Estimating and correcting for measurement error using hidden Markov models

Paulina Pankowska
p.k.pankowska@uu.nl
Assistant Professor, Department of Sociology, Utrecht University

October 16, 2023

Utrecht University

Overview

1. **Measurement error**
2. **Hidden Markov models (HMMs)**
3. **HMMs in R**
4. **HMMs in Latent Gold**
5. **Concluding remarks**

Measurement error

Definition

- Occurs when the observed/measured value differs from the true value of a variable
- Can be random (by chance) or systematic (according to some pattern)
- In categorical data referred to as misclassification

Measurement error

Definition

- Occurs when the observed/measured value differs from the true value of a variable
- Can be random (by chance) or systematic (according to some pattern)
- In categorical data referred to as misclassification

Effects

- Can bias estimates
- Lead to incorrect research findings
- Usually, systematic error worse than random error

Sources of measurement error

Surveys

- Self-reported measures affected by cognitive processes
  - Recall bias
  - Social desirability bias
- Poor questions or questionnaire design
  - Misunderstanding, wrong interpretation
  - Respondent fatigue

Sources of measurement error

Surveys

- Self-reported measures affected by cognitive processes
  - Recall bias
  - Social desirability bias
- Poor questions or questionnaire design
  - Misunderstanding, wrong interpretation
  - Respondent fatigue

Digital trace data

- Incomplete data from tracking apps
  - Not installed on all devices
  - Temporarily switched off
- Data not representing individual usage
  - Multiple HH members using same device/accoung

Hidden Markov models (HMMs)

Background

- Latent class models used to correct for error in categorical, longitudinal data
- Do not require any data source to be error-free
- Use repeated measures of indicator(s) to extract information about the error from the data

Hidden Markov models (HMMs)

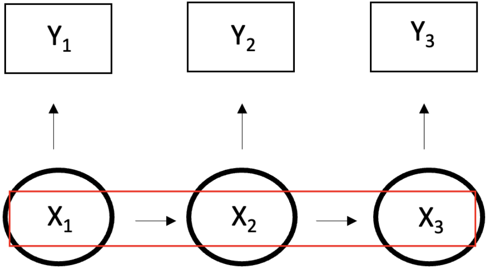
Background

- Latent class models used to correct for error in categorical, longitudinal data
- Do not require any data source to be error-free
- Use repeated measures of indicator(s) to extract information about the error from the data

The standard HMM

- **Markov assumption:** true state at time t only depends on true state at time t-1
- **Local independence assumption:** observed state at time t only depends on true state at time t

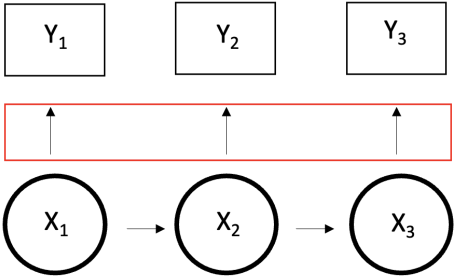
Standard HMM - Markov assumption



$$Pr(X_t = x_t | X_1 = x_1, ..., X_{t-1} = x_{t-1}) = Pr(X_t = x_t | X_{t-1} = x_{t-1})^1$$

[1] where $Pr(X_t = x_t)$ denotes the probability of the latent state X_t taking on a specific value x_t out of k possible categories.

Standard HMM - local independence assumption



$$Pr(A_1 = a_1, \ldots, A_T = a_T | X_1 = x_1, ..., X_T = x_T) = \prod_{t=1}^{T} Pr(A_t = a_t | X_t = x_t)^2$$

[2] where $Pr(A_1 = a_1, \ldots, A_T = a_T)$ denotes the probability of observing a specific sequence of states, where each state – A_1 , . . . , A_T – takes on a specific value – a_1 , . . . , a_T – out of k possible categories.

