An introduction to sequence analysis using the TraMineR R package

Matthias Studer,
Alexis Gabadinho, Nicolas S. Müller, Gilbert Ritschard

Institute for Demographic and Life Course Studies, University of Geneva
http://mephisto.unige.ch/traminer

Workshop on Trajectories, Paris, October 14th, 2011

Objectives

- Presentation of the “TraMineR” software.
- Illustrate the main features using an example.
  - Descriptive approach and visualization.
  - Discrepancy analysis.
- Present the main resources available to learn and start using “TraMineR”.

Plan

1. TraMineR
2. A TraMineR preview
3. Other features
4. Getting started

TraMineR

- TraMineR is an R package for sequence analysis.
- Specially designed for the social sciences.
- TraMineR: Trajectory Miner in R

  (Not inspired by our liking for the Gewürztraminer wine)
- Freely available on the CRAN (Comprehensive R Archive Network)

  `http://cran.r-project.org/web/packages/TraMineR/`
- To install: `install.packages("TraMineR")`

Why TraMineR is an R package?

- R is a environment for statistical computing and graphics.
- It is open source and multi-platform.
- Provides basic and advanced statistical methods.
- Since TraMineR is an R package, you may:
  - Use all data file format supported by R (SPSS, Stata, ...).
  - Analyze results produced by TraMineR using methods provided by other packages (optimized clustering procedure, MDS, multi-level models, ...).
  - Use results produced by other packages directly in TraMineR (build sequence of network properties, for instance).

The TraMineR project

- TraMineR was launched as a FNS project:
- Mining event histories: Towards new insights on personal Swiss life courses.
- Project FN-113998 et FN-122230 de février 2007 à janvier 2011
- Development continues within the LIVES NCCR: “overcoming vulnerability: life course perspectives” (IP 14).

Scientific committee

- Gilbert Ritschard, professor of statistics
- Alexis Gabadinho, demography
- Nicolas S. Müller, sociology and information systems
- Matthias Studer, economy and sociology

Sequence Analysis in the Social Sciences

- TraMineR is designed to answer questions arising in the social sciences.
- Sequences describe life trajectories or more generally social processes.
  - Professional carriers.
  - Cohabitational life courses.
  - History of organizations.
- Sequence analysis provide an holistic view of the trajectories.
- Unlike "event-oriented" approach (Billari, 2001), states and transitions are analyzed in the context of the whole process.

Common questions

Abbott (1990) identifies three common questions in sequence analysis.

- Are there typical (recurrent) patterns of trajectories? What are those patterns?
- How are the trajectories related to explanatory factors? Which factors influence the trajectory followed by an individual? Do we observe differences according to cohort, social origin or sex?
- How is a given outcome, such as health status or income, related to a previous trajectory?

TraMineR features

- Description of states sequences
    - Visualization of a set of states sequences.
    - Compute descriptive statistics.
- Compute dissimilarities between states sequences.
- Build and visualize a typology of states sequences (using other R packages).
- Analyze the links between states sequences and explanatory covariates using discrepancy analysis (Studer et al., 2011).

Plan

1. TraMineR
2. A TraMineR preview
3. Other features
4. Getting started

- Study from McVicar and Anyadike-Danes (2002) on transition from school to work in Northern Ireland.
- Aim: identify which young people are more at risk to experience unsuccessful transitions into the labour market.
- Use TraMineR to:
    - Visualize the sequences and descriptive methods.
    - Building a typology.
    - Discrepancy analysis.

Example: The mvad data set

- 712 individuals
- Follow-up starting at the end of the compulsory education (July 1993)
- Time series of 70 status variables: September 1993 to June 1999.
- The alphabet is made of the following statuses: EM (Employment), FE (Further Education), HE (Higher Education), JL (Joblessness), SC (School), TR (Training).
- Included in the TraMineR library.

