Study of the effect of the university degree on the work path by a latent Markov model

Francesco Bartolucci
Department of Economics
University of Perugia
bart@stat.unipg.it

Fulvia Pennoni
University of Milano-Bicocca
fulvia.pennoni@unimib.it

Giorgio Vittadini
University of Milano-Bicocca
giorgio.vittadini@unimib.it

Bolzano, March 25, 2014

Table of contents

Outline

- A model is proposed to investigate the relative *effectiveness* of certain degree programs on the work path after graduation
- the approach is motivated by the availability of a *dataset about graduates* from some universities in Milan for whom some outcomes are repeatedly observed
- the type of degree program may be considered as a *treatment* which is not controlled for (self assigned)
- the outcomes should be able to describe how much the university degrees improve *personal opportunities* in the labor market
- the treatment groups *differ* prior to treatment in a way that can influence the outcomes and therefore bias may arise if standard statistical tools are used for data analysis

- We consider the work path of a graduate as the manifestation of his/her *Human Capital* (HC) development

- HC is assumed to affect (and then it measured by) different response variables; in the application we consider *3 ordinal response variables* observed for 4 quarters after graduation:
  - (*i*) *employment contract type* with categories: none (for unemployed), temporary, and permanent
  - (*ii*) *employment skill* with categories: none, low/medium, and high
  - (*iii*) *gross income* in Euros with categories: none, $\leq$3,750, and $>$3,750 (we consider the threshold of €15,000 yearly)

- the model we formulate takes into account:
  - the latent nature of HC (the evolution of which is represented by a sequences of *latent variables* or latent process)
  - the *multivariate and longitudinal* nature of the response variables
  - the need to *balance among groups* according to background covariates

Causal latent Markov (LM) model

- Basic notation:
  - n: *number of subjects* (e.g., graduates)
  - T: *number of periods* of observation (e.g., 4 quarters after graduation)
  - r: *number of response variables* (e.g., 3 corresponding to contract type, skill level, gross income)
  - Y_{ijt}: *response variable* of type j for subject i at occasion t (with c_j categories)
  - $\mathbf{Y}_{it} = (Y_{i1t}, \ldots, Y_{irt})$: *response vector* for unit i at occasion t
  - $\mathbf{X}_i$: column vector of the *pretreatment covariates* for subject i
  - Z_i: *treatment indicator variable* with I levels (e.g., 1 technical degree, 2 architecture, 3 business, 4 humanistic, 5 scientific)

Model formulation

- The key point of the proposed model is representing the HC level at the different occasions by a *sequence of discrete latent variables*:
$$\mathbf{H}_i = (H_{i1}, \ldots, H_{iT})$$
where each H_{it} has support $\{1, \ldots, k\}$
- to give a *causal interpretation* of the proposed model we also consider
$$\mathbf{H}_i^{(z)} = (H_{i1}^{(z)}, \ldots, H_{iT}^{(z)})$$
with $H_{it}^{(z)}$ denoting the HC level for subject i at occasion t if he/she has received treatment (e.g., degree) of type z
- each $H_{it}^{(z)}$ is a sort of *potential outcome* in the terminology of Rubin (1974, 2005), the main difference is that it is a potential version of a variable that is not directly observable even for the selected treatment

Assumptions

A1: every *latent process* $\mathbf{H}_i^{(z)}$ is assumed to follow a *first-order homogeneous Markov chain* with state space $\{1, \dots, k\}$, initial probabilities $p(H_{i1}^{(z)} = h)$, and transition probabilities $p(H_{it}^{(z)} = h | H_{i,t-1}^{(z)} = \bar{h})$ suitably parametrized

- the *initial probabilities* are parametrized as:

$$\log \frac{p(H_{i1}^{(z)} = h)}{p(H_{i1}^{(z)} = 1)} = \alpha_h + I(z > 1)\beta_{hz}, \quad h = 2, \dots, k$$

- $I(\cdot)$: *indicator function*
- α_h: intercept (*effect of the first treatment*)
- β_{hz} is *average treatment effect (ATE)* of the z-th treatment $(z > 1)$ with respect to the first treatment on the logit scale
- to make *comparisons between treatments* we can also consider the difference $p(H_{i1}^{(z)} = h) - p(H_{i1}^{(1)} = h)$