Standard HMM

Full HMM

Combining the Markov and local independence assumptions leads to the following probability of observing a certain path $A = (A_1, ..., A_T)$ in the data:

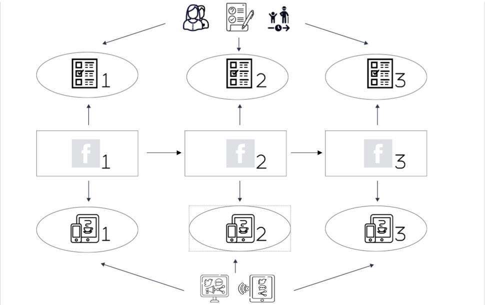
$$Pr(A = a) = \sum_{x_0=1}^{k} \cdots \sum_{x_t=1}^{k} Pr(X_0 = x_0) \prod_{t=1}^{T} Pr(X_t = x_t | X_{t-1} = x_{t-1})$$
$$\prod_{t=1}^{T} Pr(A_t = a_t | X_t = x_t)$$

- $Pr(X_0 = x_0)$ represents the initial state latent probabilities
- $Pr(X_t = x_t | X_{t-1} = x_{t-1})$ represents the latent transition probabilities, which follow a Markov process
- $Pr(A_t = a_t | X_t = x_t)$ denotes the classification error (or emission) probabilities, which satisfy the local independence assumption

Extended HMMs

Multiple indicator HMMs

- Correct for error in all sources simultaneously
- Allow to relax the local independence assumption
  - Model more realistic error scenarios, e.g., systematic error



HMMs in R

Available packages

- HMM (http://CRAN.R-project.org/package=HMM)
- hmm.discnp (http://CRAN.R-project.org/ package=hmm.discnp)
- msm (https://doi.org/10.18637/jss.v038.i08)
- depmixS4 (https://doi.org/10.18637/jss.v036.i07)
- mhsmm (https://doi.org/10.18637/jss.v039.i04)
- LMest (http://CRAN.R-project.org/package=LMest)
- seqHMM (https://doi.org/10.18637/jss.v088.i03)

HMMs in R

Package limitations

- Only standard HMMs are allowed (no covariates, one indicator)
- Only continuous time processes
- Must specify starting values
- Cannot handle missing values
- Cannot properly use a three-step approach

The seqHMM package - preparing data

Data

Frequency of FB use in 3 time points (3 survey waves and corresponding periods from tracking apps).

```
# Data survey
 head(data.fb.wide.survey)

new.id fb.survey1             fb.survey2             fb.survey3
1      1.several times a day  2.daily                1.several times a day
3      2.daily                2.daily                2.daily
4      2.daily                3.weekly               2.daily
5      5.less than monthly    5.less than monthly    5.less than monthly
10     1.several times a day  2.daily                1.several times a day
```

```
# Data tracking app
 head(data.fb.wide.trackig)

new.id fb.tracking1           fb.tracking2           fb.tracking3
1      1.several times a day  1.several times a day  1.several times a day
3      2.daily                2.daily                2.daily
4      5.less than monthly    5.less than monthly    5.less than monthly
5      4.monthly              3.weekly               4.monthly
10     2.daily                2.daily                3.weekly
```