The mvad data set - Variable list

Table: List of Variables in the `MVAD` data set

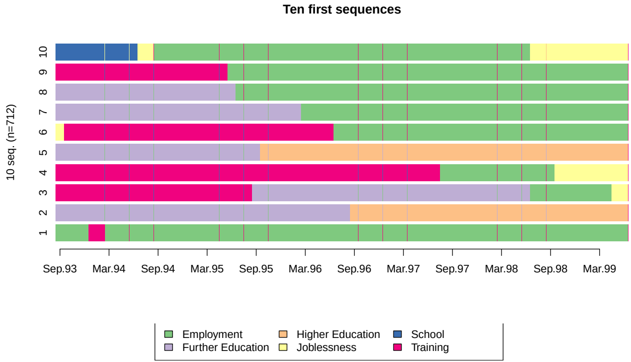
id	unique individual identifier
weight	sample weights
sex	binary dummy for gender, 1=male
religion	protestant or catholic
region	location of school, one of five Education and Library Board areas in Northern Ireland
Grammar	binary dummy indicating type of secondary education, 1=grammar school
funemp	binary dummy indicating father's employment status at time of survey, 1=father unemployed
gcse5eq	binary dummy indicating qualifications gained by the end of compulsory education, 1=5+ GCSEs at grades A-C, or equivalent
fmpr	binary dummy indicating SOC code of father's current or most recent job,1=SOC1 (professional, managerial or related)
livboth	binary dummy indicating living arrangements at time of first sweep of survey (June 1995), 1=living with both parents
jul93	Monthly Activity Variables are coded 1-6, 1=school, 2=FE, 3=employment, 4=training, 5=joblessness, 6=HE
:	"
jun99	"

Definition of a state sequences

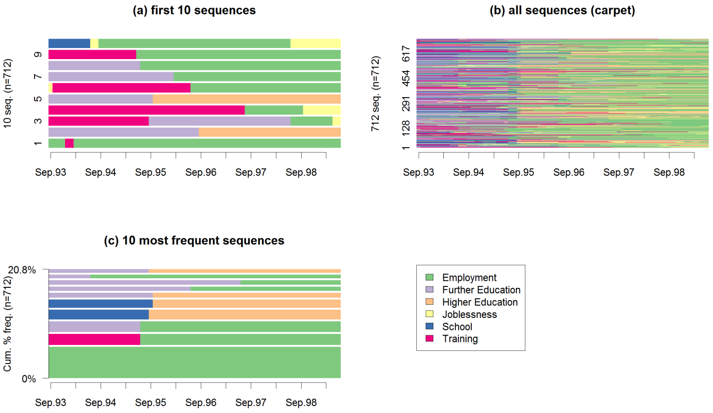
Definitions

- Alphabet A: finite set of possible states.
- Sequence of length k: ordered list of k elements taken from A.

```
R> mvad.seq <- seqdef(mvad, 17:86, labels = mvad.lab, xtstep = 6)
R> seqiplot(mvad.seq, border = NA, title = "Ten first sequences")
```



Visualization based on individual sequences



(a) first 10 sequences

(b) all sequences (carpet)

(c) 10 most frequent sequences

Transversal statistics

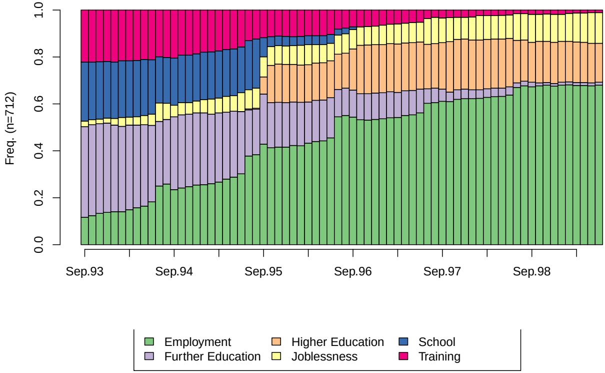
- Summarize a set of sequences using a sequence of transversal statistics.

