UNIVERSITÄT BIELEFELD
Faculty of Business Administration and Economics

BiGSEM

Mitigating consequences of the Markov property

Jan-Ole Koslik, Universität Bielefeld, jan-ole.koslik@uni-bielefeld.de

Abstract

*To understand complex real-world phenomena, so-called **hidden Markov models (HMMs)** are a powerful instrument for statistically modelling time series data with **underlying sequential dependencies**. While HMMs have gained popularity in various fields, they have also faced criticism for their reliance on the **Markov assumption**, suggesting that the present can entirely describe future events without consideration of the past. Traditionally, HMMs assume **homogeneity** in the underlying process, where the most probable time spent in a hidden state is one time unit. However, recent years have seen a growing interest in **inhomogeneous models**, allowing for time-varying state-transition dynamics, such as seasonality. We investigate whether the common criticism of HMMs as being overly simplistic in capturing real-world processes remains valid for more complex models by deriving important properties of periodically inhomogeneous Markov chains. Our contribution establishes novel tools for inference and model checking, and a case study reveals that inhomogeneous HMMs hold significant potential to mitigate unrealistic consequences of the Markov assumption.*

1 Motivation

1.1 Basic model formulation

- A basic HMM comprises an **observed state-dependent process** $\{X_t\}$ which is driven by an **unobserved state process** $\{S_t\}$, an N-state Markov chain with transition probability matrix (t.p.m.) $\boldsymbol{\Gamma} = (\gamma_{ij})$, where $\gamma_{ij} = \Pr(S_{t+1} = j \mid S_t = i)$, and an initial distribution $\boldsymbol{\delta}$, where $\delta_i = \Pr(S_1 = i)$.
- Conditional on $S_t = i$, the observed process is assumed to be independent of X_k and S_k for all $k \neq t$ and generated by a **state-dependent distribution** $f_i(x_t)$.

1.2 Problem

- In the case of a homogeneous Markov chain, it is easy to see that once a state $i \in \{1, \dots, N\}$ is entered, leaving it can be interpreted as a **repeated Bernoulli-trial** with probability $p = 1 - \gamma_{ii}$.
- Thus, the time spent in a state, also called the **state dwell time**, is **geometrically distributed** with probability mass function

$$d_i(r) = (1 - \gamma_{ii})\gamma_{ii}^{r-1}, \quad r \in \mathbb{N}.$$

- The geometric distribution is characterised by being **monotonously decreasing** and **memoryless**.
- Both properties may be considered **unrealistic assumptions** when modelling real processes.
- Example "sleeping": The distribution of sleep duration will, in general, not be geometric where the most probable dwell time is one time unit but exhibit a mode greater than one, corresponding to the most probable sleep duration.

2 Periodically inhomogeneous Markov chains

- Recent years have seen growing interest in **inhomogeneous HMMs** where the t.p.m. $\boldsymbol{\Gamma}^{(t)}$ is allowed to vary over time, by linking it to external covariates.
- Here, we focus on a special case of inhomogeneity, namely **periodic variation** or **seasonality**. More formally:

$$\boldsymbol{\Gamma}^{(t)} = \boldsymbol{\Gamma}^{(t+L)}, \quad \text{for all } t = 1, \dots, T, \qquad (1)$$

where L denotes the length of one cycle (e.g. for hourly data and time-of-day variation $L = 24$).

2.1 Periodic stationarity

- Instead of interpreting the t.p.m. as a function of time, it has become common practice to consider a simpler summary statistic, namely the periodically varying **unconditional state distribution**.
- This is usually approximated by the **hypothetical stationary distribution** $\boldsymbol{\rho}^{(t)}$ that would result if the Markov chain was homogeneous with t.p.m. $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^{(t)}$, i.e. $\boldsymbol{\rho}^{(t)}$ is the solution to $\boldsymbol{\rho}^{(t)}\boldsymbol{\Gamma} = \boldsymbol{\rho}^{(t)}$ subject to $\sum_{i=1}^N \rho_i^{(t)} = 1$ (Patterson et al., 2009).
- This **approximation** will, in general, be **biased** because it ignores the preceding process dynamics and instead pretends that the process has been following the dynamics as implied by a constant $\boldsymbol{\Gamma}^{(t)}$ for a considerable time.
- For periodically inhomogeneous Markov chains as defined in (1), there is no need for such an approximation. Consider for fixed t the **thinned Markov chain** $S_t, S_{t+L}, S_{t+2L}, \dots$ which is homogeneous with constant L-*step* t.p.m.

$$\tilde{\boldsymbol{\Gamma}}_t = \boldsymbol{\Gamma}^{(t)}\boldsymbol{\Gamma}^{(t+1)} \dots \boldsymbol{\Gamma}^{(t+L-1)}. \qquad (2)$$

- Provided that this thinned Markov chain is irreducible and aperiodic, it has a unique **stationary distribution** $\boldsymbol{\delta}^{(t)}$, which is the solution to

$$\boldsymbol{\delta}^{(t)}\tilde{\boldsymbol{\Gamma}}_t = \boldsymbol{\delta}^{(t)} \qquad (3)$$

(see also Ge, Jiang, and Qian, 2006; Kargapolova and Ogorodnikov, 2012; Touron, 2019).

