Methods for Longitudinal Data Categorical Response

Gilbert Ritschard

Institute for demographic and life course studies, University Geneva
http://mephisto.unige.ch

LIVES Swiss National Center of Competence in Research

Doctoral Program, Lausanne, May 20, 2011

Typology of methods for life course data

	Issues	
Questions	duration/hazard	state/event sequencing
descriptive	• Survival curves: Parametric (Weibull, Gompertz, ...) and non parametric (Kaplan-Meier, Nelson-Aalen) estimators.	• Sequence clustering • Frequencies of given patterns • Discovering typical episodes
causality	• Hazard regression models (Cox, ...) • Survival trees	• Markov models • Mobility trees • Association rules among episodes

Survival Approaches

Event history analysis

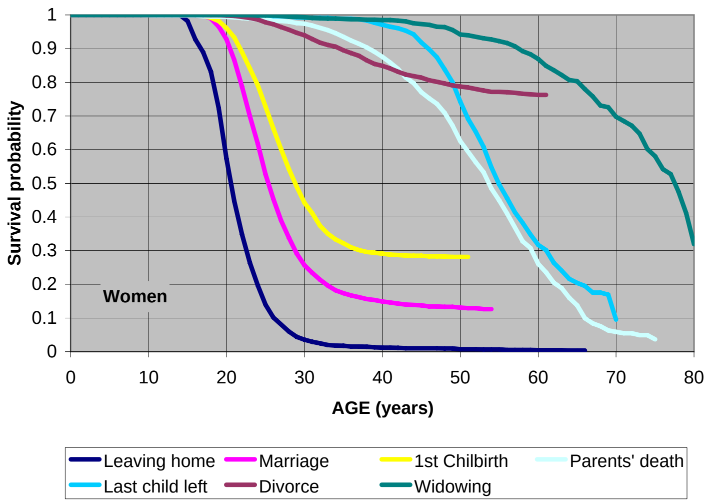
- Survival or Event history analysis (Mills, 2011)(Blossfeld and Rohwer, 2002)
  - Focuses on one event.
  - Concerned with duration until event occurs or with hazard of experiencing event.
- Survival curves: Distribution of duration until event occurs

$$S(t) = p(T \geq t) \ .$$

- Hazard models: Regression like models for $S(t, \mathbf{x})$ or hazard $h(t) = p(T = t \mid T \geq t)$

$$h(t, \mathbf{x}) = g\Big(t, \beta_0 + \beta_1 x_1 + \beta_2 x_2(t) + \cdots\Big) \ .$$

Survival curves (Switzerland, SHP 2002 biographical survey)



SHP biographical retrospective survey

http://www.swisspanel.ch

- SHP retrospective survey: 2001 (860) and 2002 (4700 cases).
- We consider only data collected in 2002.
- Data completed with variables from 2002 wave (language).

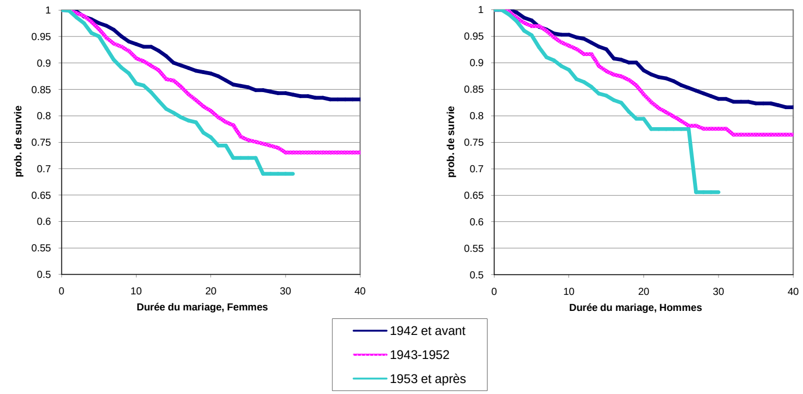
Characteristics of retained data for divorce

(individuals who get married at least once)

	men	women	Total
Total	1414	1656	3070
1st marriage dissolution	231	308	539
	16.3%	18.6%	17.6%

