Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani Piano Nazionale di Ripresa e Resilienza
Università degli Studi di Milano Bicocca

Dealing with hidden state separation in hidden Markov models via penalized maximum likelihood estimation

Luca Brusa[1], Fulvia Pennoni[1],
Romina Peruilh Bagolini[2], Francesco Bartolucci[2]
(luca.brusa@unimib.it)

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

Outline

Hidden Markov model: notation and formulation

- **Univariate binary response variables** $Y_i^{(t)}$, observed for unit i at time occasion t
- **Hidden process** $U_i^{(t)}$, following a first-order Markov chain with state-space $\{1, \ldots, k\}$
- **Covariates** $\boldsymbol{x}_i^{(t)}$, representing the vector of observed individual covariates for unit i at time t
- Composed of two sub-models:
  1. **Measurement model**: conditional distribution of the response variable $Y_i^{(t)}$ given the latent variable $U_i^{(t)}$ and the possible influence of individual time-varying and time-fixed covariates $\boldsymbol{x}_i^{(t)}$
  2. **Latent model**: non-parametric distribution of the latent process; accounts for the **unobserved heterogeneity** between individuals

Hidden Markov model: parameters

- **Initial probabilities**: $\pi_u = p(U_i^{(1)} = u)$
- **Transition probabilities**: $\pi_{u|\bar{u}} = p(U_i^{(t)} = u | U_i^{(t-1)} = \bar{u})$
- **Conditional response probabilities**:

$$\phi_{y|u\boldsymbol{x}}^{(t)} = p\big(Y_i^{(t)} = y | U_i^{(t)} = u, \boldsymbol{X}_i^{(t)} = \boldsymbol{x}\big),$$

based on the following **global logit parameterization**:

$$\log \frac{\phi_{y|u\boldsymbol{x}}^{(t)} + \cdots + \phi_{c-1|u\boldsymbol{x}}^{(t)}}{\phi_{0|u\boldsymbol{x}}^{(t)} + \cdots + \phi_{y-1|u\boldsymbol{x}}^{(t)}} = \alpha_u + \boldsymbol{x}'\boldsymbol{\beta},$$

with:

- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$: support points of the latent variables
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$: regression parameters for the covariates

Maximum likelihood estimation

- The **expectation-maximization** (EM) algorithm is 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 $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$
- Standard errors of the estimates are obtained as minus the second derivative of the expected value of the complete-data log-likelihood

Outline

Motivation

- **separation problem**: the support points α_u may be very large and far apart, leading to **widely separated latent states**

- This may result 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** and **large standard errors**

Penalized maximum likelihood estimation

- Proposed **penalty**: to **reduce latent states separation**:

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

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

- **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 through the **EM algorithm**:
  - the **E-step** remains unaltered
  - the **M-step** requires to **revise the Newton-Raphson** iteration for $\boldsymbol{\alpha}$

Cross-validated log-likelihood

- A **cross validation** (**CV**) 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

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

Outline

Simulation study

- **Different scenarios** (48) 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 level** (five different behaviors: $\boldsymbol{\alpha}^j$, $j = 1, \dots, 6$)

- **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 standard and penalized approaches

Comparison criteria

- **Percentage variation** in the following quantities for the penalized estimation approach with respect to the standard one:
  1. **root mean squared relative error between real (β_j) and estimated ($\hat{\beta}_{js}$) model parameters**, defined as:

$$\text{MSE}(\beta_j) = \frac{1}{50}\sum_{s=1}^{50}(\beta_j - \hat{\beta}_{js})^2, \quad j = 1, 2, 3, 4$$

  2. **standard errors of the covariate regression parameters**
  3. **computational time**

Fixed tuning parameter ($\lambda = 0.01$) - MSE

- In the majority of scenarios (40 out of 48), the **penalized estimation method yields lower MSE** values for all regression coefficients β_j (**higher estimation accuracy**)
- Scenarios characterized by highly separated states show the **greatest improvement**; the percentage decreases using penalized estimation often exceeds 90%
- The penalization approach is **less effective** in scenario characterized by **closely spaced states** and in cases with a **high value of T** (20 time occasions)
- In the **few instances** where the **penalized estimation does not improve** accuracy, the **increase in MSE is extremely low** (typically below 1%)

Fixed tuning parameter ($\lambda = 0.01$) - Other results

- **Standard errors**:
  - In most scenarios, **penalization reduces the estimated standard errors** (greater precision of the parameter estimates)
  - The proposed approach is **less effective** in scenarios characterized by **poorly separated latent states**
  - In all other cases, the **percentage decrease is significant**, reaching very high values (over **90%** in scenarios with **highly separated** states)
- **Computational time**:
  - Estimation with the **penalty** approach often **reduces the average computational time**
  - Benefits are **particularly evident** (percentage reduction up to 50%) when **hidden states are widely separated**

Cross validation

- Settings:
  - 10-fold cross validation
  - Grid search: $\lambda \in \{0.00, 0.01, \ldots, 0.05\}$
- Results:
  - Consistent with the fixed-λ analysis, the CV approach yields lower MSE, standard errors, and computational times compared to standard estimation method in most scenarios
  - **High separation** scenarios: CV results are similar to those obtained with λ fixed at 0.01 $\rightarrow$ Penalization, regardless of precise tuning, is sufficient to improve estimation
  - **Low separation** scenarios: CV results outperform those obtained with the fixed penalty approach $\rightarrow$ Selecting the correct tuning parameter becomes crucial