Running HMMs with varying k

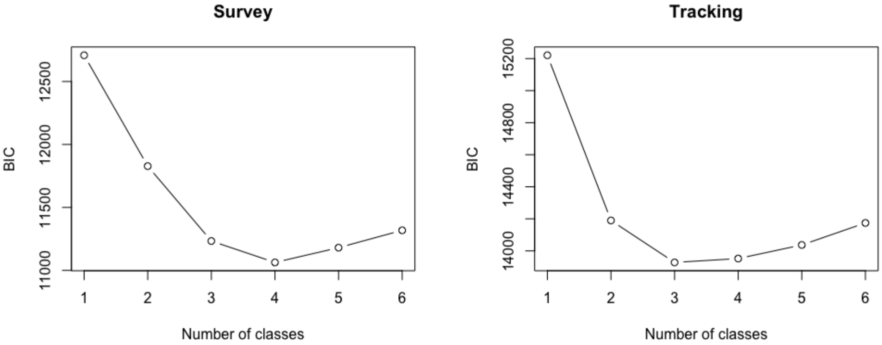
Model selection

Three criteria: model fit (BIC/AIC), class sizes, intepretative value (unless fixed cat.).

```
data.fb.seq.survey <- seqdef(data.fb.wide.survey[,2:4], start=1)
results <- NA
for(i in 2:7) {
    model.class <- build.hmm(observations=data.fb.seq.survey, n.states=i)
    model.fit <- fit.model(model.class,
                         control.em=list(restart=list(times = 5)))
    results <- append(results, BIC(model.fit$model))

plot(results[-1], type="b")
```

```
data.fb.seq.tracking <- seqdef(data.fb.wide.tracking[,2:4], start=1)
results.tracking <- NA
for(i in 2:7) {
    model.class.tracking <- build.hmm(observations=data.fb.seq.tracking, n.states=i)
    model.fit.tracking <- fit.model(model.class.tracking,
                         control.em=list(restart=list(times = 5)))
    results.tracking <- append(results.tracking, BIC(model.fit.tracking$model))

plot(results.tracking[-1], type="b")
```

Selecting HMM



Running selected HMM

```r
model.4class <- build.hmm(observations=data.fb.seq.survey, n.states=4)
model.4fit <- fit.model(model.4class,
                control.em=list(restart=list(times = 5)))
model.4fit
```

```r
model.4class.tracking <- build.hmm(observations=data.fb.seq.tracking,
                          n.states=4)
model.4fit.tracking <- fit.model(model.4class.tracking,
                         control.em=list(restart=list(times = 5)))
model.4fit.tracking
```

Results selected HMM - survey data

```
$model

Initial probabilities :
State 1   State 2 State 3 State 4
  0.323   0.207   0.164   0.306

Transition probabilities :
        State 1  State 2 State 3 State 4
State 1 0.95898  0.0000  0.0000  0.0410
State 2 0.00020  0.9134  0.0116  0.0748
State 3 0.00000  0.0000  1.0000  0.0000
State 4 0.03668  0.0374  0.0000  0.9259


Emission probabilities :
state names 1.several times a day 2.daily 3.weekly 4.monthly 5.less than monthly
State 1                   0.92096  0.0754  0.0036   0.00000             0.00000
State 2                   0.00656  0.0503  0.8063   0.13676             0.00000
State 3                   0.00000  0.0000  0.0235   0.43569             0.54079
State 4                   0.10774  0.8260  0.0607   0.00246             0.00308
```

Results selected HMM - survey data

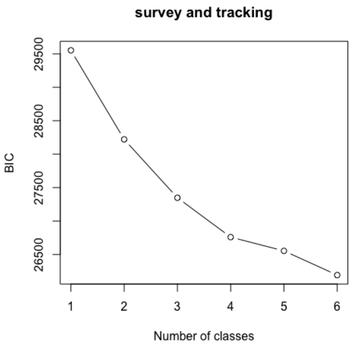
```
logLik
-5484.077

$em_results
$em_results$logLik
-5484.077

$em_results$iterations
490

$em_results$change
9.926655e-11

$em_results$best_opt_restart
-5484.077 -5484.077 -5484.077 -5484.077 -5484.077 -5484.077


$global_results
NULL

$local_results
NULL
$call
fit_model(model = model_4class, control_em = list(restart = list(times = 5)))
```

Results selected HMM - tracking data

```
$model

Initial probabilities :
State 1 State 2 State 3 State 4
  0.249   0.262   0.419   0.070



Transition probabilities :
        State 1 State 2   State 3 State 4
State 1 0.8930   0.0951   0.0000   0.0123
State 2 0.0352   0.9648   0.0000   0.0000
State 3 0.0114   0.0346   0.9360   0.0175
State 4 0.0000   0.0000   0.0300   0.9700



Emission probabilities :
state names 1.several times a day  2.daily 3.weekly 4.monthly  5.less than monthly
State 1                  0.03431  0.00000  0.5538   0.2400              0.17176
State 2                  0.90190  0.00000  0.0898   0.0000              0.00823
State 3                  0.00243  0.00291  0.0354   0.1010              0.85812
State 4                  0.00000  0.61530  0.3664   0.0115              0.00686
```

Running two-indicator HMM

```
data.fb.seq.combo <- list(data.fb.seq.survey, data.fb.seq.tracking)
results.combo <- NA
for(i in 2:7) {
    model.class.combo <- build.hmm(observations=data.fb.seq.combo, n.states=i)
    model.fit.combo <- fit.model(model.class.combo,
                            control.em=list(restart=list(times = 5)))
    results.combo <- append(results.combo, BIC(model.fit.combo$model))

plot(results.combo[-1], type="b")
```