◇ the *transition probabilities* are modeled as:

$$\log \frac{p(H_{it}^{(z)} = h | H_{i,t-1}^{(z)} = \bar{h})}{p(H_{it}^{(z)} = 1 | H_{i,t-1}^{(z)} = \bar{h})} = \gamma_{\bar{h}h} + I(z > 1)\delta_{hz},$$

for $\bar{h}, h = 1, \ldots, k,\ h \neq \bar{h},\ t = 2, \ldots, T$

- $\gamma_{\bar{h}h}$: intercept (effect of the *first treatment* on the transition probabilities)
- δ_{hz}: *differential ATE* for treatment z ($z > 1$) referred to the transition from level $\bar{h}$ to level h of HC measured on the logit scale
- a differential ATE can also be directly measured on the *probability scale*, that is, as difference between transition probabilities

A2: *consistency*: $H_{it} = H_{it}^{(z_i)}$ where z_i is the observed treatment of subject i, implying that

$$p(H_{it} = H_{it}^{(z_i)} | Z_i = z_i) = 1$$

for $i = 1, \ldots, n,\ t = 1, \ldots, T$

A3: *positivity*: $0 < p(Z_i = z | \mathbf{X}_i = \mathbf{x}_i) < 1$ for $z = 1, \ldots, l$ and any possible configuration $\mathbf{x}_i$ of the pretreatment covariates

A4: *absence of unobserved confounding*: $Z_i \perp\!\!\!\perp \mathbf{H}_i^{(z)} | \mathbf{X}_i$, $z = 1, \ldots, l$

A5: *local independence*: every response variable Y_{ijt} is conditionally independent of any other variable given H_{it}, for $i = 1, \ldots, n$, $j = 1, \ldots, r$, and $t = 1, \ldots, T$

- the *adopted parameters* are:

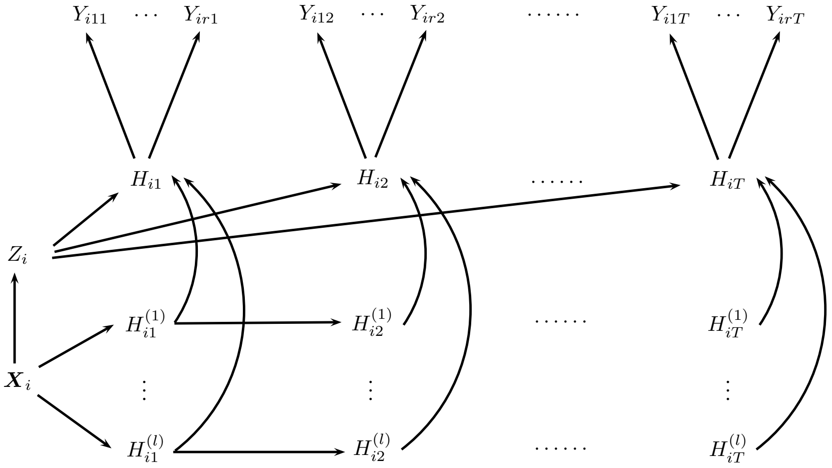
$$\phi_{jy|h} = p(Y_{ijt} = y|H_{it} = h),$$

for $h = 1, \ldots, k$, $i = 1, \ldots, n$, $j = 1, \ldots, r$, $t = 1, \ldots, T$, and $y = 0, \ldots, c_j - 1$; therefore:

$$p(\mathbf{Y}_{it} = \mathbf{y}|H_{it} = h) = \prod_{j=1}^{r} \phi_{jy_j|h},$$

where $\mathbf{y} = (y_1, \ldots, y_r)$ is a generic configuration of $\mathbf{Y}_{it}$

Path diagram for the proposed causal LM model



Two-step maximum likelihood estimation

- In order to *estimate the proposed causal model*, we follow a two-step approach similar to that proposed by Lanza *et al.* (2013) for a causal latent class model
- at the *first step* we estimate a weight for each subject depending on the pretreatment covariates according to the propensity score approach (Rosenbaum and Rubin, 1983)
- at the *second step* we maximize a weighted log-likelihood of the LM model with weights computed at the first step
- *propensity score* is the probability that a unit takes a certain treatment given the pretreatment covariates (standard tool of causal inference with non-experimental studies)

