A penalized maximum likelihood estimation for hidden Markov models to address latent state separation

Luca Brusa[1], Francesco Bartolucci[2],
Fulvia Pennoni[1], Romina Peruilh Bagolini[2]

(luca.brusa@unimib.it)

[1]University of Milano-Bicocca - Department of Statistics, and Quantitative Methods
[2]University of Perugia - Department of Economics

Gießen, 28.08.2024

Outline

1. Hidden Markov models
2. Penalized maximum likelihood approach
3. Simulation study
4. Application
5. References

Hidden Markov model: notation

- **Univariate binary response variables** $\boldsymbol{Y}_i = (Y_i^{(1)}, \ldots, Y_i^{(T)})$, with

$$Y_i^{(t)} = \begin{cases} 1 & \text{if the event of interest is observed at time } t \text{ for unit } i \\ 0 & \text{otherwise} \end{cases}$$

- **Time-fixed and time-varying covariates**: $\boldsymbol{x}_i = (\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(T)})$, with $\boldsymbol{x}_i^{(t)}$ representing the vector of observed individual covariates for unit i at time t

- **Hidden process**: $\boldsymbol{U}_i = (U_i^{(1)}, \ldots, U_i^{(T)})$, following a first-order Markov chain with state-space $\{1, \ldots, k\}$

Model formulation

1. **Measurement model**: $p\left(y_i^{(t)} \mid u_i^{(t)}, \boldsymbol{x}_i^{(t)}, y_i^{(t-1)}\right)$
   - represents the conditional distribution of the response variable $Y_i^{(t)}$ given the latent process $U_i^{(t)}$, with covariates $\boldsymbol{x}_i^{(t)}$ and lagged response variable $Y_i^{(t-1)}$
   - covariates directly influence the response variable
   - the lagged response among covariates allows for serial dependence between observed responses over time, thus relaxing the conditional independence of $\boldsymbol{Y}$ given $\boldsymbol{U}$ and $\boldsymbol{x}$

2. **Latent model**: $p\left(\boldsymbol{u}_i\right)$
   - represents the non-parametric distribution of the latent process
   - is not affected by covariates: the same latent model holds for all units
   - accounts for unobserved heterogeneity between individuals, which remains when observed covariates in the measurement model cannot fully explain the variability

Model parameters

1. **Conditional response probabilities**, given the latent state, the covariate configuration, and the lagged response:

$$\phi_{u\boldsymbol{x}y}^{(t)} = \mathbb{P}\left(Y_i^{(t)} = 1 | u_i^{(t)}, \boldsymbol{x}_i^{(t)}, y_i^{(t-1)}\right),$$

such that:

$$\log \frac{\phi_{uxy}^{(t)}}{1 - \phi_{uxy}^{(t)}} = \mu + \alpha_u + \boldsymbol{x}_i' \boldsymbol{\beta} + y_i^{(t-1)} \gamma$$

- μ: intercept
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$: support points corresponding to the latent states
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$: regression parameters for the covariates
- γ: parameter for the lagged response variable

2. **Initial** and **transition probabilities**, denoted as π_u and $\pi_{u|\bar{u}}$, respectively

Maximum likelihood estimation

- **Expectation-maximization** (EM) algorithm (Dempster et al., 1977) is often employed to perform **maximum likelihood estimation**
- It maximizes the observed-data log-likelihood function $\ell(\boldsymbol{\theta})$ relying on the **complete-data log-likelihood** function $\ell^*(\boldsymbol{\theta})$
- It alternates the following steps until convergence:
  - **E-step**: **compute the conditional expected value** of $\ell^*(\boldsymbol{\theta})$ given the value of the parameters at the previous step and the observed data
  - **M-step**: **update the model parameters** by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$:
    - explicit solutions are available for π_u and $\pi_{u|\bar{u}}$
    - a Newton-Raphson algorithm is used for updating μ, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ

Outline

1 Hidden Markov models

2 Penalized maximum likelihood approach

3 Simulation study

4 Application

5 References

Motivation

- When the available **covariates do not fully explain the heterogeneity** between individuals, the support points α_u may be very large, leading to **widely separated latent states**

- This may results in:
  - **excessively higher relevance of one or more latent states** than others
  - **reduced importance of the available covariates** whose estimated effects may become negligible and insignificant
  - **instability of the estimates**