Running two-indicator HMM

```
model.4class.combo <- build.hmm(observations=data.fb.seq.combo,
                                 n.states=4)
model.4fit.combo <- fit.model(model.4class.combo,
                               control.em=list(restart=list(times = 5)))
model.4fit.combo
```

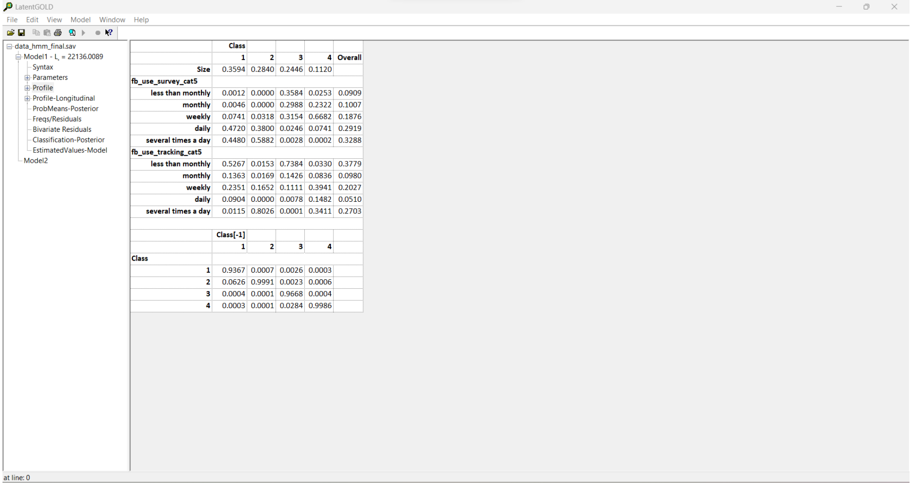
Results two-indicator HMM

```
$model

Initial probabilities :
State 1 State 2 State 3 State 4
  0.189   0.332   0.303   0.176

Transition probabilities :
        State 1   State 2  State 3  State 4
State 1 0.99356  0.00000  0.00644  0.0000
State 2 0.00438  0.90001  0.07024  0.0253
State 3 0.00000  0.00298  0.99702  0.0000
State 4 0.00000  0.00000  0.00000  1.0000

Emission probabilities :
Survey:
state names 1.several times a day 2.daily 3.weekly 4.monthly  5.less than monthly
State 1                    0.0000  0.0000  0.1020  0.4318            0.4662
State 2                    0.3560  0.3830  0.2530  0.0060            0.0030
State 3                    0.5040  0.3450  0.1250  0.0221            0.0035
State 4                    0.3250  0.3610  0.2630  0.0499            0.0014
Tracking:
state names 1.several times a day 2.daily 3.weekly 4.monthly  5.less than monthly
State 1                    0.0018  0.0270  0.1439  0.1389            0.6884
State 2                    0.0057  0.0013  0.0766  0.1310            0.7853
State 3                    0.8196  0.0000  0.1475  0.0173            0.0156
State 4                    0.0122  0.2399  0.5491  0.1423            0.0565
```

HMMs in Latent Gold

```
option
   algorithm
      tolerance=1e-007 emtolerance=0.01 emiterations=10000 nriterations=300;
   startvalues
      seed=0 sets=30 tolerance=1e-005 iterations=100;
   bayes
      categorical=1 variances=1 latent=1 poisson=1;
   missing  includeall;
   output
      parameters=effect  betaopts=wl standarderrors profile probmeans=posterior
      frequencies loadings bivariateresiduals classification estimatedvalues=model
      reorderclasses;
 outfile  'C:\Users\Paulina\Downloads\trace_out\trace_me_experiment_13Oct\fb_classfication4_extravar.sav'
      classification
      keep fb_donated_use, ed_secondary_cat, voc_training_cat,  migration.background, fb_devices, pol.int,left_right, pop_att, issue.assimilation, issue.climate,
      trust_ms_media, trust_politics, demcont, dembest;

variables
   caseid new_id;
   dependent fb_use_survey_cat5 nominal, fb_use_tracking_cat5 nominal;
   independent female nominal, age numeric, wave numeric, device nominal;
latent
      Class dynamic nominal 4;
equations
   Class[=0] <- 1;
   Class<- (~tra) 1 | Class[-1];
   fb_use_survey_cat5 <- 1 | Class +  female +  wave +  age;
   fb_use_tracking_cat5 <-  1 | Class + device ;
```