- The *first step* consists of estimating a multinomial logit model based on the following parametrization:

$$\log \frac{p(Z_i = z|\mathbf{X}_i = \mathbf{x}_i)}{p(Z_i = 1|\mathbf{X}_i = \mathbf{x}_i)} = \eta_z + \mathbf{x}_i'\boldsymbol{\lambda}_z, \quad z = 2, \dots, l$$

- the *individual weights* are then computed as

$$\hat{w}_i = n \frac{1/\hat{p}(Z_i = z_i|\mathbf{X}_i = \mathbf{x}_i)}{\sum_{m=1}^n 1/\hat{p}(Z_m = z_m|\mathbf{X}_i = \mathbf{x}_i)}, \quad i = 1, \dots, n$$

- at the *second step* we maximize the weighted log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \hat{w}_i \log p(\mathbf{Y}_i^{(1)} = \mathbf{y}_i^{(1)}, \dots, \mathbf{Y}_i^{(T)} = \mathbf{y}_i^{(T)}|Z_i = z_i),$$

  - $\boldsymbol{\theta}$: vector of all LM model parameters arranged in a suitable way

- The weighted log-likelihood is maximized with respect to $\boldsymbol{\theta}$ by the *EM algorithm* (Baum *et al.*, 1970, Dempster *et al.*, 1977)
- it is based on the *complete data log-likelihood* (that we could compute if we knew the latent process of each subject):

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) &= \sum_{h=1}^{k}\sum_{j=1}^{r}\sum_{t=1}^{T}\sum_{y=0}^{c_j-1} a_{hjty}\log\phi_{jy|h} + \sum_{h=1}^{k}\sum_{i=1}^{n}\hat{w}_i b_{hi1}\log p(H_{i1}=h|Z_i=z_i)\\ &+ \sum_{\bar{h}=1}^{k}\sum_{h=1}^{k}\sum_{i=1}^{n}\sum_{t=2}^{T}\hat{w}_i b_{\bar{h}hit}\log p(H_{it}=h|H_{i,t-1}=\bar{h},Z_i=z_i)\end{aligned}$$

  - a_{juty}: weighted frequency of subjects responding by y to the j-th response variable and belonging to latent state h at occasion t
  - b_{hit}: indicator variable equal to 1 if subject i belongs to latent class h at occasion t
  - $b_{\bar{h}hit} = b_{\bar{h}i,t-1}b_{hit}$: indicator variable equal to 1 if the same subject moves from state $\bar{h}$ to state h at occasion t

- The EM algorithm consists of alternating two steps until convergence in $\ell(\boldsymbol{\theta})$:
  - E-step: compute the *expected value* of the complete data log-likelihood given the current $\boldsymbol{\theta}$ and the observed data
  - M-step: *maximize* this expected value with respect to $\boldsymbol{\theta}$

- a *nonparametric bootstrap* procedure (Davison and Hynkley, 1997) is used to obtain standard errors for the estimates: we repeatedly resample from the observed sample and compute the maximum likelihood estimates for every bootstrap sample

- the number of latent classes (k) is *selected* according to Bayesian Information Criteria (BIC) (Schwarz, 1978):

$$BIC = -2\,\hat{\ell} + \log(n)\#\mathrm{par}$$

  - ◇ $\hat{\ell}$: maximum of the log-likelihood
  - ◇ $\#\mathrm{par}$: number of free parameters

Application

- The application is based on data coming from the integration of certain *administrative archives*:[1]
  - archive of the federal observatory of the labour market in Lombardy concerning the compulsory communications given by employers from 2000 to nowadays, regarding activation and termination of the employment relationship
  - archive of the graduates of four universities in Milan, concerning the academic performance for all students earning a university degree between 2003 and 2008
  - archive of the Italian office of revenues relative to the annual gross earned income of all residents declaring income in Lombardy (available years: 2007-2008 for residents in Milan)
  - archive of the Milan's City Hall recording annually the personal information about citizens