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

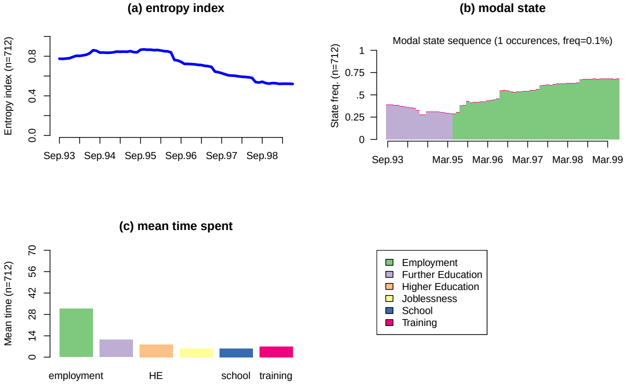
- On may use:
  - The states distribution.
  - The entropy index (Billari, 2001).
  - The modal state.
- Use overall statistics: mean time spent in each state.

Chronogram

```
R> seqdplot(mvad.seq)
```



Transversal statistics



Building a typology

- Compute dissimilarity between sequences.
- Regroup similar sequences using cluster analysis.
- Visualize the results.

Dissimilarity: Concepts

- Most of advanced sequence analysis methods rely on a dissimilarity measure.
- A dissimilarity is a quantification of how far two objects are.
- For instance, consider two incomes x and y:
  - $d(x, y) = |x - y|$
  - $d(x, y) = \log(1 + |x - y|)$
  - $d(x, y) = (x - y)^2$
- How to do it with categorical sequences?
- Depending on the issue, we want our dissimilarity measure to account for:
  - Order of the states and transitions in each sequence.
  - Temporality of the transitions.
  - Duration of stay in each state.

Dissimilarity: Concepts

- Most of advanced sequence analysis methods rely on a dissimilarity measure.
- A dissimilarity is a quantification of how far two objects are.
- For instance, consider two incomes x and y:
    - $d(x, y) = |x - y|$
    - $d(x, y) = \log(1 + |x - y|)$
    - $d(x, y) = (x - y)^2$
- How to do it with categorical sequences?
- Depending on the issue, we want our dissimilarity measure to account for:
    - Order of the states and transitions in each sequence.
    - Temporality of the transitions.
    - Duration of stay in each state.

Dissimilarity: Concepts

- Most of advanced sequence analysis methods rely on a dissimilarity measure.
- A dissimilarity is a quantification of how far two objects are.
- For instance, consider two incomes x and y:
  - $d(x, y) = |x - y|$
  - $d(x, y) = \log(1 + |x - y|)$
  - $d(x, y) = (x - y)^2$
- How to do it with categorical sequences?
- Depending on the issue, we want our dissimilarity measure to account for:
  - Order of the states and transitions in each sequence.
  - Temporality of the transitions.
  - Duration of stay in each state.

Dissimilarity: Concepts

- Most of advanced sequence analysis methods rely on a dissimilarity measure.
- A dissimilarity is a quantification of how far two objects are.
- For instance, consider two incomes x and y:
  - $d(x, y) = |x - y|$
  - $d(x, y) = \log(1 + |x - y|)$
  - $d(x, y) = (x - y)^2$
- How to do it with categorical sequences?
- Depending on the issue, we want our dissimilarity measure to account for:
  - Order of the states and transitions in each sequence.
  - Temporality of the transitions.
  - Duration of stay in each state.

Dissimilarity: Concepts

- Most of advanced sequence analysis methods rely on a dissimilarity measure.
- A dissimilarity is a quantification of how far two objects are.
- For instance, consider two incomes x and y:
  - $d(x, y) = |x - y|$
  - $d(x, y) = \log(1 + |x - y|)$
  - $d(x, y) = (x - y)^2$
- How to do it with categorical sequences?
- Depending on the issue, we want our dissimilarity measure to account for:
  - Order of the states and transitions in each sequence.
  - Temporality of the transitions.
  - Duration of stay in each state.