Penalization term

- **Proposed penalization** term (**aimed at reducing separation among latent states**):

$$\mathcal{A} = \sum_{u=1}^{k} (\alpha_u - \bar{\alpha})^2,$$

where $\bar{\alpha} = \frac{1}{k} \sum_{u=1}^{k} \alpha_u$

- In matrix notation (computationally convenient):

$$\mathcal{A} = \boldsymbol{\alpha}' \boldsymbol{J} \boldsymbol{\alpha},$$

where $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{k}\mathbf{1}\mathbf{1}'$, $\boldsymbol{I}$ is the identity matrix, and $\mathbf{1} = (1, \ldots, 1)'$

Penalized maximum likelihood estimation

- The **proposed penalization** term is **applied to** both the **observed-data log-likelihood** $\ell(\boldsymbol{\theta})$ and the **complete-data log-likelihood** $\ell^*(\boldsymbol{\theta})$:

$$\tilde{\ell}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \lambda\mathcal{A} \qquad \text{and} \qquad \tilde{\ell}^*(\boldsymbol{\theta}) = \ell^*(\boldsymbol{\theta}) - \lambda\mathcal{A},$$

  where $\lambda \in \mathbb{R}^+$ is a **tuning parameter** controlling the penalization

- Penalized estimation is performed using the **EM algorithm**, where:
  - the **E-step** remains unaltered
  - the **M-step** requires to **revise the Newton-Raphson** iteration for $\boldsymbol{\alpha}$

Outline

1. Hidden Markov models
2. Penalized maximum likelihood approach
3. Simulation study
4. Application
5. References

Simulation study

- **Different scenarios** (32) to explore the performance of the proposal with $k = 3$ hidden states: **sample size** ($n = 250, 500$), **number of time occasions** ($T = 10, 20$), hidden **state persistence** (high or low) and **separation** (four different behaviors: $\boldsymbol{\alpha}^j$, $j = 1, \dots, 4$)

- **Four covariates** also including the **lagged response variable**; the corresponding vector of regression coefficients is $\boldsymbol{\beta} = (1, -1, 1, 1)'$

- Extensive **Monte Carlo simulation study**; for each scenario:
  - we randomly draw 50 samples
  - we estimate the HM model using both the standard approach and the penalized approach with $\lambda = 0.01$ and $\lambda = 0.05$

Comparison criteria

- **Percentage variation in** the following quantities for both procedures:

  1. **root mean squared relative error between true and estimated model parameters** defined as:

  $$\text{RMSRE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{\theta}_m - \theta_m}{\theta_m}\right)^2}$$

  2. **standard errors of the covariate regression parameters** $(\boldsymbol{\beta}, \gamma)$, obtained as minus the second derivative of the expected value of the complete-data log-likelihood

  3. **computational time**

Results - RMSRE

- In the majority of cases, both values of λ ensure a **lower RMSRE when using the penalized estimation** method. This indicates that the estimated parameters are closer to the true values (**higher estimation precision**)
- The **fourth scenario** (characterized by highly separated states) shows the **greatest improvement**; the percentage decreases using penalized estimation are often exceeding 90%
- The penalization approach appears **less effective in the first scenario** (characterized by closely spaced states) and in cases with a high value of T (20 time occasions)
- **Only in two cases** the penalized estimation method exhibits **no improvement** with either value of λ, showing very slight increases in the RMSRE value

Other results

- **Standard errors**:
  - In most cases, **penalization reduces the estimated standard errors**
  - The proposed approach is **less effective under the first scenario**
  - In all other cases, the **percentage decrease is significant**, often reaching very high values

- **Computational time**:
  - Estimation with the **penalty** approach often **reduces the average computational time**
  - Benefits are **particularly evident in the fourth scenario** where hidden states are widely separated

Outline

1 Hidden Markov models

2 Penalized maximum likelihood approach

3 Simulation study

4 Application

5 References

Hypotension during spinal anesthesia

- Data[1] refer to **375 patients** undergoing spinal anesthesia during a surgery; they cover the period **from January 2008 to January 2011**

- Measurements are taken 8 times, at equally spaced intervals over a period of 40 minutes

- For each patient-time observation, a binary variable indicates whether or not the patient has experienced **hypotension** (decrease in mean systolic blood pressure)

- Approximately **25% (**$n = 94$**) of patients** recorded **at least one hypotensive status**

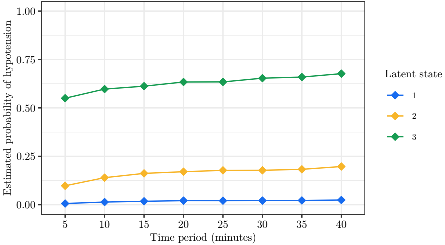
[1] Data are freely available at https://peerj.com/articles/648/.

