Non-parametric Observation Driven Hidden Markov Model

Hanna Bacave[(1)], Pierre-Olivier Cheptou[(2)], Nikolaos Limnios[(3)], Nathalie Peyrard[(1)]

[(1)] INRAE, UR MIAT, Université de Toulouse, Castanet-Tolosan, France.

[(2)] CEFE-CNRS, Université de Montpellier, France.

[(3)] Sorbonne University Alliance, Université de Technologie de Compiègne, LMAC, France.

June 2023

Introduction

Hidden Markov Models : Uses and results

Example (HMM uses)

Field of study	Uses
Medicine	Analyse epidemiologic surveillance data (Le Strat and Carrat 1999)
Ecology	Reconstruct hidden or partially observed ecological dynamics (McClintock et al. 2020)
Finance	Predict the regime of a monetary system thanks to the exchange rate (Engel and Hamilton 1990)

Results established for HMM :

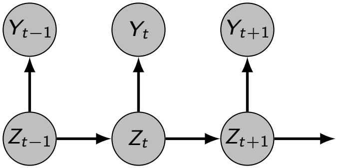
- Model identifiability (Allman et al. 2009; Cappé et al. 2005)
- Asymptotic properties of the Maximum Likelihood Estimator (MLE, Cappé et al. 2005)
- Expectation-Maximization algorithm for HMM

Hidden Markov Models : Limits

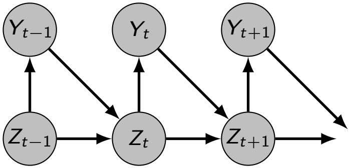
Example (Cases in which HMM is not sufficient)

- In the case of plants, due to the fact that the seeds produced by the grown plants enter the soil each year an HMM is not sufficient (Pluntz et al. 2018; Le Coz et al. 2019).
- In finance, in an extension of the Hamilton's Markov-switching Model (Hamilton 1989) the hidden regime depends on the observed financial data.

So, we propose the **Observation-Driven HMM**.



(a) for an HMM.



(b) for an OD-HMM.

OD-HMM

Model definition

We work on discrete time, with discrete and finite state spaces.

- initial probability :

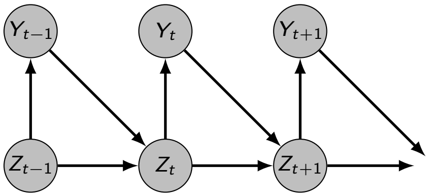
$$\pi(z_0) = \mathbb{P}(Z_0 = z_0);$$

- emission probability :

$$R(z_t, y_t) = \mathbb{P}(Y_t = y_t | Z_t = z_t);$$

- transition matrix :

$$P_{y_{t-1}}(z_{t-1}, z_t) = \mathbb{P}(Z_t = z_t | Z_{t-1} = z_{t-1}, Y_{t-1} = y_{t-1}).$$



Specificity of OD-HMM

There are as many transition matrices as observed states.

Generic identifiability

Generic identifiability for OD-HMM adapted from Allman et al. (2009)

The parameters of an OD-HMM with $|\Omega_Z|$ hidden states and $|\Omega_Y|$ observable states are generically identifiable from the marginal distribution of $2L+1$ consecutive variables provided L satisfies:

$$\binom{L + |\Omega_Y| - 1}{|\Omega_Y| - 1} \geqslant |\Omega_Z||\Omega_Y|,$$

where $|\Omega_Z|$ is the size of the hidden state space and $|\Omega_Y|$ is the size of the observed state space.

Example

For $\Omega_Z = 2$ and $\Omega_Y = 2$, the set of parameters θ are identifiable as soon as the chain has more than $2L + 1 = 7$ observations.

Maximum Likelihood Estimation

Consistency of MLE

- OD-HMM is a particular case of *Non Homogeneous Markov-Switching Auto-Regressive models* as defined in Ailliot and Pène (2015)
- We traduce their general results on consistency of the MLE for NHMS-AR

Sufficient conditions for OD-HMM

- The elements of P_y and R are continuous in θ, for any y in Ω_Y.
- Assumption satisfyed for finite state space and non-parametric case

EM algorithm for HMM

EM algorithm for HMM

Iterative algorithm: each iteration is composed of two steps.

1. **Expectation Step (E step):** computation of the marginal distributions involved in the expression of the intermediate quantity $Q(\theta|\theta^{(m)}) = \mathbb{E}_{\theta^{(m)}}\left[\ln \mathbb{P}_{\theta}(Y_{0:M}, Z_{0:M})|Y_{0:M} = y_{0:M}\right]$.
2. **Maximization step (M step):** updating of the set of parameters θ by resolving $\theta^{(m+1)} = \arg\max_{\theta} Q(\theta|\theta^{(m)})$.

For OD-HMM, the forward-backward equations of the E step must be adapted to take into account the fact that the transition depends on the observations.