Dissimilarity measures

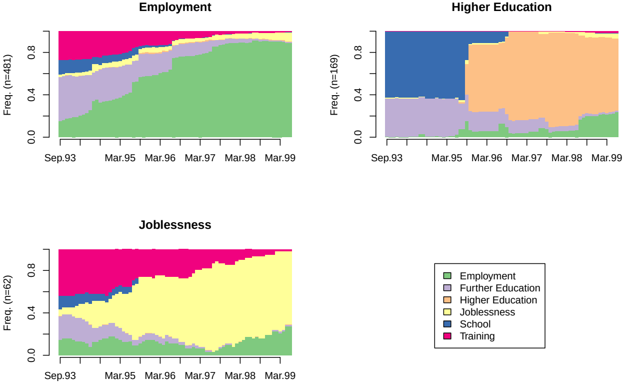
Dissimilarity measures provided by TraMineR:

- Optimal Matching (OM).
- Longest Common Prefix (LCP).
- Longest Common Suffix (RLCP).
- Longest Common Subsequence (LCS).
- Hamming distance (HAM).
- Dynamic Hamming Distance (DHD) (Lesnard, 2010).
- Other measures are planned.

Cluster Analysis

- Keep three clusters (best silhouette width).
- Quality is poor and may be an artifact (average silhouette width=0.41) (Kaufman and Rousseeuw, 1990).

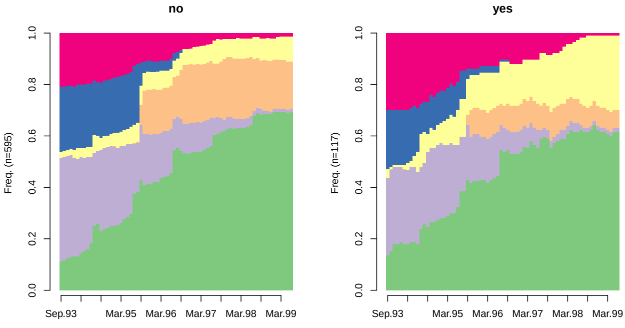
```
R> seqdplot(mvad.seq, group = pamclustfac, border = NA)
```



Comparing groups of sequences

- Do the sequences differ according to father unemployment status?
- Are those differences *significant*? And what happens if we control for the region?

```
R> seqdplot(mvad.seq, group = mvad$funemp, border = NA)
```



🟩 Employment 🟧 Higher Education 🟦 School
🟪 Further Education 🟨 Joblessness 🟥 Training

Discrepancy Analysis of State Sequences

Discrepancy analysis (Studer et al., 2011):

- Allow to study the links between state sequences and explanatory covariate.
- Measure the strength of the association using the share of the discrepancy of the sequences “explained” by a given explanatory covariate.
- Attest the significance of the association by estimating the “p-value”.

Main principles

- In the Euclidean case:

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}
\end{aligned}
$$

- Replacing d_{ij} with OM, LCP, LCS ... we define a discrepancy measure of a set of sequence.
- We may then use the ANOVA principles to compute the R^2.
- R^2 is the share of the total discrepancy explained by a given covariate.
- Estimate significance using permutation tests.
- Permutation test provides an empirical estimation of the probability that a random partition of the sequences explains a biggest part of the discrepancy than our covariate.

Main principles

- In the Euclidean case:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}$$

- Replacing d_{ij} with OM, LCP, LCS ... we define a discrepancy measure of a set of sequence.
- We may then use the ANOVA principles to compute the R^2.
- R^2 is the share of the total discrepancy explained by a given covariate.
- Estimate significance using permutation tests.
- Permutation test provides an empirical estimation of the probability that a random partition of the sequences explains a biggest part of the discrepancy than our covariate.

Main principles

- In the Euclidean case:

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i-y_j)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij} \end{aligned}$$

- Replacing d_{ij} with OM, LCP, LCS ... we define a discrepancy measure of a set of sequence.
- We may then use the ANOVA principles to compute the R^2.
- R^2 is the share of the total discrepancy explained by a given covariate.
- Estimate significance using permutation tests.
- Permutation test provides an empirical estimation of the probability that a random partition of the sequences explains a biggest part of the discrepancy than our covariate.

