim is
    - Characterizing sets of sequences
    - Identifying typical (sequence) patterns
    - Studying relationship with individual characteristics and environment

Numerical longitudinal data: Essentially modeling approaches

- Multilevel models (Fixed and random effects) (Gelman and Hill, 2007; Frees, 2004)
  - Can handle mixed longitudinal-cross-sectional data, but do not really describe dynamics
- Growth curve models (specialized SEM) (McArdle, 2009)
- But also, distance-based analysis (DTW, ...)

Categorical longitudinal data

- Multilevel models for nominal and ordinal data (Hedeker, 2007; Müller, 2011)
- Survival approaches (descriptive survival curves and hazard regression models) (Therneau and Grambsch, 2000)
- Markov chain models and Probabilistic suffix trees (Berchtold and Raftery, 2002; Bejerano and Yona, 2001)
- Aligning techniques (biology) (Sharma, 2008)

Nature of sequences

- Chronological order?
- If yes, we can study timing and duration.
- Information conveyed by position \( j \) in the sequence
- If position is a time stamp, differences between positions reflect durations.
- Nature of the elements of the alphabet
- states, transitions or events, letters, proteins, ...
Exploring Sequential Data: Tutorial

Introduction

What is sequence analysis (SA)?

State versus event sequences

- An important distinction for chronological sequences is between state sequences and event sequences
  - A State, such as 'living with a partner' or 'being unemployed', lasts the whole unit of time
  - An event, such as 'moving in with a partner' or 'ending education', does not last but provokes a state change, possibly in conjunction with other events.

Time stamped events

<table>
<thead>
<tr>
<th></th>
<th>Sandra</th>
<th>Jack</th>
</tr>
</thead>
<tbody>
<tr>
<td>1979</td>
<td>Education</td>
<td>Education</td>
</tr>
<tr>
<td>1980</td>
<td>Education</td>
<td>End working in 1980</td>
</tr>
<tr>
<td>1981</td>
<td>End working in 1981</td>
<td></td>
</tr>
<tr>
<td>1982</td>
<td>End working in 1982</td>
<td></td>
</tr>
<tr>
<td>1983</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- There can be simultaneous events (see Sandra)
- Elements at same position do not occur at same time

State sequence view

<table>
<thead>
<tr>
<th>year</th>
<th>Sandra</th>
<th>Jack</th>
</tr>
</thead>
<tbody>
<tr>
<td>1979</td>
<td>Education</td>
<td>Education</td>
</tr>
<tr>
<td>1980</td>
<td>Education</td>
<td>Employed</td>
</tr>
<tr>
<td>1981</td>
<td>Employed</td>
<td>Employed</td>
</tr>
<tr>
<td>1982</td>
<td>Employed</td>
<td>Employed</td>
</tr>
<tr>
<td>1983</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Only one state at each observed time
- Position conveys time information: All states at position 2 are states in 1980.

Typical questions

- Are there standard sequences, types of sequences?
- How are those standards linked to covariates such as sex, birth cohort, ... ?
- How does some target variable (e.g., social status) depend on the followed sequence (lived trajectory)?
- . . .

Sequencing, timing and duration

- For chronological sequences (with time dimension)
- SA can answer questions about:
  - Sequencing: Order in which the different elements occur.
  - Timing: When do the different elements occur?
  - Duration: How long do we stay in the successive states?
Overview of sequence analysis outcomes

Aim:
- Show what kind of results can be obtained
- as well as how to get the results with our TraMineR package for R
- TraMineR: Trajectory Miner for R (Gabadinho et al., 2011)

The mvad example dataset

The mvad sequences are in STS form

The mvad sequences are organized in STS form, i.e., each sequence is given as a (row) vector of consecutive states.

```R
head(mvad[, 1:22])
```

- Sep.93 Oct.93 Nov.93 Dec.93 Jan.94 Feb.94
  1 employment employment employment employment training training
  2 FE FE FE FE FE FE
  3 training training training training training training
  4 FE FE FE FE FE FE
  5 joblessness training training training training training
  6 training training training training training training
  7 FE FE FE FE FE FE
  8 joblessness training training training training training
  9 FE FE FE FE FE FE