Marriage duration until divorce

Survival curves



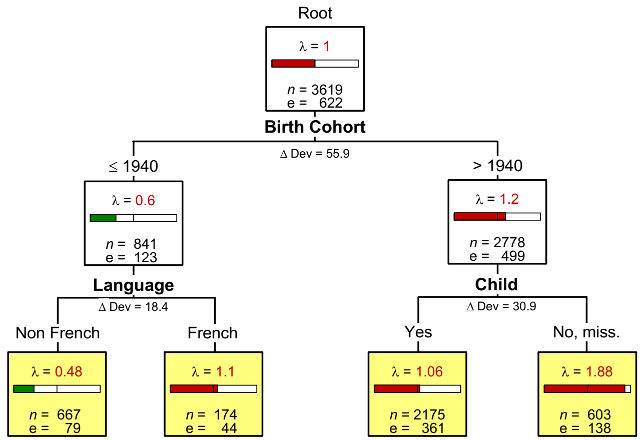
Marriage duration until divorce

Hazard model

- Discrete time model (logistic regression on person-year data)
- $\exp(B)$ gives the Odds Ratio, i.e. change in the odd $h/(1-h)$ when covariate increases by 1 unit.

		exp(B)	Sig.
birthyr		1.0088	0.002
university		1.22	0.043
child		0.73	0.000
language	unknwn	1.47	0.000
	French	1.26	0.007
	German	1	ref
	Italian	0.89	0.537
Constant		0.0000000004	0.000

Divorce, Switzerland, Relative risk



Hazard model with interaction

- Adding interaction effects detected with the tree approach
- improves significantly the fit (sig $\Delta\chi^2 = 0.004$)

		exp(B)	Sig.
born after 1940		1.78	0.000
university		1.22	0.049
child		0.94	0.619
language	unknwn	1.50	0.000
	French	1.12	0.282
	German	1	ref
	Italian	0.92	0.677
b_before_40*French		1.46	0.028
b_after_40*child		0.68	0.010
Constant		0.008	0.000

Illustrative mvad data set

- McVicar and Anyadike-Danes (2002)'s study of transition from school to employment in North Ireland.
  - Survey of 712 Irish youngsters.
  - Sequences describe their follow-up during the 6 years after the end of compulsory school (16 years old) and are formed by 70 successive monthly observed states between September 1993 and June 1999.
  - Sates are:

EM	Empoyement
FE	Further education
HE	Higher education
JL	Joblessness
SC	School
TR	Training.

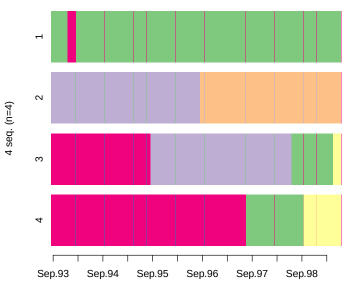
Sate sequences - mvad data set

- First sequences (first 20 months)

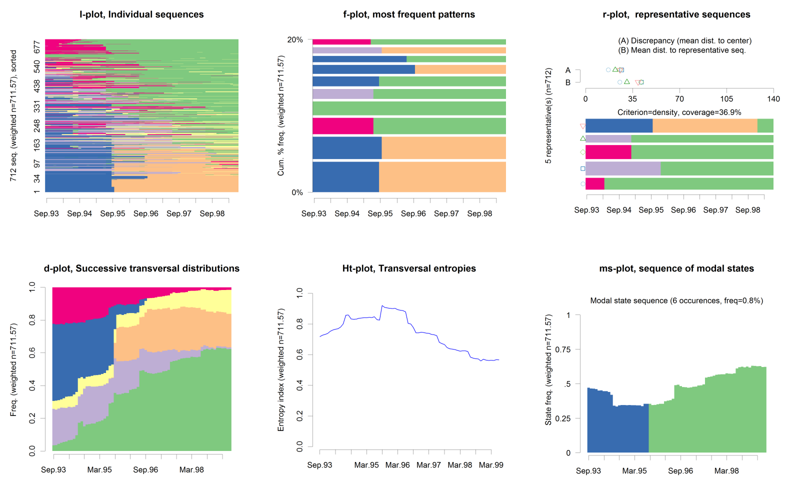
```
  Sequence
1 EM-EM-EM-EM-TR-TR-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM-EM
2 FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE-FE
3 TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR
4 TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR-TR
```

- compact representation (SPS format)