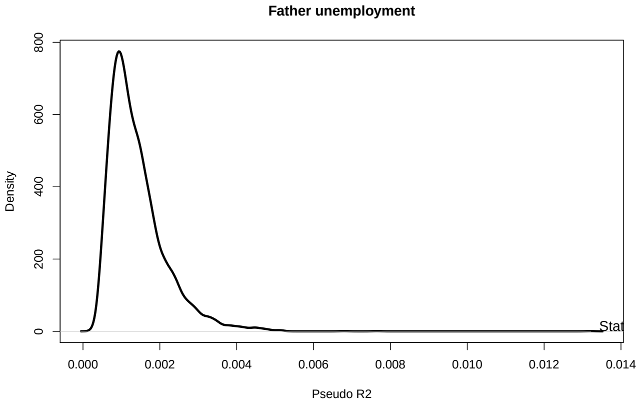
Main principles

- In the Euclidean case:

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(y_i - y_j)^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d_{ij}
\end{aligned}
$$

- Replacing d_{ij} with OM, LCP, LCS ... we define a discrepancy measure of a set of sequence.
- We may then use the ANOVA principles to compute the R^2.
- R^2 is the share of the total discrepancy explained by a given covariate.
- Estimate significance using permutation tests.
- Permutation test provides an empirical estimation of the probability that a random partition of the sequences explains a biggest part of the discrepancy than our covariate.

Null distribution of R^2



Discrepancy Analysis

- Bivariate association with each explanatory factor.

```
R> funemp_assoc <- dissassoc(mvad.dist, mvad$funemp, R = 5000)
```

	F	R^2	p-value
gcse5eq	67.69	0.087	0.000
Grammar	23.16	0.032	0.000
funemp	9.51	0.013	0.000
fmpr	8.76	0.012	0.000
sex	6.84	0.010	0.000
region	5.50	0.030	0.000
religion	2.75	0.004	0.012
livboth	2.23	0.003	0.035

Multi-factors analysis

- Allow to control for the influence of another covariate.
- For instance, controlling for the region.

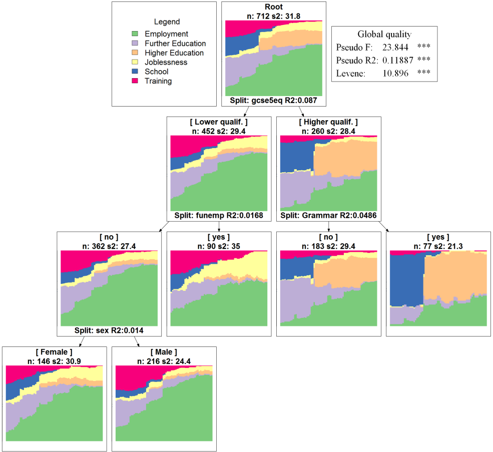
```
R> dmfac <- dissmfacw(mvad.dist ~ funemp + region, data = mvad,
+       R = 5000, squared = FALSE)
```

	F	R^2	p-value
funemp	7.13	0.010	0.000
region	4.90	0.027	0.000
Total	5.87	0.040	0.000

Induction tree

- Recursively select the “best” explanatory covariate.
- Use only significant split.

```
R> st <- seqtree(mvad.seq ~ gcse5eq + Grammar + funemp + sex,
+        data = mvad, R = 5000, diss = mvad.dist)
R> seqtreedisplay(st, type = "d", border = NA)
```



Plan

1. TraMineR
2. A TraMineR preview
3. Other features
4. Getting started

Other TraMineR features

- Handling and conversions between various longitudinal data format.
- Support for weights and missing values.
- Extraction and visualization of representative sequences of a set of sequence.
- Other descriptives statistics (transition rates, ...).
- Compute longitudinal characteristic of individual sequence (Complexity index, longitudinal entropy, turbulence, time spent in each state, ...)
- Homogeneity of discrepancy
- Analysis of event sequences.
  - Extraction of frequent event subsequences.
  - Identification of discriminant subsequences.