There are many other ways of organizing sequences data and TraMineR supports most of them.
Creating the state sequence object

- Most TraMineR functions for state sequences require a **state sequence object** as input argument.
- The state sequence object contains
  - the sequences
  - and their attributes (alphabet, labels, colors, weights, ...)
- Hence, we first have to create this object

### Starting TraMineR and creating a state sequence object

1. Load TraMineR and themvad data.
   - `library(TraMineR)`
   - `data(mvad)`
2. Check the alphabet (from Sept 93 to June 99; i.e., positions 17 to 86: We skip July-August 93)
   - `mvad.alph <- seqstatl(mvad[, 17:86])`
   - `[1] "employment" "FE" "HE" "joblessness" "school"
   - `[6] "training"
3. Create the 'state sequence' object
   - `mvad.lab <- c("employment", "further education", "higher education", "joblessness", "school", "training")`
   - `mvad.shortlab <- c("EM", "FE", "HE", "JL", "SC", "TR")`
   - `mvad.seq <- seqdef(mvad[, 17:86], alphabet = mvad.alph, states = mvad.shortlab, labels = mvad.lab, weights = mvad$weight, xtstep = 6)`

### Rendering sequences

- `seqfplot(mvad.seq, withlegend = FALSE, title = "f-plot", border = NA)`
- `seqdplot(mvad.seq, withlegend = FALSE, title = "d-plot", border = NA)`
- `seqIplot(mvad.seq, withlegend = FALSE, title = "I-plot", sortv = "from.end")`
- `seqlegend(mvad.seq, position = "bottomright", fontsize = 1.2)`

### Main sequence object attributes and seqdef arguments

<table>
<thead>
<tr>
<th>Attribute name</th>
<th>Description</th>
<th>Argument</th>
<th>Default</th>
<th>Retrieve/Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>alphabet</td>
<td>list of states</td>
<td>states=</td>
<td>from input data</td>
<td>alphabet()</td>
</tr>
<tr>
<td>cpal</td>
<td>color palette</td>
<td>cpal=</td>
<td>from RColorBrewer</td>
<td>cpal()</td>
</tr>
<tr>
<td>labels</td>
<td>long state labels</td>
<td>labels=</td>
<td>from input data</td>
<td>stlab()</td>
</tr>
<tr>
<td>cnames</td>
<td>position names</td>
<td>cnames=</td>
<td>from input data</td>
<td>names()</td>
</tr>
<tr>
<td>xtstep</td>
<td>jumps between tick marks</td>
<td>xtstep=</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>row.names</td>
<td>row (sequence) labels</td>
<td>id=</td>
<td>from input data</td>
<td>rownames()</td>
</tr>
<tr>
<td>weights</td>
<td>optional case weights</td>
<td>weights=</td>
<td>NULL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>missing handling</td>
<td>left=</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>gaps=</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>right=</td>
<td>&quot;DEL&quot;</td>
<td></td>
</tr>
</tbody>
</table>

- seqfplot(mvad.seq, withlegend = FALSE, title = "f-plot", border = NA)
- seqdplot(mvad.seq, withlegend = FALSE, title = "d-plot", border = NA)
- seqIplot(mvad.seq, withlegend = FALSE, title = "I-plot", sortv = "from.end")
- seqlegend(mvad.seq, position = "bottomright", fontsize = 1.2)
Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Rendering sequences
Rendering sequences by group (sex)
seqplot(mvad.seq, group = mvad$male, sortv = "from.start", title = "Sex")

Characterizing set of sequences

- Sequence of transversal measures (modal state, between entropy, ...)
- Summary of longitudinal measures (within entropy, transition rates, mean duration ...)
- Other global characteristics: sequence medoid, diversity of sequences, ...