HMMs in Latent Gold



	Class				
	1	2	3	4	Overall
Size	0.3594	0.2840	0.2446	0.1120	
fb_use_survey_cat5					
less than monthly	0.0012	0.0000	0.3584	0.0253	0.0909
monthly	0.0046	0.0000	0.2988	0.2322	0.1007
weekly	0.0741	0.0318	0.3154	0.6682	0.1876
daily	0.4720	0.3800	0.0246	0.0741	0.2919
several times a day	0.4480	0.5882	0.0028	0.0002	0.3288
fb_use_tracking_cat5					
less than monthly	0.5267	0.0153	0.7384	0.0330	0.3779
monthly	0.1363	0.0169	0.1426	0.0836	0.0980
weekly	0.2351	0.1652	0.1111	0.3941	0.2027
daily	0.0904	0.0000	0.0078	0.1482	0.0510
several times a day	0.0115	0.8026	0.0001	0.3411	0.2703

	Class[-1]			
	1	2	3	4
Class				
1	0.9367	0.0007	0.0026	0.0003
2	0.0626	0.9991	0.0023	0.0006
3	0.0004	0.0001	0.9668	0.0004
4	0.0003	0.0001	0.0284	0.9986

HMMs in Latent Gold

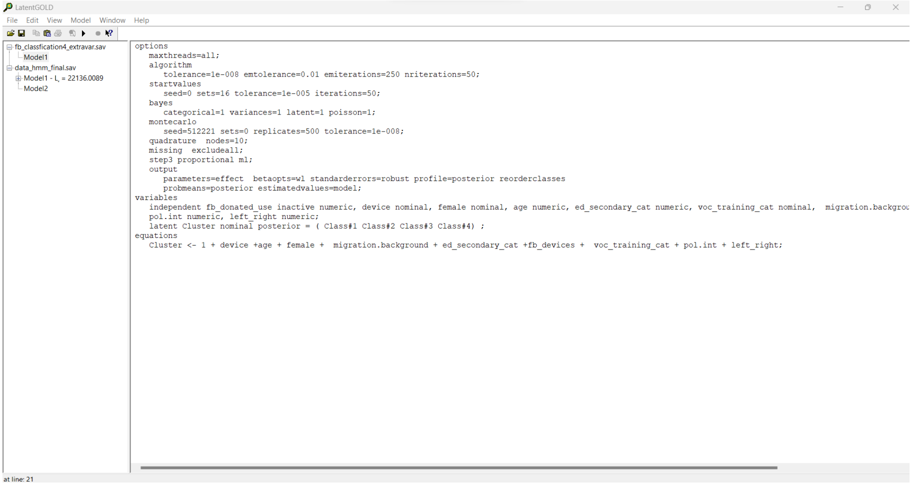
LatentGOLD

File Edit View Model Window Help

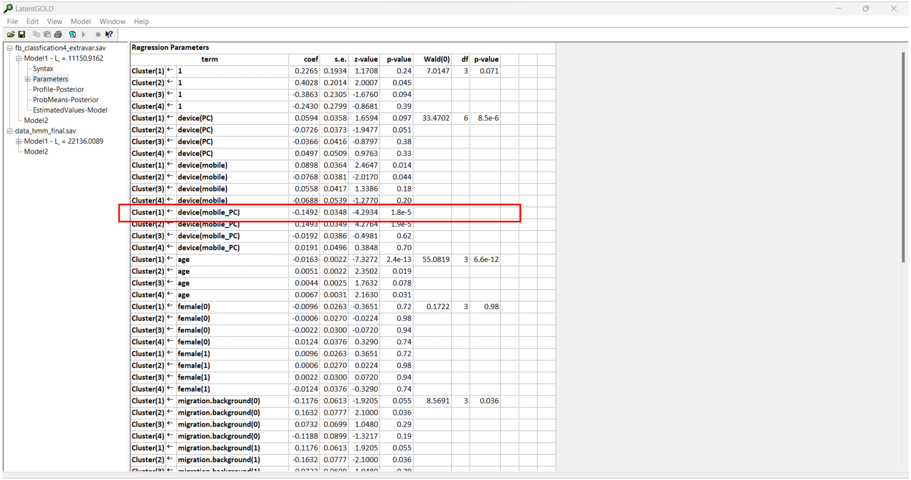
- data_hmm_final.sav
  - Model1 - L² = 22136.0089
    - Syntax
    - Parameters
    - Profile
    - Profile-Longitudinal
    - ProbMeans-Posterior
    - Freqs/Residuals
    - Bivariate Residuals
    - Classification-Posterior
    - EstimatedValues-Model