Outline

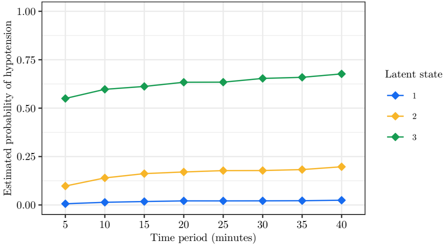
Hypotension during spinal anesthesia

- 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
- Variable $Y_i^{(t)}$ indicates whether or not patient i has experienced **hypotension** (decrease in mean systolic blood pressure) at time t
- Approximately **25% ($n = 94$) of patients** recorded **at least one hypotensive status**
- Time-fixed covariates, time-varying covariates, and lagged response
- Using a **CV approach**, we select **$k = 3$ hidden states** and a **penalization parameter of $\lambda = 0.02$** as optimal

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

Estimated conditional hypotension probability

- Patients in the **first hidden state** ($\hat{\alpha}_1 = -0.821$) 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
Gender (Female)	1.526 *	0.752	0.042
Position (Supine)	0.797	0.508	0.116
Operation (Urogoly)	0.374	1.073	0.727
Operation (General surgery)	-0.058	0.897	0.949
ECG (Normal)	-0.858	0.893	0.337
Age (year)	0.036 *	0.019	0.042
DBP	-0.166 **	0.022	0.000
Pulse rate	-0.002	0.011	0.858
Marcain-heavy	-0.049	0.040	0.220
Midazolam	0.242 *	0.115	0.036
Chirocaine	-0.049	0.036	0.180
Fentanyl	1.210	2.350	0.608
Hypotension $(t-1)$	2.678 **	0.011	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
Gender (Female)	1.526 *	0.752	0.042
Position (Supine)	0.797	0.508	0.116
Operation (Urogoly)	0.374	1.073	0.727
Operation (General surgery)	-0.058	0.897	0.949
ECG (Normal)	-0.858	0.893	0.337
Age (year)	0.036 *	0.019	0.042
DBP	-0.166 **	0.022	0.000
Pulse rate	-0.002	0.011	0.858
Marcain-heavy	-0.049	0.040	0.220
Midazolam	0.242 *	0.115	0.036
Chirocaine	-0.049	0.036	0.180
Fentanyl	1.210	2.350	0.608
Hypotension ($t-1$)	2.678 **	0.011	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
Gender (Female)	1.526 *	0.752	0.042
Position (Supine)	0.797	0.508	0.116
Operation (Urogoly)	0.374	1.073	0.727
Operation (General surgery)	-0.058	0.897	0.949
ECG (Normal)	-0.858	0.893	0.337
Age (year)	0.036 *	0.019	0.042
DBP	-0.166 **	0.022	0.000
Pulse rate	-0.002	0.011	0.858
Marcain-heavy	-0.049	0.040	0.220
Midazolam	0.242 *	0.115	0.036
Chirocaine	-0.049	0.036	0.180
Fentanyl	1.210	2.350	0.608
Hypotension $(t-1)$	2.678 **	0.011	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
Gender (Female)	1.526 *	0.752	0.042
Position (Supine)	0.797	0.508	0.116
Operation (Urogoly)	0.374	1.073	0.727
Operation (General surgery)	-0.058	0.897	0.949
ECG (Normal)	-0.858	0.893	0.337
Age (year)	0.036 *	0.019	0.042
DBP	-0.166 **	0.022	0.000
Pulse rate	-0.002	0.011	0.858
Marcain-heavy	-0.049	0.040	0.220
Midazolam	0.242 *	0.115	0.036
Chirocaine	-0.049	0.036	0.180
Fentanyl	1.210	2.350	0.608
Hypotension $(t-1)$	2.678 **	0.011	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
Gender (Female)	1.526 *	0.752	0.042
Position (Supine)	0.797	0.508	0.116
Operation (Urogoly)	0.374	1.073	0.727
Operation (General surgery)	-0.058	0.897	0.949
ECG (Normal)	-0.858	0.893	0.337
Age (year)	0.036 *	0.019	0.042
DBP	-0.166 **	0.022	0.000
Pulse rate	-0.002	0.011	0.858
Marcain-heavy	-0.049	0.040	0.220
Midazolam	0.242 *	0.115	0.036
Chirocaine	-0.049	0.036	0.180
Fentanyl	1.210	2.350	0.608
Hypotension ($t-1$)	2.678 **	0.011	0.000

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

Outline

Limitations and future works

- Implementation of the estimation procedure in **C++** to **improve computational speed**
- **Parallel computation** of the CV approach, distributing the grid search across multiple cores
- **Adoption of genetic algorithms** as an alternative to the exhaustive grid search in CV
- **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
- **Theoretical investigation** to better understand the circumstances under which latent state separation occurs

Outline

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.

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Acknowledgments

The authors acknowledge the financial support from the grant "Hidden Markov Models for Early Warning Systems" of Ministero dell'Università e della Ricerca (PRIN 2022TZEXKF) funded by European Union - Next Generation EU, Mission 4, Component 2, CUP J53D23004990006.

State separation: a motivating example

- Data generation:
  - one binary response variable for $n = 250$ units over $T = 10$ time points
  - four covariates (including the lagged response), with $\boldsymbol{\beta} = [1, -1, 1, 1]'$
  - $k = 3$ latent states with support points $\alpha = [-20, -5, 5]'$
  - 1^{st} time point: 88 units in the 1^{st} state, 71 in the 2^{nd}, and 91 in the 3^{rd}

- Model estimation using standard maximum likelihood:

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	-250.13	146.80	457.96	24.14	-80.21	56.69	103.48
Standard error	-	-	-	23.28	69.43	50.91	90.14
p-value	-	-	-	0.30	0.25	0.27	0.25

  - poor accuracy in parameter estimation for both α_u and β_j
  - 1^{st} time point: 159 units in the 1^{st} state and 91 in the 3^{rd}
  - conditional response probabilities equal to 0 (1^{st} state) or 1 (3^{rd} state)