Transition rates

round(trate <- seqtrate(mvad.seq), 3)

Mean time in each state by qualification gained at end of compulsory school

seqmtplot(mvad.seq, group = mvad$gcse5eq, title = "End CS qualification")
Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Characterizing set of sequences

**Sequence of transversal distributions**
For bad qualification at end of compulsory school, 9 months

```
seqstad(mvad.seq[mvad$gcse5eq == "bad", 6:15])
```

**State frequencies**
- Feb.94: EM 0.08, FE 0.18, HE 0.00, JL 0.10, SC 0.33, TR 0.31
- Mar.94: EM 0.094, FE 0.181, HE 0.00, JL 0.093, SC 0.316, TR 0.316
- Apr.94: EM 0.10, FE 0.176, HE 0.00, JL 0.11, SC 0.316, TR 0.315
- May.94: EM 0.11, FE 0.16, HE 0.00, JL 0.15, SC 0.28, TR 0.31
- Jun.94: EM 0.13, FE 0.14, HE 0.00, JL 0.16, SC 0.17, TR 0.32
- Jul.94: EM 0.22, FE 0.17, HE 0.00, JL 0.15, SC 0.094, TR 0.32
- Aug.94: EM 0.23, FE 0.212, HE 0.00, JL 0.091, SC 0.171, TR 0.316
- Sep.94: EM 0.211, FE 0.211, HE 0.00, JL 0.084, SC 0.171, TR 0.295
- Oct.94: EM 0.231, FE 0.209, HE 0.00, JL 0.084, SC 0.171, TR 0.292
- Nov.94: EM 0.244, FE 0.209, HE 0.00, JL 0.084, SC 0.171, TR 0.292

**Valid states**

**Entropy index**
- Feb.94: 0.82, Mar.94: 0.83, Apr.94: 0.83, May.94: 0.84, Jun.94: 0.85, Jul.94: 0.87, Aug.94: 0.86, Sep.94: 0.86

End CS qualification − bad
Freq. (weighted n=429.56)
- Sep.93 to Sep.98: employment, further education, higher education, joblessness, school, training

End CS qualification − good
Freq. (weighted n=282.01)
- Sep.93 to Sep.98: employment, further education, higher education, joblessness, school, training

Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Characterizing set of sequences

**Sequence of modal states**
by qualification gained at end of compulsory school

```
seqmsplot(mvad.seq, group = mvad$gcse5eq, title = "End CS qualification", border = NA)
```

End CS qualification − bad
Modal state sequence (0 occurrences, freq=0%)
- Sep.93 to Sep.98: employment, further education, higher education, joblessness, school, training

End CS qualification − good
Modal state sequence (0 occurrences, freq=0%)
- Sep.93 to Sep.98: employment, further education, higher education, joblessness, school, training

Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Characterizing set of sequences

**Transversal entropies**
Time evolution of the transversal state diversity

```
seqplot.tentrop(mvad.seq, title = "End CS qualification", group = mvad$gcse5eq)
```

End CS qualification
Entropy
- Sep.93 to Mar.99: bad, good
- Sep.93 to Sep.98: employment, further education, higher education, joblessness, school, training

Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Characterizing set of sequences
Exploring Sequential Data: Tutorial  
Overview of what sequence analysis can do  
Longitudinal characteristics

### Longitudinal Characteristics

- **Characteristics of individual sequences**
  - `seqlength()` length of the sequence
  - `seqtransn()` number of transitions
  - `seqsubsn()` number of sub-sequences
  - `seqdss()` list of the distinct successive states (DSS)
  - `seqdur()` list of the durations in the states of the DSS
  - `seqistatd()` time in each state (longitudinal distribution)
  - `seqient()` Longitudinal entropy
  - `seqST()` Turbulence (Elzinga and Liefbroer, 2007)
  - `seqici()` Complexity index (Gabadinho et al., 2011)

### Complexity of the sequences

- To evaluate the complexity of a sequence we may consider
  - **Longitudinal entropy**
    - does not account for the sequencing of the states
    - (AABB and ABAB have same entropy)
  - **Turbulence** (Elzinga and Liefbroer, 2007)
    - composite measure based on
      - the number of sub-sequences of the DSS sequence
      - the variance of the durations of the successive states
    - sensitive to state sequencing
  - **Index of complexity** (Gabadinho et al., 2010, 2011)
    - composite measure based on
      - the number of transitions
      - the longitudinal entropy
    - sensitive to state sequencing

---

**Computing the sequence complexity measures**

```r
mvad.ient <- seqient(mvad.seq)
mvad.cplx <- seqici(mvad.seq)
mvad.turb <- seqST(mvad.seq)
tab <- data.frame(mvad.ient, mvad.cplx, mvad.turb)
```
Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Longitudinal characteristics