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



State sequences: Graphical display

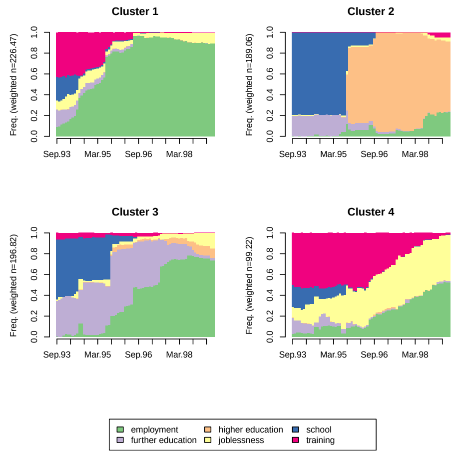


Pairwise dissimilarities and cluster analysis

- Different metrics permit to compute pairwise dissimilarities between sequences
  - of which optimal matching (Abbott and Forrest, 1986) is perhaps the most popular in social sciences
- Once you have pairwise dissimilarities, you can do
  - cluster analysis of sequences
  - principal coordinate analysis
  - measure the discrepancy between sequences
  - Find representative sequences, either most central or with highest density neighborhood (Gabadinho et al., 2011b)
  - ANOVA-like analysis and Regression trees (Studer et al., 2011)

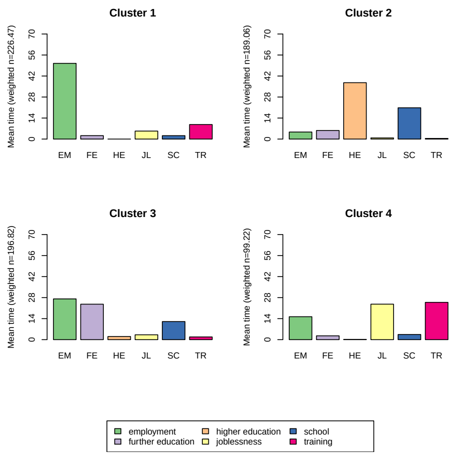
Cluster analysis: Outcome

- Rendering the cluster contents: transversal state distributions

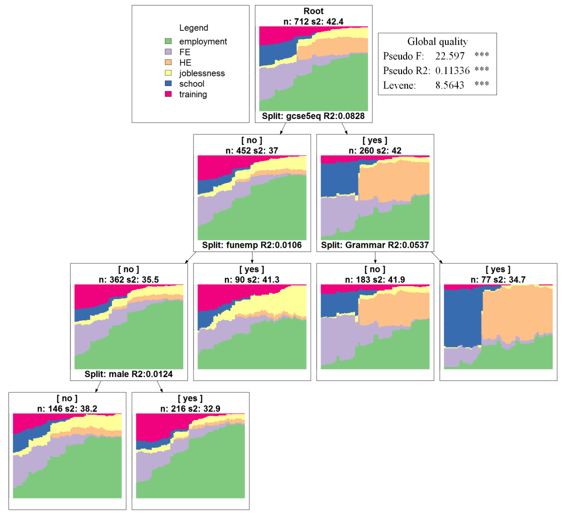


Cluster analysis: Outcome (2)

- Mean time per state by cluster



Regression tree



Markov process: Principle

(Brémaud, 1999; Berchtold and Raftery, 2002)

- Assume we have a sequence of states (not necessarily panel data)
- How is state in position t related to previous states?
- What is the probability to switch to state B in t when we are in state A in $t-1$?
  - Probability to fall next year into joblessness when we have a partial time job.
  - Probability to stay unemployed next t when we are currently unemployed.
  - Probability to recover from illness next month.

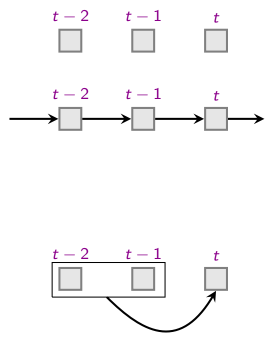
Homogenous Markov process: Assumptions

- transition probability is the same whatever t (homogeneity)
- a few lagged states summarize all the sequence before t
- 1st order: state in $t-1$ summarizes all the sequence before t; i.e.; state in t depends only on state in $t-1$
- 2nd order: states in $t-1$ and $t-2$ summarize all the sequence before t; i.e.; state in t depends only on states in $t-1$ and $t-2$
- ...

Markov process: Illustration

- Blossfeld and Rohwer (2002) sample of 600 job episodes extracted from the German Life History Study
- Job episodes partitioned into 3 job length categories
  - short (1) = ≤ 3 years
  - medium (2) = $(3; 10]$ years
  - long (3) = > 10 years
- Data reorganized into 162 sequences of 2 to 9 job episodes (units with single episode not considered)
- How does present episode length depend upon those of preceding jobs?