Validation of the estimation procedure - Protocol

Protocol for the tests (repeated 50 times)

1. We simulated C independent chains (Z_t) and (Y_t) of length $M = 500$.
2. We used the EM algorithm for OD-HMM to obtain the parameters estimator $\hat{\theta}$ from the C observed chains.
3. We quantified the difference between the true parameters θ^* and the estimated ones $\hat{\theta}$ by using the following distance :

$$dist(\theta_i^*, \hat{\theta}_i) = \frac{1}{K} \sum_{k=1}^{K} \frac{|\theta_{i,k}^* - \hat{\theta_{i,k}}|}{\theta_{i,k}^*},$$

where K is the length of θ_i and $\theta_{i,k}$ is the k-th element of the θ i-th row.

Validation of the estimation procedure - Results

	True parameters	Average error for $C = 100$ chains
Test 1	$P_0^* = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}, P_1^* = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}, R^* = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$	0.068
Test 2	$P_0^* = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}, P_1^* = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}, R^* = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}$	0.126
Test 3	$P_0^* = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}, P_1^* = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}, R^* = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}$	0.166

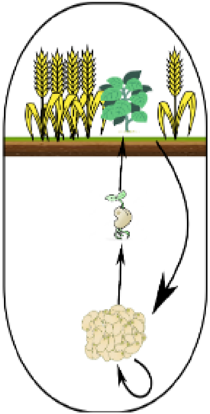
Conclusion :

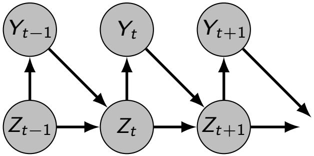
- If the matrices are contrasted, or at least the emission matrix, estimation of the OD-HMM is of good quality.
- Else, if the rows of the emission matrix are similar, we observe label switching phenomenon and a degradation of the estimation quality.
- the R are always better estimated than the P.

Illustration on simulated dynamics of a real ecological system

Model for annual plants dynamics inspired from Pluntz et al. (2018)

Here, Z_t is hidden seed bank, and Y_t is standing flora.





Specificities

- The emission matrix is expressed as a function of the germination, g ;
- The transition matrix depends on the seed survival, s, and the seed production, d.

Estimation of the average time of seeds stock persistence

Protocol for the estimation (repeated on 50 data set with the same θ^*)

1. We simulated with the OD-HMM, C independent chains (Z_t) and (Y_t) of length $M = 500$.
2. We estimated an HMM and an OD-HMM on these data.
3. For both models, we estimated the average time of continuous presence of seeds in the soil using simulated trajectories.

Results :

	Min.	**1st Decile**	**Median**	**Mean**	**9th Decile**	**Max.**
Empirical estimate from data	1	1	2	2.973	6	12
EM for OD-HMM	1	1	2	3.155	7	22
EM for HMM	1	2	9	12.166	27	113

Conclusion

Conclusion

Main results obtained for the OD-HMM :

Perspectives

- Parametric OD-HMM: to be able to handle larger state spaces
- OD-HSMM: extension to the case where the hidden chain is a semi-Markov chain (Barbu and Limnios 2008; Yu 2016), for more realism $\longrightarrow$ how to integrate the impact of the observation on the sojourn time?

References I

Ailliot, P., and F. Pène. 2015. "Consistency of the maximum likelihood estimate for non-homogeneous Markov-switching models." *ESAIM: PS* 19:268–292.

Allman, E., C. Matias, and J. Rhodes. 2009. "Identifiability of parameters in latent structure models with many observed variables." *Annals of Statistics* 37 (6A):3099–3132.

Barbu, V-S., and N. Limnios. 2008. *Semi-Markov Chains and Hidden Semi-Markov Models towards Applications.* Springer.

Cappé, O., E. Moulines, and T. Rydén. 2005. *Inference in Hidden Markov Models.* Springer.

Engel, C., and J.-D. Hamilton. 1990. "Long Swings in the Dollar: Are They in the Data and Do Markets Know It?" *The American Economic Review* 80 (4): 689–713.

Hamilton, J.-D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57 (2): 357–384.

References II

Le Coz, S., P. O. Cheptou, and N. Peyrard. 2019. "A spatial Markovian framework for estimating regional and local dynamics of annual plants with dormancy." *Theoretical Population Biology* 127:120–132.

Le Strat, Y., and F. Carrat. 1999. "Monitoring epidemiologic surveillance data using hidden Markov models." *Statistics in Medicine* 18 (24): 3377–3513.

McClintock, B.-T., R. Langrock, O. Gimenez, E. Cam, D.-L. Borchers, R. Glennie, and T.-A. Patterson. 2020. "Uncovering ecological state dynamics with hidden Markov models." *Ecology Letters* 23 (12): 1878–1903.

Pluntz, M., S. Le Coz, N. Peyrard, R. Pradel, R. Coquet, and P. O. Cheptou. 2018. "A general method for estimating seed dormancy and colonisation in annual plants from the observation of existing flora." *Ecology Letters* 21:1311–1318.

References III

Yu, S-Z. 2016. *Hidden Semi-Markov Models Theory, Algorithms and Applications.* Elsevier.