Comparing the measures

```
plot(ctab)
```

![Graph showing plots of Entropy, C, and Turbulence](image)

Analyzing how complexity is related to covariates
Regressing complexity on covariates

```
lm.ici <- lm(mvad.cplx ~ male + fumemp + gcse6eq, data = mvad)
```

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 0.109    | 0.004      | 28.01   | 0.000    |
| male       | -0.013   | 0.004      | -3.04   | 0.002    |
| father unemployed | 0.007   | 0.006      | 1.24    | 0.216    |
| good ECS grade | 0.010   | 0.005      | 2.20    | 0.028    |

Distribution of complexity by sex

```
boxplot(mvad.cplx ~ mvad$male, col = "lightsteelblue")
```

![Box plot showing distribution of complexity by sex](image)

Pairwise dissimilarities between sequences

- Distance between sequences
  - Different metrics (LCP, LCS, OM, HAM, DHD)
- Once we have pairwise dissimilarities, we can
  - Partition a set of sequences into homogeneous clusters
  - Identify representative sequences (medoid, densest neighborhood)
  - Self-organizing maps (SOM) of sequences (Massoni et al., 2009)
  - MDS scatterplot representation of sequences
  - Measure the discrepancy between sequences
  - Discrepancy analysis of a set of sequences (ANOVA)
  - Grow regression trees for explaining the sequence discrepancy
Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Dissimilarity-based analyses

Summary of available distances

<table>
<thead>
<tr>
<th>Distance</th>
<th>Method</th>
<th>Position-wise</th>
<th>Additional arguments</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Count of common attributes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Hamming</td>
<td>HAM</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Longest Common Prefix</td>
<td>LCP</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Longest Common Suffix</td>
<td>RLC</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Longest Common Subsequence</td>
<td>LCS</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edit distances</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal Matching</td>
<td>OM</td>
<td>No</td>
<td>Insertion/deletion costs (indel) and substitution costs matrix (sm)</td>
<td></td>
</tr>
<tr>
<td>Hamming</td>
<td>HAM</td>
<td>Yes</td>
<td>substitution costs matrix (sm)</td>
<td></td>
</tr>
<tr>
<td>Dynamic Hamming</td>
<td>DHD</td>
<td>Yes</td>
<td>substitution costs matrix (sm)</td>
<td></td>
</tr>
</tbody>
</table>

Dissimilarity matrix

- **TraMineR** provides the `seqdist` function

```r
# OM distances with custom indel and substitution costs
# costs used by McVicar and Anyadike-Danes (2012).
sbm.custom <- matrix(
  c(0,1,1,2,1,1,
   1,0,1,2,1,2,
   1,1,0,3,1,2,
   2,2,3,0,3,1,
   1,1,3,0,2,1,
   1,2,2,1,2,0),
  nrow = 6, ncol = 6, byrow = TRUE,
  dimnames = list(mvad.shortlab, mvad.shortlab))
mvad.dist <- seqdist(mvad.seq, method="OM", indel=4, sm=sbm.custom)
dim(mvad.dist)
```

```
[1] 712 712
```

Dissimilarity matrix

- Other distances

  - There exist many other distances not yet implemented in **TraMineR**.
    - Distances based on counts of common subsequences (Elzinga, 2003, 2007b)
    - Distances based on counts of common subsequences of length 2 (Oh and Kim, 2004)
    - Distances based on scores of multiple correspondence analysis (Grelet, 2002)
    - Distances accounting for the common future (Rousset et al., 2011)
    - Plenty of variants of Optimal Matching (Hollister, 2009; Halpin, 2010; Gauthier et al., 2009)
    - OM of transitions instead of states (Biemann, 2011)
  - Matthias Studer compares over 30 distances in his PhD thesis (Studer, 2012a).
Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Dissimilarity-based analyses

Cluster analysis

- Can run any clustering method which accepts a dissimilarity matrix as input.
- Many solutions in R:
  - For hierarchical clustering
    - `hclust()` base function (can account for weights)
    - Package `cluster` (does not accept weights!):
      - `agnes()`: agglomerative nesting (average, UPGMA, WPGMA, ward, beta-flexible, ...)
      - `diana()`: divisive partitioning
  - For PAM and other direct partitioning methods
    - Packages: `cluster`, `fastclust`, `flashClust`, ...
    - `WeightedCluster` (currently only available from R-Forge, Studer 2012b)