fb_use_tracking_cat5(less than monthly)	←	1	Class(1)	1.6797	0.0821	20.4606	4.8e-93	2636.9888	16	3.4e-555	735.6740	12	1.0e-149
fb_use_tracking_cat5(monthly)	←	1	Class(1)	0.3342	0.0851	3.9286	8.5e-5						
fb_use_tracking_cat5(weekly)	←	1	Class(1)	0.8114	0.0801	10.1321	4.0e-24						
fb_use_tracking_cat5(daily)	←	1	Class(1)	-0.4743	0.1090	-4.3511	1.4e-5						
fb_use_tracking_cat5(several times a day)	←	1	Class(1)	-2.3510	0.2628	-8.9456	3.7e-19						
fb_use_tracking_cat5(less than monthly)	←	1	Class(2)	0.1287	1.8409	0.0699	0.94						
fb_use_tracking_cat5(monthly)	←	1	Class(2)	0.3501	1.8230	0.1920	0.85						
fb_use_tracking_cat5(weekly)	←	1	Class(2)	2.6936	1.8137	1.4852	0.14						
fb_use_tracking_cat5(daily)	←	1	Class(2)	-7.4411	7.2402	-1.0277	0.30						
fb_use_tracking_cat5(several times a day)	←	1	Class(2)	4.2688	1.8138	2.3535	0.019						
fb_use_tracking_cat5(less than monthly)	←	1	Class(3)	3.7389	0.7702	4.8543	1.2e-6						
fb_use_tracking_cat5(monthly)	←	1	Class(3)	2.0556	0.7709	2.6665	0.0077						
fb_use_tracking_cat5(weekly)	←	1	Class(3)	1.6971	0.7705	2.2027	0.028						
fb_use_tracking_cat5(daily)	←	1	Class(3)	-1.3273	0.8373	-1.5851	0.11						
fb_use_tracking_cat5(several times a day)	←	1	Class(3)	-6.1644	3.0527	-2.0193	0.043						
fb_use_tracking_cat5(less than monthly)	←	1	Class(4)	-1.5291	0.3574	-4.2789	1.9e-5						
fb_use_tracking_cat5(monthly)	←	1	Class(4)	-0.4858	0.1548	-3.1380	0.0017						
fb_use_tracking_cat5(weekly)	←	1	Class(4)	1.1164	0.1129	9.8888	4.7e-23						
fb_use_tracking_cat5(daily)	←	1	Class(4)	-0.0370	0.1547	-0.2388	0.81						
fb_use_tracking_cat5(several times a day)	←	1	Class(4)	0.9355	0.1463	6.3928	1.6e-10						
fb_use_tracking_cat5(less than monthly)	←	device(PC)		0.6813	0.0743	9.1689	4.8e-20	190.7827	8	5.6e-37			
fb_use_tracking_cat5(monthly)	←	device(PC)		0.3958	0.0703	5.6336	1.8e-8						
fb_use_tracking_cat5(weekly)	←	device(PC)		0.0250	0.0506	0.4944	0.62						
fb_use_tracking_cat5(daily)	←	device(PC)		-0.9648	0.1197	-8.0608	7.6e-16						
fb_use_tracking_cat5(several times a day)	←	device(PC)		-0.1373	0.1044	-1.3156	0.19						
fb_use_tracking_cat5(less than monthly)	←	device(mobile)		0.0173	0.0608	0.2845	0.78						
fb_use_tracking_cat5(monthly)	←	device(mobile)		-0.0985	0.0626	-1.5746	0.12						
fb_use_tracking_cat5(weekly)	←	device(mobile)		-0.0568	0.0426	-1.3314	0.18						
fb_use_tracking_cat5(daily)	←	device(mobile)		0.4137	0.0829	4.9919	6.0e-7						
fb_use_tracking_cat5(several times a day)	←	device(mobile)		-0.2757	0.0937	-2.9439	0.0032						
fb_use_tracking_cat5(less than monthly)	←	device(mobile_PC)		-0.6986	0.0566	-12.3326	6.0e-35						
fb_use_tracking_cat5(monthly)	←	device(mobile_PC)		-0.2973	0.0599	-4.9631	6.9e-7						
fb_use_tracking_cat5(weekly)	←	device(mobile_PC)		0.0317	0.0406	0.7826	0.43						
fb_use_tracking_cat5(daily)	←	device(mobile_PC)		0.5511	0.0792	6.9551	3.5e-12						
fb_use_tracking_cat5(several times a day)	←	device(mobile_PC)		0.4130	0.0844	4.8912	1.0e-6						