[1] held by the Interuniversity Research Centre `http://www.crisp-org.it/`

The data

- We choose the *graduates in 2007 from four different universities* with five years of university education: (pre-reform and post-reform only laurea magistralis)
- we *exclude* faculties such as Law and Health often characterized by institutionalized stages for advancement in the associated professional careers
- from the merging of the archives we obtain a *dataset concerning 1,624 graduates* resident in the area surrounding Milan who declared income in 2009: they have been followed along four quarters after the graduation, covering one year

Descriptive statistics

- *Percentage of graduates* for each type of treatment

Treatment		(%)
Degree type	technical	22.78
	architecture	8.93
	business	13.85
	humanistic	41.26
	science	13.18

- the available *pretreatment covariates* are: *gender*, *district of birth*, *family income*, *number of family members*, *final grade at high school*, *type of high school*, *year of high school diploma*

Covariate		University degree				
		techn.	arch.	econ.	human.	scien.
gender:	male	0.802	0.483	0.547	0.202	0.577
	female	0.198	0.517	0.453	0.798	0.423
district of birth:	Milan	0.729	0.727	0.789	0.778	0.859
	Lombardy	0.073	0.093	0.063	0.044	0.028
	Italy	0.175	0.102	0.135	0.149	0.085
	others	0.023	0.078	0.013	0.029	0.028
income of the family:		67.316	62.236	68.374	56.584	55.563
number of family members:	1	0.182	0.239	0.161	0.178	0.099
	2	0.125	0.166	0.161	0.137	0.155
	3	0.350	0.254	0.314	0.360	0.451
	4	0.264	0.288	0.247	0.263	0.254
	≥ 5	0.079	0.054	0.117	0.061	0.042
final score high school:		87.479	77.649	77.588	79.908	82.183
type of high school:	lyceum	0.878	0.795	0.700	0.827	0.901
	others	0.122	0.205	0.300	0.173	0.099
year of high school diploma:	1999	0.294	0.337	0.368	0.444	0.282
	2000	0.106	0.220	0.363	0.292	0.197
	2001	0.383	0.390	0.224	0.193	0.338
	2002	0.218	0.054	0.045	0.070	0.183

- *Frequency distribution* of the response variables

	Quarter (t)			
Contract type	1st	2nd	3rd	4th
none	61.58	53.51	50.68	47.23
temporary	26.72	31.53	31.83	33.44
permanent	11.70	14.96	17.49	19.33

	Quarter (t)			
Skill	1st	2nd	3rd	4th
none	61.58	53.51	50.68	47.23
medium/low	14.72	15.21	16.07	17.00
high	23.72	31.28	33.25	35.78

	Quarter (t)			
Gross Income	1st	2nd	3rd	4th
none	59.73	50.62	47.35	44.95
≤ 3750	31.28	29.74	27.34	25.31
> 3750	8.99	19.64	25.31	29.74

First step (propensity score)

- A weight is associated to each student which is computed by fitting a *multinomial logit model* based on a suitable selected set of pretreatment covariates

		Degree (vs technical)			
		arch.	econ.	human.	scien
intercept		5.677**	6.061**	4.932**	2.074*
final score:		-0.083**	-0.086**	-0.076**	-0.044**
gender:	female	1.878**	1.628**	3.156**	1.323**
type of high school:	others	0.799**	1.497**	0.694**	-0.091
district of birth:	Lombardy	0.139	-0.338	-0.762†	-1.182
	Italy	-0.507†	-0.311	-0.297	-0.867†
	others	0.853	-1.299†	-0.130	0.007

(†significant at 10%, *significant at 5%, **significant at 1%)

- we checked that there is a *much higher balance* between groups corresponding to different treatment

Second step (maximum likelihood estimation)

- According to the BIC index a suitable model for the data is based on $k = 4$ *latent states*
- the *maximum log-likelihood* of the model is equal to $\hat{\ell}$ =-5215.625 with 42 parameters; the corresponding value of BIC is 10874.91
- *interpretation of the latent classes* is based on the conditional response probabilities ($\phi_{jy|h}$), whereas estimates of the parameters β_{hz} and δ_{hz} measure the *causal effects* of the different treatments