Example: Hierarchical clustering (Ward)

```r
mvad.clusterward <- hclust(as.dist(mvad.dist), method = "ward",
                           members = mvad$weight)
plot(mvad.clusterward, labels = FALSE)
```

PAM clustering

- PAM much faster, but must set a priori number $k$ of clusters.
- `WeightedCluster` offers nice tools to help selecting $k$.
- $k = 4$ was found to be good choice.
- PAM with function `wcKMedoids` from `WeightedCluster`

```r
library(WeightedCluster)
set.seed(4)
pam.mvad <- wcKMedoids(mvad.dist, k = 4, weight = mvad$weight)
```

Cluster membership is in `pam.mvad$clustering`

```r
mvad.cl4 <- pam.mvad$clustering
table(mvad.cl4)
```

Labeling the PAM clusters

```r
seqdplot(mvad.seq, group = group.p(mvad.cl4), border = NA)
```

Rearranging cluster order and defining labels

```r
                 "Joblessness")
mvad.cl4.factor <- factor(mvad.cl4, levels = c(467, 66, 607, 641), labels = c14.labels)
```
Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Dissimilarity-based analyses

Mean time in each state

```
seqmtplot(mvad.seq, group = mvad.cl4.factor)
```

Most frequent sequences

```
seqfplot(mvad.seq, group = mvad.cl4.factor, border = NA)
```

Individual sequences (sorted by states from start)

```
seqIplot(mvad.seq, group = mvad.cl4.factor, sortv = "from.start")
```

Sorted by states from the end

```
seqIplot(mvad.seq, group = mvad.cl4.factor, sortv = "from.end")
```
Overview of what sequence analysis can do

Dissimilarity-based analyses

Representative sequences (Gabadinho et al., 2011)
Smallest set of patterns with given percentage of sequences in their neighborhood

seqrplot(mvad.seq, group = mvad.cl4.factor, dist.matrix = mvad.dist, trep = 0.6, sim = 0.15, border = NA, cex.legend = 1.6)

Computing the dispersion

- For the whole set of sequences
dissvar(mvad.dist)
  [1] 32.06
- By cluster (dissvar.grp from library TraMineRextras)
data.frame(Dispersion = dissvar.grp(mvad.dist, group = mvad.cl4.factor))

Analysis of sequence discrepancy

- Running an ANOVA-like analysis for gcse5eq
da <- dissassoc(mvad.dist, group = mvad$gcse5eq, R = 1000)
print(da)

Discrepancy of sequences

- Sum of squares SS can be expressed in terms of distances between pairs
  \[ SS = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (y_i - y_j)^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{ij} \]

- Setting \( d_{ij} \) equal to OM, LCP, LCS ... distance, we get SS.
- From which we can measure the dispersion with the pseudo-variance \( SS/n \).
- And run ANOVA analyses (Studer et al., 2011, 2010, 2009).
Exploring Sequential Data: Tutorial

Overview of what sequence analysis can do

Dissimilarity-based analyses

ANOVA output

Pseudo ANOVA table:

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp</td>
<td>1952</td>
<td>1</td>
<td>1952.4</td>
</tr>
<tr>
<td>Res</td>
<td>20871</td>
<td>710</td>
<td>29.4</td>
</tr>
<tr>
<td>Total</td>
<td>22823</td>
<td>711</td>
<td>32.1</td>
</tr>
</tbody>
</table>

Test values (p-values based on 1000 permutation):

- $t_0$: p.value = 0.001
- Pseudo F: $66.41934$ (p.value: 0.001)
- Pseudo Fbf: $67.37188$ (p.value: 0.001)
- Pseudo $R^2$: $0.08555$ (p.value: 0.001)
- Bartlett: $0.14693$ (p.value: 0.001)
- Levene: $0.77397$ (p.value: 0.001)

Inconclusive intervals:

- $0.00383 < 0.01 < 0.0162$
- $0.03649 < 0.05 < 0.0635$

Discrepancy per level:

- bad: 452, $29.76$
- good: 260, $28.53$
- Total: 712, $32.06$

Exploring Sequential Data: Tutorial

Overview of what sequence analysis can do

Dissimilarity-based analyses

Differences over time

- How do differences between groups vary over time?
- At which age do trajectories most differ across birth cohorts?
- Compute $R^2$ for short sliding windows (length 2)
- We get thus a sequence of $R^2$, which can be plotted
- Similarly, we can plot series of
  - total within (residual) discrepancy ($SS_W$)
  - within discrepancy of each group ($SS_C$)
Overview of what sequence analysis can do

Dissimilarity-based analyses

Grade at end of compulsory school

mvad.diff <- seqdiff(mvad.seq, group = mvad$gcseSeq)

mvad.diff$stat[c(1, 13, 25, 37), ]

<table>
<thead>
<tr>
<th></th>
<th>Pseudo F</th>
<th>Pseudo Fbf</th>
<th>Pseudo R²</th>
<th>Bartlett</th>
<th>Levene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep.93</td>
<td>41.46</td>
<td>44.64</td>
<td>0.05520</td>
<td>9.87187</td>
<td>76.271</td>
</tr>
<tr>
<td>Sep.94</td>
<td>72.00</td>
<td>77.42</td>
<td>0.09213</td>
<td>9.49256</td>
<td>104.501</td>
</tr>
<tr>
<td>Sep.95</td>
<td>50.52</td>
<td>50.37</td>
<td>0.06646</td>
<td>0.06569</td>
<td>1.041</td>
</tr>
<tr>
<td>Sep.96</td>
<td>104.80</td>
<td>103.06</td>
<td>0.12869</td>
<td>0.76633</td>
<td>2.748</td>
</tr>
</tbody>
</table>

mvad.diff$discrepancy[c(1, 13, 25, 37), ]

<table>
<thead>
<tr>
<th></th>
<th>bad</th>
<th>good</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep.93</td>
<td>0.3620</td>
<td>0.2561</td>
<td>0.3387</td>
</tr>
<tr>
<td>Sep.94</td>
<td>0.3876</td>
<td>0.2761</td>
<td>0.3783</td>
</tr>
<tr>
<td>Sep.95</td>
<td>0.3590</td>
<td>0.3691</td>
<td>0.3888</td>
</tr>
<tr>
<td>Sep.96</td>
<td>0.2862</td>
<td>0.3147</td>
<td>0.3415</td>
</tr>
</tbody>
</table>

Plotting R-squares over time

plot(mvad.diff, lwd = 3, col = "darkred", xtstep = 6)

Plotting within discrepancies over time

plot(mvad.diff, lwd = 3, stat = "discrepancy", xtstep = 6,
     legendposition = "bottomleft")

Tree structured discrepancy analysis

Objective: Find the most important predictors and their interactions.

Iteratively segment the cases using values of covariates (predictors)

Such that groups be as homogenous as possible.

At each step, we select the covariate and split with highest $R^2$.

Significance of split is assessed through a permutation $F$ test.

Growing stops when the selected split is not significant.
Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Dissimilarity-based analyses

Growing the tree

```
dt <- seqtree(mvad.seq ~ male + Grammar + funemp + gcseSeq +
              fmpr + livboth, weighted = FALSE, data = mvad, diss = mvad.dist,
              R = 5000)
print(dt, gap = 3)
```

Graphical tree

  
```
R> seqtreedisplay(dt, filename = "fg_mvadseqtree.png",
                    type = "d", border = NA)
```

- The plot is produced as a png file and displayed with the default program associated to this extension.

Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Dissimilarity-based analyses

Tree in text form

```
Dissimilarity tree:
Parameters: minSize=35.6, maxdepth=5, R=5000, pval=0.01
Formula: mvad.seq ~ male + Grammar + funemp + gcseSeq + fmpr + livboth
Global R2: 0.12

Fitted tree:
|-- Root (n: 712 disc: 32)
  |-> gcseSeq 0.086
  |   |-- [ bad ] (n: 452 disc: 30)
  |     |--> funemp 0.017
  |     |   |-- [ no ] (n: 362 disc: 26)
  |     |     |--> male 0.014
  |     |     |   |-- [ female ] (n: 146 disc: 31)[(FE,2)-(EM,68)] *
  |     |     |   |-- [ male ] (n: 216 disc: 25)[(EM,70)] *
  |     |     |   |-- [ yes ] (n: 90 disc: 36)[(EM,70)] *
  |     |   |-- [ good ] (n: 260 disc: 29)
  |     |   |-> Grammar 0.048
  |     |     |-- [ no ] (n: 183 disc: 30)[(FE,22)-(EM,48)] *
  |     |     |-- [ yes ] (n: 77 disc: 21)[(SC,25)-(HE,45)] *
```