Markov matrices of order 0, 1 and 2

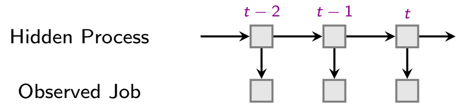


		job length at t 1	2	3	half conf. interval
Indep		.50	.35	.15	.07
	$t-1$				
	1	.57	.30	.13	.10
	2	.43	.42	.15	.13
	3	.20	.53	.27	.29
$t-2$	$t-1$				
1	1	.55	.30	.15	.11
2	1	.60	.30	.10	.20
3	1	1	0	0	.65
1	2	.37	.45	.18	.18
2	2	.50	.41	.09	.20
3	2	.45	.33	.22	.38
1	3	.33	.17	.50	.46
2	3	0	.87	.13	.40
3	3	1	0	0	1

Main findings

- First order:
  - Probability to start short job (1) after a short one (1) is much higher than starting a medium (2) or long job (3)
  - not the case after a medium or long job
- Second order:
  - No clear evidence about impact of lag 2 job
  - Main difference concerns long job (3) (but not significant)
  - Confirmed by MTD model, which gives weight 0 to second lag

Two state hidden Markov model



Hidden state at t / $t-1$	1	2	half conf. interval
1	.78	.22	.12
2	.53	.47	.19
initial	.56	.44	.11

Hidden state	Job length 1	2	3	half conf. interval
1	.75	.23	.02	.12
2	.05	.58	.37	.18

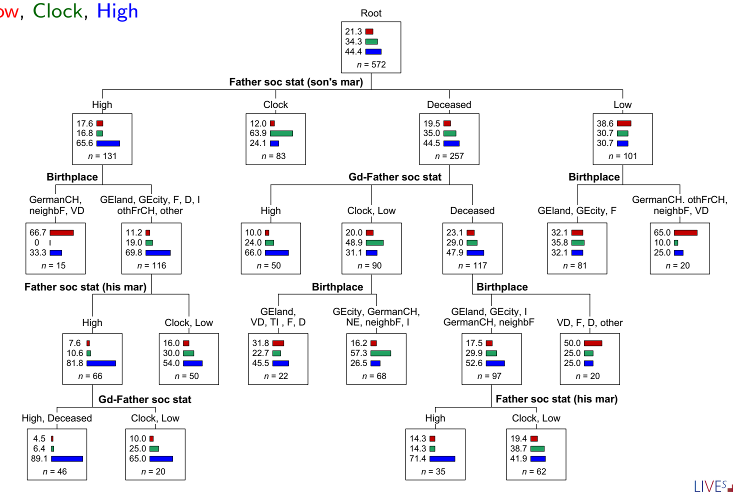
Hidden Markov Model (HMM)

- Relaxing homogeneity assumption with HMM
- Fitting a HMM with 2 hidden states
  - distribution of initial state of hidden variable
  - transition matrix of hidden process
  - distribution of transitions to the job length categories associated to each hidden state

Mobility tree

Social transition tree with birth place covariate (Ritschard and Oris, 2005)

Low, Clock, High



Conclusion

- Now, it is your turn!
- To chose a method, you first have to
  - Clarify what you are looking for
    - typical patterns, departures from standards, ...
    - specific transitions or holistic view
    - relationships with context (covariates)
    - ...
  - Identify the nature of your data
    - Categorical vs numerical
    - Direct or indirect measures of variable of interest
    - Long or short sequences
    - ...

Thank You!

References I

Abbott, A. and J. Forrest (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History 16*, 471–494.

Berchtold, A. and A. E. Raftery (2002). The mixture transition distribution model for high-order Markov chains and non-gaussian time series. *Statistical Science 17*(3), 328–356.

Blossfeld, H.-P. and G. Rohwer (2002). *Techniques of Event History Modeling, New Approaches to Causal Analysis* (2nd ed.). Mahwah NJ: Lawrence Erlbaum.

Brémaud, P. (1999). *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues*. New york: Springer Verlag.

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

References II

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2011b). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, and J. Filipe (Eds.), *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Volume 128 of *Communications in Computer and Information Science (CCIS)*, pp. 94–106. Springer-Verlag.

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.

Mills, M. (2011). *Introducing Survival and Event HistoryAnalysis*. London: Sage. (Chap. 11 about Sequential analysis and TraMineR).

Ritschard, G., A. Gabadinho, N. S. Müller, and M. Studer (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management 1*(1), 68–90.

Ritschard, G. and M. Oris (2005). Life course data in demography and social sciences: Statistical and data mining approaches. In R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, and E. Widmer (Eds.), *Towards an Interdisciplinary Perspective on the Life Course*, Advances in Life Course Research, Vol. 10, pp. 289–320. Amsterdam: Elsevier.

References III

Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research*. In press.