- The true periodically stationary distribution and the biased approximation are shown in Figure 2.

2.2 Dwell-time distributions

- We want to apply the knowledge gained from Section 2.1 to characterise state dwell-time distributions in a periodically inhomogeneous setting.
- As a first step, we derive the **time-varying state dwell-time distribution**, i.e. the distribution of the dwell time in state i, when the transition into state i is **known to be at time t**.
- For each state i and time point $t = 1, \dots, L$, this distribution is defined by its probability mass function

$$d_i^{(t)}(r) = (1 - \gamma_{ii}^{(t+r-1)}) \cdot \prod_{j=1}^{r-1} \gamma_{ii}^{(t+j-1)}, \quad r \in \mathbb{N}. \qquad (4)$$

- It can be regarded as a **generalisation** of the geometric distribution to a **time-varying success probability**, for periodic settings.
- Consequently, it is not characterised by a strictly decreasing monotonic pattern (see Figure 1).

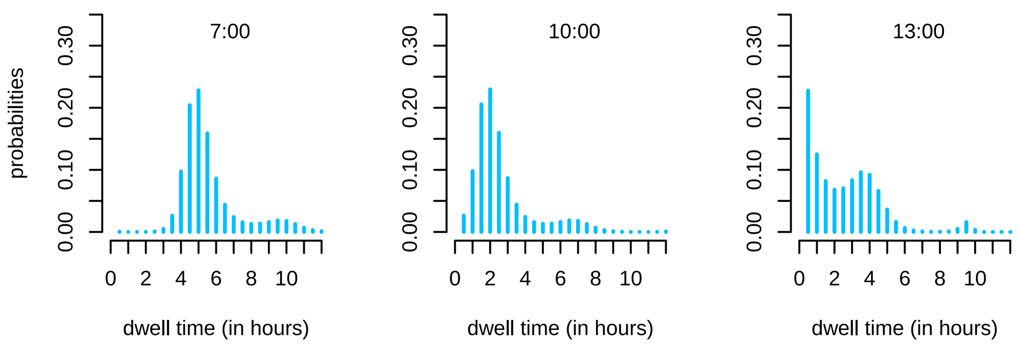

Figure 1: Time-varying dwell-time distribution of an example chain, visualised for three distinct times.

- The time-varying dwell-time distribution allows for **fine-grained inference** on the state process. However, it may be cumbersome to interpret all time-varying dwell-time distributions.
- In addition, the inferential focus concerning state dynamics will often be on the **unconditional distribution**, not explicitly conditioning on the start time of the stay.
- This **overall dwell-time distribution** can be derived as a mixture of the time-varying dwell-time distributions and is defined by its probability mass function

$$d_i(r) = \sum_{t=1}^{L} w_i^{(t)} d_i^{(t)}(r), \quad r \in \mathbb{N}, \qquad (5)$$

with the mixture weights defined as

$$w_i^{(t)} = \frac{\sum_{l \in \mathcal{S} \setminus i} \delta_l^{(t-1)} \gamma_{li}^{(t-1)}}{\sum_{t=1}^{L} \sum_{l \in \mathcal{S} \setminus i} \delta_l^{(t-1)} \gamma_{li}^{(t-1)}}, \quad t = 1, \dots, L,$$

where $\mathcal{S} = \{1, \dots, N\}$, $\boldsymbol{\Gamma}^{(0)} = \boldsymbol{\Gamma}^{(L)}$, $\boldsymbol{\delta}^{(0)} = \boldsymbol{\delta}^{(L)}$, and $\boldsymbol{\delta}^{(t)}$ as in Equation (3).

- It particularly serves as a **model-checking tool** for the state process when comparing it to the **empirical** dwell-time distribution obtained from the **decoded state sequence** (see Figure 3).

3 Application: Drosophila melanogaster

- Fruit flies have a pronounced **circadian rhythm**, and researchers are interested in its **reaction to external variation**. Therefore, flies were trained under light-dark (LD) and constant darkness (DD) conditions.
- We model the half-hourly time series ($L = 48$) using a 2-state HMM to describe a **low- and high-activity state**, where for $i \neq j$ the transition probabilities are modelled by trigonometric functions with increasing frequencies:

$$\text{logit}(\gamma_{ij}^{(t)}) = \beta_0^{(ij)} + \sum_{k=1}^{3} \beta_{1k}^{(ij)} \sin\left(\frac{2\pi k t}{48}\right) + \beta_{2k}^{(ij)} \cos\left(\frac{2\pi k t}{48}\right).$$

- The periodically stationary distribution varies substantially over the course of one day, and there is a clear **discrepancy** between the two lighting schedules (see Figure 2).