Exploring Sequential Data: Tutorial
Overview of what sequence analysis can do
Dissimilarity-based analyses

Graphical Tree

- Legend:
  - Male:
  - Female:
  - Grammar:
  - Funemp:
  - Gcse5eq:
  - Fmpr:
  - Livboth:

- Global null model R2: 0.12
- Legend: R2: 0.12


- The plot is produced as a png file and displayed with the default program associated to this extension.
Exploring Sequential Data: Tutorial

Overview of what sequence analysis can do

Dissimilarity-based analyses

Graphical Tree, using I-plots and showdepth=TRUE

About TraMineR

TraMineR: What is it?

TraMineR

- Trajectory Miner in R: a toolbox for exploring, rendering and analyzing categorical sequence data

- Developed within the SNF (Swiss National Fund for Scientific Research) project Mining event histories 1/2007-1/2011

- ... development goes on within IP 14 methodological module of the NCCR LIVES: Overcoming vulnerability: Life course perspectives (http://www.lives-nccr.ch).

TraMineR: Where and why in R?

- Package for the free open source R statistical environment
  - freely available on the CRAN (Comprehensive R Archive Network) http://cran.r-project.org
  - `R> install.packages("TraMineR", dependencies=TRUE)`

- TraMineR runs in R, it can straightforwardly be combined with other R commands and libraries. For example:
  - dissimilarities obtained with TraMineR can be inputted to already optimized processes for clustering, MDS, self-organizing maps, ...  
  - TraMineR’s plots can be used to render clustering results;  
  - complexity indexes can be used as dependent or explanatory variables in linear and non-linear regression, ...

TraMineR: Who?

- Under supervision of a scientific committee:
  - Gilbert Ritschard (Statistics for social sciences)
  - Alexis Gabadinho (Demography)
  - Nicolas S. Müller (Sociology, Computer science)
  - Matthias Studer (Economics, Sociology)

- Additional members of the development team:
  - Reto Bürgin (Statistics)
  - Emmanuel Rousseaux (KDD and Computer science)

both PhD students within NCCR LIVES IP-14
TraMineR’s features

- Handling of longitudinal data and conversion between various sequence formats
- Plotting sequences (distribution plot, frequency plot, index plot and more)
- Individual longitudinal characteristics of sequences (length, time in each state, longitudinal entropy, turbulence, complexity and more)
- Sequence of transversal characteristics by position (transversal state distribution, transversal entropy, modal state)
- Other aggregated characteristics (transition rates, average duration in each state, sequence frequency)
- Dissimilarities between pairs of sequences (Optimal matching, Longest common subsequence, Hamming, Dynamic Hamming, Multichannel and more)
- Representative sequences and discrepancy measure of a set of sequences
- ANOVA-like analysis and regression tree of sequences
- Rendering and highlighting frequent event sequences
- Extracting frequent event subsequences
- Identifying most discriminating event subsequences
- Association rules between subsequences

Other programs for sequence analysis

- Optimize (Abbott, 1997)
  - Computes optimal matching distances
  - No longer supported
- TDA (Rohwer and Pötter, 2002)
  - Free statistical software, computes optimal matching distances
- Stata, SQ-Ados (Brzinsky-Fay et al., 2006)
  - Free, but license required for Stata
  - Optimal matching distances, visualization and a few more
  - See also the add-ons by Brendan Halpin
    http://teaching.sociology.ul.ie/seqanal/
- CHESA free program by Elzinga (2007a)
  - Various metrics, including original ones based on non-aligning methods
  - Turbulence

References I

References II


References III


References IV


References V


Exploring Sequential Data: Tutorial
About TraMineR
References VI


References VII


References VIII