- Estimates of the *conditional response probabilities*:

Contract type ($j = 1$)	Latent class (h) 1	2	3	4
none	1.000	0.000	0.000	0.000
temporary	0.000	0.995	0.671	0.000
permanent	0.000	0.005	0.329	1.000

Skill ($j = 2$)	Latent class (h) 1	2	3	4
none	1.000	0.000	0.000	0.000
low/medium	0.000	0.040	1.000	0.000
high	0.000	0.996	0.000	1.000

Gross income ($j = 3$)	Latent class (h) 1	2	3	4
none	1.000	0.032	0.021	0.013
$\leq$ 3750	0.000	0.652	0.563	0.265
$>$3750	0.000	0.316	0.416	0.721

- According to the estimated conditional probabilities the *latent states are characterized as*:
  1. unemployed subjects who may have income from other sources (*lowest HC level*)
  2. employed graduates for whom there is prevalence of high skill
  3. graduates with less level of skill compared with the previous class but with more stable contracts and higher income levels
  4. graduates with high quality jobs and more appropriate income to the job qualification (*highest HC level*)

- Estimates of the parameters affecting the *initial probabilities* (ATE on the logit scale):

	Latent class (h)		
Treatment	2	3	4
technical ($\hat{\alpha}_h$)	-0.586**	-0.988**	-1.256**
architecture vs. technical ($\hat{\beta}_{h2}$)	-1.253**	-1.092*	-1.635
economic vs. technical ($\hat{\beta}_{h3}$)	-0.216	0.292	-0.110
humanistic vs. technical ($\hat{\beta}_{h4}$)	-0.766**	-0.584†	-2.107**
scientific vs. technical ($\hat{\beta}_{h5}$)	-0.975	-0.415	-1.153
economic vs. architecture ($\hat{\beta}_{h3} - \hat{\beta}_{h2}$)	1.036**	1.384**	1.525
humanistic vs. architecture ($\hat{\beta}_{h4} - \hat{\beta}_{h2}$)	0.486	0.508†	-0.472
scientific vs. architecture ($\hat{\beta}_{h5} - \hat{\beta}_{h2}$)	0.278	0.677	0.482
humanistic vs. economic ($\hat{\beta}_{h4} - \hat{\beta}_{h3}$)	-0.550†	-0.876**	-1.997**
scientific vs. economic ($\hat{\beta}_{h5} - \hat{\beta}_{h3}$)	-0.758	-0.707†	-1.043
scientific vs. humanistic ($\hat{\beta}_{h6} - \hat{\beta}_{h3}$)	-0.209	0.168	0.954

- *Estimated initial probabilities* of the hidden Markov chain:

	Latent class (h)			
Treatment	1	2	3	4
technical	0.452	0.251	0.168	0.129
architecture	0.747	0.119	0.093	0.041
economic	0.454	0.203	0.226	0.116
humanistic	0.666	0.172	0.138	0.023
scientific	0.647	0.136	0.159	0.058

- At the beginning of the period, there is a statistical significant difference of technical and economic degrees in terms of effect on HC with respect to architecture and humanistic degrees
- significant differences are not observed between technical and economic degrees and between architecture and humanistic degrees, whereas scientific degrees seem to be a compromise
- based on both the first and the last class, we obtain the same raking of the degree types in terms of effect on the initial HC level

- Estimates of the parameters affecting the *transition probabilities* (ATE on the logit scale):