Covariates

- **Time-fixed covariates**: age, gender, type of surgical hospital unit (general surgery, urology, obstetrics and gynecology), position of the patient during the surgery (lithotomy, supine), electrocardiography status (normal, abnormal), and doses of medication in the blood (Marcain-heavy, chirocaine, fentanyl and midazolam)
- **Time-varying covariates**: diastolic blood pressure, and patient pulse rate
- **Lagged response variable** to relax the conditional independence assumption and measure state dependence

Estimated conditional hypotension probability

- Patients in the **first hidden state** ($\hat{\alpha}_1 = -0.827$) have an **almost negligible probability of hypotension** during the surgery
- Patients in the **second hidden state** ($\hat{\alpha}_2 = 3.147$) experience a **low probability of hypotension**, ranging approximately from 0.10 to 0.20
- Patients in the **third hidden state** ($\hat{\alpha}_3 = 7.359$) have a **high probability of hypotension** during surgery, ranging from 0.54 to 0.68



Estimated regression coefficients

Covariate	$\hat{\beta}$	$\hat{se}$	p-value
Intercept	3.597 **	3.485	0.000
Gender (Female)	1.541 *	2.032	0.042
Position (Supine)	0.813	1.597	0.110
Operation (Urogoly)	0.363	0.338	0.735
Operation (General surgery)	-0.061	-0.068	0.946
ECG (Normal)	-0.878	-0.978	0.328
Age (year)	0.037 *	1.967	0.049
DBP	-0.167 **	-7.642	0.000
Pulse rate	-0.002	-0.216	0.829
Marcain-heavy	-0.049	-1.235	0.217
Midazolam	0.243 *	2.104	0.035
Chirocaine	-0.049	-1.346	0.178
Fentanyl	1.215	0.515	0.607
Hypotension ($t-1$)	2.675 **	10.289	0.000

- **Gender** (female) has a **significant positive effect** on the response variable, indicating that the conditional **probability of experimenting hypotension** given the latent state is **higher for females**

Estimated regression coefficients

Covariate	$\hat{\beta}$	$\hat{se}$	p-value
Intercept	3.597 **	3.485	0.000
Gender (Female)	1.541 *	2.032	0.042
Position (Supine)	0.813	1.597	0.110
Operation (Urogoly)	0.363	0.338	0.735
Operation (General surgery)	-0.061	-0.068	0.946
ECG (Normal)	-0.878	-0.978	0.328
Age (year)	0.037 *	1.967	0.049
DBP	-0.167 **	-7.642	0.000
Pulse rate	-0.002	-0.216	0.829
Marcain-heavy	-0.049	-1.235	0.217
Midazolam	0.243 *	2.104	0.035
Chirocaine	-0.049	-1.346	0.178
Fentanyl	1.215	0.515	0.607
Hypotension ($t-1$)	2.675 **	10.289	0.000

- **Older individuals** exhibit **higher log-odds** of being diagnosed with **hypotension** compared to younger individuals

Estimated regression coefficients

Covariate	$\hat{\beta}$	$\hat{se}$	p-value
Intercept	3.597 **	3.485	0.000
Gender (Female)	1.541 *	2.032	0.042
Position (Supine)	0.813	1.597	0.110
Operation (Urogoly)	0.363	0.338	0.735
Operation (General surgery)	-0.061	-0.068	0.946
ECG (Normal)	-0.878	-0.978	0.328
Age (year)	0.037 *	1.967	0.049
DBP	-0.167 **	-7.642	0.000
Pulse rate	-0.002	-0.216	0.829
Marcain-heavy	-0.049	-1.235	0.217
Midazolam	0.243 *	2.104	0.035
Chirocaine	-0.049	-1.346	0.178
Fentanyl	1.215	0.515	0.607
Hypotension ($t-1$)	2.675 **	10.289	0.000

- **Diastolic blood pressure** has a significant **negative effect** on the log-odds of hypotension: lower pressure is associated with higher probabilities of experiencing hypotension

Estimated regression coefficients

Covariate	$\hat{\beta}$	$\hat{se}$	p-value
Intercept	3.597 **	3.485	0.000
Gender (Female)	1.541 *	2.032	0.042
Position (Supine)	0.813	1.597	0.110
Operation (Urogoly)	0.363	0.338	0.735
Operation (General surgery)	-0.061	-0.068	0.946
ECG (Normal)	-0.878	-0.978	0.328
Age (year)	0.037 *	1.967	0.049
DBP	-0.167 **	-7.642	0.000
Pulse rate	-0.002	-0.216	0.829
Marcain-heavy	-0.049	-1.235	0.217
Midazolam	0.243 *	2.104	0.035
Chirocaine	-0.049	-1.346	0.178
Fentanyl	1.215	0.515	0.607
Hypotension $(t-1)$	2.675 **	10.289	0.000