HMMs in Latent Gold



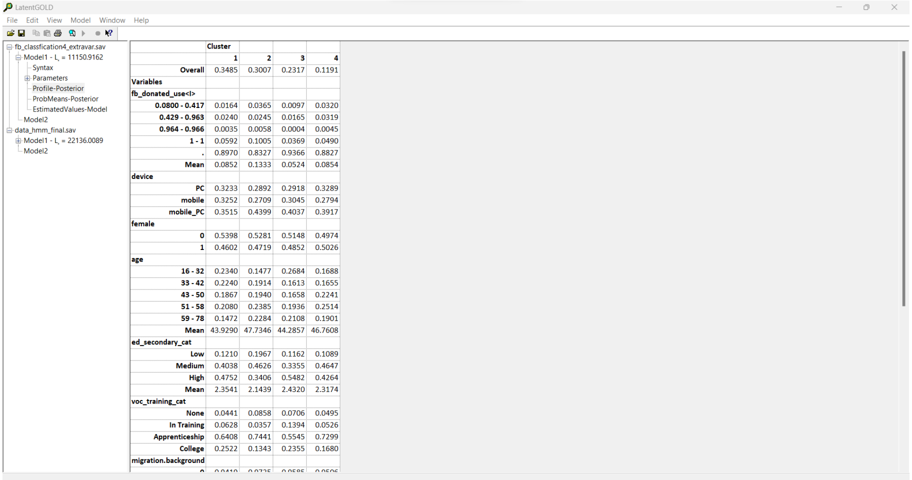
HMMs in Latent Gold



Regression Parameters

		term	coef	s.e.	z-value	p-value	Wald(0)	df	p-value
Cluster(1)	←	1	0.2265	0.1934	1.1708	0.24	7.0147	3	0.071
Cluster(2)	←	1	0.4028	0.2014	2.0007	0.045			
Cluster(3)	←	1	-0.3863	0.2305	-1.6760	0.094			
Cluster(4)	←	1	-0.2430	0.2799	-0.8681	0.39			
Cluster(1)	←	device(PC)	0.0594	0.0358	1.6594	0.097	33.4702	6	8.5e-6
Cluster(2)	←	device(PC)	-0.0726	0.0373	-1.9477	0.051			
Cluster(3)	←	device(PC)	-0.0366	0.0416	-0.8797	0.38			
Cluster(4)	←	device(PC)	0.0497	0.0509	0.9763	0.33			
Cluster(1)	←	device(mobile)	0.0898	0.0364	2.4647	0.014			
Cluster(2)	←	device(mobile)	-0.0768	0.0381	-2.0170	0.044			
Cluster(3)	←	device(mobile)	0.0558	0.0417	1.3386	0.18			
Cluster(4)	←	device(mobile)	-0.0688	0.0539	-1.2770	0.20			
Cluster(1)	←	device(mobile_PC)	-0.1492	0.0348	-4.2934	1.8e-5			
Cluster(2)	←	device(mobile_PC)	0.1493	0.0349	4.2764	1.9e-5			
Cluster(3)	←	device(mobile_PC)	-0.0192	0.0386	-0.4981	0.62			
Cluster(4)	←	device(mobile_PC)	0.0191	0.0496	0.3848	0.70			
Cluster(1)	←	age	-0.0163	0.0022	-7.3272	2.4e-13	55.0819	3	6.6e-12
Cluster(2)	←	age	0.0051	0.0022	2.3502	0.019			
Cluster(3)	←	age	0.0044	0.0025	1.7632	0.078			
Cluster(4)	←	age	0.0067	0.0031	2.1630	0.031			
Cluster(1)	←	female(0)	-0.0096	0.0263	-0.3651	0.72	0.1722	3	0.98
Cluster(2)	←	female(0)	-0.0006	0.0270	-0.0224	0.98			
Cluster(3)	←	female(0)	-0.0022	0.0300	-0.0720	0.94			
Cluster(4)	←	female(0)	0.0124	0.0376	0.3290	0.74			
Cluster(1)	←	female(1)	0.0096	0.0263	0.3651	0.72			
Cluster(2)	←	female(1)	0.0006	0.0270	0.0224	0.98			
Cluster(3)	←	female(1)	0.0022	0.0300	0.0720	0.94			
Cluster(4)	←	female(1)	-0.0124	0.0376	-0.3290	0.74			
Cluster(1)	←	migration.background(0)	-0.1176	0.0613	-1.9205	0.055	8.5691	3	0.036