Longitudinal data formats

Code	Example										
STS	*Id*	*a18*	*a19*	*a20*	*a21*	*a22*	*a23*	*a24*	*a25*	*a26*	*a27*
	101	S	S	S	M	M	MC	MC	MC	MC	D
	102	S	S	S	MC	MC	MC	MC	MC	MC	MC
DSS	*Id*	*s1*	*s2*	*s3*	*s4*						
	101	S	M	MC	D						
	102	S	MC								
SPS	*Id*	*s1*	*s2*	*s3*	*s4*						
	101	(S,3)	(M,2)	(MC,4)	(D,1)						
	102	(S,3)	(MC,7)								
SPELL	*Id*	*Index*	*From*	*To*	*State*						
	101	1	18	20	Single (S)						
	101	2	21	22	Married (M)						
	101	3	23	26	Married w Children (MC)						
	101	4	27	27	Divorced (D)						
	102	1	18	20	Single (S)						
	102	2	21	27	Married w Children (MC)						
TSE	*Id*	*Time*	*Event*								
	101	21	M (Marriage)								
	101	23	C (Childbirth)								
	101	26	C (Childbirth)								
	101	27	D (Divorce)								
	102	21	M (Marriage)								
	102	21	C (Childbirth)								

Plan

1. TraMineR
2. A TraMineR preview
3. Other features
4. Getting started

Getting started: selected bibliography

- Introduction to states sequences analysis using TraMineR:
  - 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.
- Discrepancy analysis:
  - Studer, M., G. Ritschard, A. Gabadinho and N. S. Müller (2011). Discrepancy Analysis of State Sequences. *Sociological Methods and Research 40*(3), 471–510.

Other important resources

- TraMineR website: `http://mephisto.unige.ch/traminer`
  - latest news.
  - A TraMineR preview with all R command needed to reproduce this presentation.
  - Link to documetation resources.
    - A user guide (approx. 120 pages)
    - Tutorials and presentation.
    - A list of publication made by TraMineR users (preprints).
  - Information about training in Sequence Analysis and TraMineR.
- A TraMineR user mailing list: `http://mephisto.unige.ch/traminer/contrib.shtml`
- Bug report and feature request: `http://mephisto.unige.ch/traminer/contrib.shtml`

References I

Abbott, A. (1990). A primer on sequence methods. *Organization Science 1*(4), 375–392.

Billari, F. C. (2001). The analysis of early life courses: Complex description of the transition to adulthood. *Journal of Population Research 18*(2), 119–142.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding groups in data. an introduction to cluster analysis*. New York: Wiley.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods and Research 38*, 389–419.

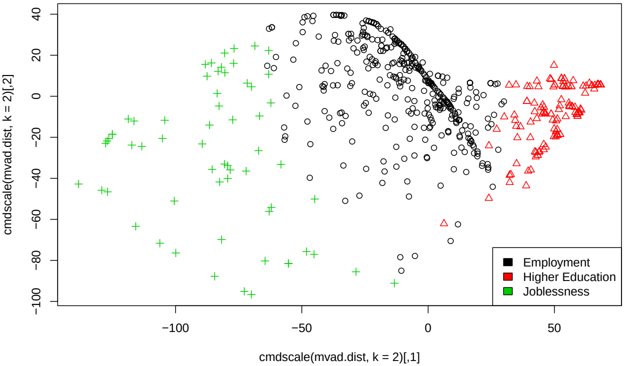
McVicar, D. and M. Anyadike-Danes (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society A 165*(2), 317–334.

References II

Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research 40*(3), 471–510.

Appendix

```
R> plot(cmdscale(mvad.dist, k = 2), col = pamclust, pch = pamclust)
R> legend("bottomright", legend = clust.labels, fill = 1:3)
```



Homogénéité des dispersions

- Le test d'homogénéité des dispersions permet de tester l'égalité des dispersions intra-groupes.
- Est-ce que la dispersion diffère significativement d'un groupe à l'autre ?
- Deux versions du test dans TraMineR
  - Basé sur une généralisation du test de Bartlett
  - Basé sur une généralisation du test de Levene (à utiliser de préférence).
- La significativité est attestée à l'aide de tests de permutation.

Test d’homogénéité des dispersions avec TraMineR

- Test d’homogénéité des dispersions selon la variable `Grammar`

```
R> catholic.assoc <- dissassoc(mvad.dist, group = mvad$catholic,
+       R = 5000)
```