- **Midazolam** has a significant **positive effect**, indicating that higher concentration of this drug in the blood is associated with increased odds of experiencing hypotension during surgery. For the other drugs, the estimated coefficients are not significant

Estimated regression coefficients

Covariate	$\hat{\beta}$	$\hat{se}$	p-value
Intercept	3.597 **	3.485	0.000
Gender (Female)	1.541 *	2.032	0.042
Position (Supine)	0.813	1.597	0.110
Operation (Urogoly)	0.363	0.338	0.735
Operation (General surgery)	-0.061	-0.068	0.946
ECG (Normal)	-0.878	-0.978	0.328
Age (year)	0.037 *	1.967	0.049
DBP	-0.167 **	-7.642	0.000
Pulse rate	-0.002	-0.216	0.829
Marcain-heavy	-0.049	-1.235	0.217
Midazolam	0.243 *	2.104	0.035
Chirocaine	-0.049	-1.346	0.178
Fentanyl	1.215	0.515	0.607
Hypotension ($t-1$)	2.675 **	10.289	0.000

- The **lagged response** has a significant **positive effect on hypotension** indicating serial correlation

Outline

1. Hidden Markov models
2. Penalized maximum likelihood approach
3. Simulation study
4. Application
5. References

References I

Aktas Samur, A., Coskunfirat, N., and Saka, O. (2014). Comparison of predictor approaches for longitudinal binary outcomes: Application to anesthesiology data. *PeerJ*, **2**, e648.

Bartolucci, F. and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent markov heterogeneity structure. *J. Am. Stat. Assoc.*, **104**, 816–831.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov models for longitudinal data*. Chapman & Hall/CRC, Boca Raton.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2014). Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test*, **23**, 433–465.

References II

BATES, D., HASTIE, T., AND TIBSHIRANI, R. (2023). Cross-validation: What does it estimate and how well does it do it? *J. Mach. Learn. Res.*, **24**, 1234–1256.

DEMPSTER, A., LAIRD, N., AND RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Series B Stat. Methodol.*, **39**, 1–38.

SMYTH, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Stat. Comput.*, **9**, 63–72.

WELCH, L. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Inform. Theory Soc. Newsl.*, **50**, 10–13.

Cross-validated log-likelihood

- A **cross-validation** approach is employed to **jointly select** the **penalization parameter** λ and the **number of states** k of the hidden chain
- We consider M **partitions of the data** D: $(D \backslash S_m,\ S_m)_{m=1,\dots,M}$
- For the m-th partition:
  - the model is estimated on the data subset $D \backslash S_m$, providing parameters estimates $\hat{\boldsymbol{\theta}}^{(k,\lambda)}(D \backslash S_m)$
  - $\ell\left(\hat{\boldsymbol{\theta}}^{(k,\lambda)}(D \backslash S_m) \mid S_m\right)$ denotes the (possibly penalized) log-likelihood function where the model parameters are estimated on the training data $D \backslash S_m$ but the log-likelihood is evaluated on the test data S_m
  - the **cross-validated likelihood** is defined as

$$\ell_{\mathrm{CV}} = \frac{1}{M} \sum_{m=1}^{M} \ell\left(\boldsymbol{\theta}^{(k,\lambda)}(D \backslash S_m) \mid S_m\right).$$

Estimation settings

- We use a **cross-validation approach** to jointly select the number of hidden states ($k = 1, \ldots, 4$) and the roughness of the penalty ($\lambda = 0.00, 0.01, 0.05$)
- To mitigate the risk of convergence to local maxima, the estimation of each HM model is **repeated 25 times**, employing both deterministic and random initialization methods
- The results indicate that **$k = 3$ hidden states** and a **penalization parameter of $\lambda = 0.01$ are optimal**

Limitations and future works

- **Implementation in C++ to improve computational speed**, particularly for datasets with a large number of repeated measurements and/or a large sample size
- **Evaluation of the model's predictive performance** and comparison with machine learning methods, also in connection with the use of HM models as early warning systems
- **Development of feature selection techniques** to identify relevant covariates, especially when many are available
- **Investigation of methods to achieve scalability**, such as parallel computation and dimension reduction, essential for handling large datasets efficiently