Treatment	Latent class (h) 2	3	4
technical $\bar{h} = 1$ ($\hat{\gamma}_{1h}$)	-1.640**	-2.257**	-2.343**
technical $\bar{h} = 2$ ($\hat{\gamma}_{2h}$)	2.163**	-1.079*	-0.246
technical $\bar{h} = 3$ ($\hat{\gamma}_{3h}$)	-0.156	2.855**	-0.015
technical $\bar{h} = 2$ ($\hat{\gamma}_{4h}$)	-0.408	-0.494	5.376**
architecture vs. technical ($\hat{\delta}_{h2}$)	-0.826**	-1.015**	-2.538**
economic vs. technical ($\hat{\delta}_{h3}$)	-0.623*	-0.504	-1.308**
humanistic vs. technical ($\hat{\delta}_{h4}$)	-0.490*	-0.928**	-1.496**
scientific vs. technical ($\hat{\delta}_{h5}$)	-0.743*	-0.901*	-1.077†
economic vs. architecture ($\hat{\delta}_{h3} - \hat{\delta}_{h2}$)	0.203	0.511†	1.230*
humanistic vs. architecture ($\hat{\delta}_{h4} - \hat{\delta}_{h2}$)	0.337	0.086	1.042*
scientific vs. architecture ($\hat{\delta}_{h5} - \hat{\delta}_{h2}$)	0.083	0.113	1.461*
humanistic vs. economic ($\hat{\delta}_{h4} - \hat{\delta}_{h3}$)	0.134	-0.425	-0.188
scientific vs. economic ($\hat{\delta}_{h5} - \hat{\delta}_{h3}$)	-0.120	-0.398	0.231
scientific vs. humanistic ($\hat{\delta}_{h5} - \hat{\delta}_{h4}$)	-0.120	-0.398	0.231

- *Estimated transition probabilities* of the hidden Markov chain:

		Latent class (h)			
Degree	$\bar{h}$	1	2	3	4
technical	1	0.717	0.139	0.075	0.069
	2	0.092	0.804	0.031	0.072
	3	0.049	0.042	0.859	0.049
	4	0.005	0.003	0.003	0.990
architecture	1	0.885	0.075	0.034	0.007
	2	0.200	0.763	0.025	0.012
	3	0.129	0.048	0.813	0.010
	4	0.054	0.016	0.012	0.919
economic	1	0.838	0.087	0.053	0.021
	2	0.164	0.767	0.034	0.035
	3	0.082	0.037	0.859	0.023
	4	0.017	0.006	0.005	0.971
humanistic	1	0.864	0.101	0.035	0.018
	2	0.151	0.803	0.020	0.026
	3	0.116	0.061	0.797	0.026
	4	0.020	0.008	0.005	0.967
scientific	1	0.857	0.079	0.036	0.028
	2	0.180	0.746	0.025	0.048
	3	0.114	0.046	0.802	0.038
	4	0.013	0.004	0.003	0.979

- There are significant differences between technical degrees and all the other types of degree during the period of observation in favor of the first (evolution of HC)
- Statistically significant differences are also observed between the degree in Architecture and the other degrees, with the first having a worse impact on the evolution of the HC level
- Economic, humanistic, and scientific degrees give rise to a group having an intermediate impact between technical degrees and Architecture, with economic degrees performing slightly better than the other two in the same group
- All transition probability matrices are characterized by a rather high persistence, with elements in the main diagonal always greater than 0.7, most of which are also greater than 0.8
- For technical degrees there is the lowest probability of remaining in the first latent class (0.72) and the highest probability of remaining in the last class (0.99)

Conclusions

- In terms of causal effects of the university degrees, a clear *ranking* results:

    1. *Technical degrees*: highest effect at the beginning and in terms of evolution of HC level
    2. *Economic degrees*: impact close to technical degrees at the beginning and worse in terms of evolution
    3. *Scientific degrees*: intermediate impact both at the beginning and in the following
    4. *Humanistic degrees*: significantly worse impact with respect to technical and economic degrees
    5. *Architecture*: impact on HC level similar to humanistic degrees, but even worse

Main References

- Bartolucci, F. and Farcomeni, A. and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*, Chapman and Hall/CRC press.
- Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics*, **41**, 164-171.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, MA.
- Dempster A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion), *JRSS-B*, **39**, 1-38.
- Grilli, L. and Mealli, F. (2008). Nonparametric bounds on the causal effect of University studies on job opportunities using principal stratification. *Journal of Educational and Behavioral Statistics*, **3**, 361–383.
- Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics*, **54**, 3-56.

- Lanza, S. T. and Coffman, D. L. and Xu, S. (2013). Causal inference in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, **3**, 361–383.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41-55.
- Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, **66**, 688-701.
- Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions, *Journal of the American Statistical Association*, **100**, 322-331.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.
- Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for time series: an introduction using R*. Springer-Verlag, New York.