Cluster(2)	←	migration.background(0)	0.1632	0.0777	2.1000	0.036			
Cluster(3)	←	migration.background(0)	0.0732	0.0699	1.0480	0.29			
Cluster(4)	←	migration.background(0)	-0.1188	0.0899	-1.3217	0.19			
Cluster(1)	←	migration.background(1)	0.1176	0.0613	1.9205	0.055			
Cluster(2)	←	migration.background(1)	-0.1632	0.0777	-2.1000	0.036			
Cluster(3)	←	migration.background(1)	0.0732	0.0699	1.0480	0.29			

HMMs in Latent Gold



	Cluster			
	1	2	3	4
Overall	0.3485	0.3007	0.2317	0.1191
Variables				
fb_donated_use<l>				
0.0800 - 0.417	0.0164	0.0365	0.0097	0.0320
0.429 - 0.963	0.0240	0.0245	0.0165	0.0319
0.964 - 0.966	0.0035	0.0058	0.0004	0.0045
1 - 1	0.0592	0.1005	0.0369	0.0490
.	0.8970	0.8327	0.9366	0.8827
Mean	0.0852	0.1333	0.0524	0.0854
device				
PC	0.3233	0.2892	0.2918	0.3289
mobile	0.3252	0.2709	0.3045	0.2794
mobile_PC	0.3515	0.4399	0.4037	0.3917
female				
0	0.5398	0.5281	0.5148	0.4974
1	0.4602	0.4719	0.4852	0.5026
age				
16 - 32	0.2340	0.1477	0.2684	0.1688
33 - 42	0.2240	0.1914	0.1613	0.1655
43 - 50	0.1867	0.1940	0.1658	0.2241
51 - 58	0.2080	0.2385	0.1936	0.2514
59 - 78	0.1472	0.2284	0.2108	0.1901
Mean	43.9290	47.7346	44.2857	46.7608
ed_secondary_cat				
Low	0.1210	0.1967	0.1162	0.1089
Medium	0.4038	0.4626	0.3355	0.4647
High	0.4752	0.3406	0.5482	0.4264
Mean	2.3541	2.1439	2.4320	2.3174
voc_training_cat				
None	0.0441	0.0858	0.0706	0.0495
In Training	0.0628	0.0357	0.1394	0.0526
Apprenticeship	0.6408	0.7441	0.5545	0.7299
College	0.2522	0.1343	0.2355	0.1680
migration.background				

Concluding remarks

- HMMs potentially a powerful tool to correct for measurement error

  But...

- Difficult to use
  - No comprehensive R packages
  - Important to know data generating process

Thank you!

Utrecht University