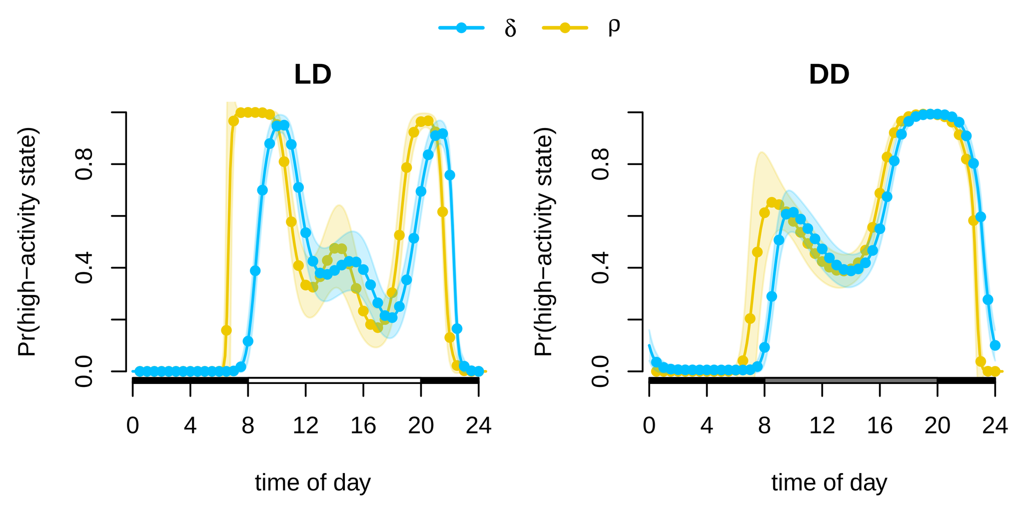

Figure 2: Periodically stationary distribution as a function of the time of day. True stationary distribution (light blue) compared to biased approximation (yellow).

- The **overall state dwell-time distribution** is non-geometric and shows bimodality, especially in the DD condition due to longer stays in the evening (see Figure 3).

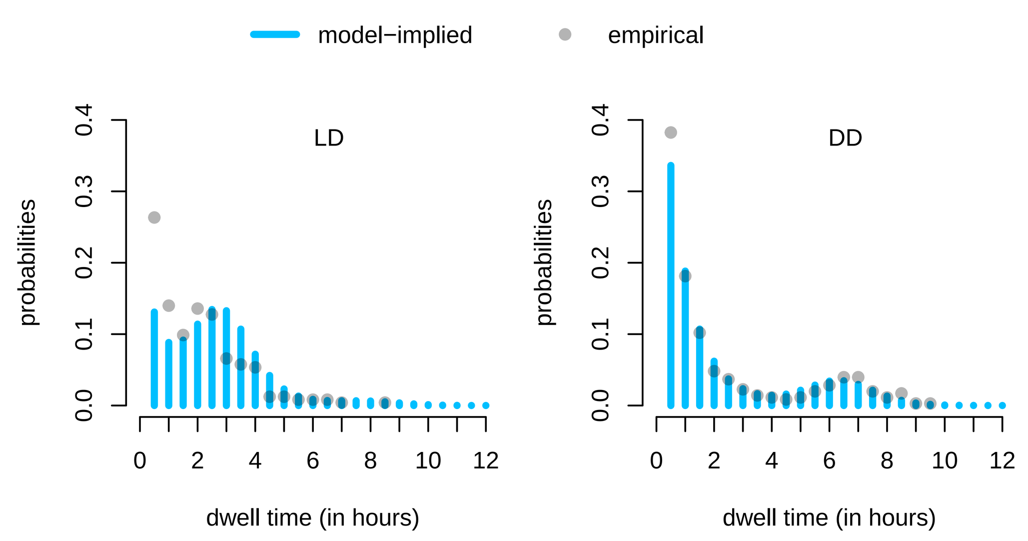

Figure 3: Overall dwell-time distribution of the high-activity state, analytically (blue bars) and empirically (grey dots) derived from the fitted HMM.

4 Discussion

- We investigated the distinct properties of periodically inhomogeneous Markov chains, allowing us to analytically derive the **periodically stationary distribution** of states and a time-varying and overall **dwell-time distribution**.
- These serve as **novel tools for inference and model checking** in scenarios characterised by periodic variation.
- When there are other, **unobserved reasons** for non-geometric state dwell-time distributions, periodically inhomogeneous HMMs are unable to describe the state process accurately.
- In such cases, **hidden semi-Markov models**, explicitly designed to model arbitrary dwell-time distributions, should be considered.

References

Ge, H., D. Jiang, and M. Qian (2006). "A Simple Discrete Model of Brownian Motors: Time-periodic Markov Chains". In: *Journal of Statistical Physics* 123.4, pp. 831–859.
Kargapolova, N. A. and V. A. Ogorodnikov (2012). "Inhomogeneous Markov chains with periodic matrices of transition probabilities and their application to simulation of meteorological processes". In: *Russian Journal of Numerical Analysis and Mathematical Modelling* 27.3, pp. 213–228.
Patterson, T. A. et al. (2009). "Classifying movement behaviour in relation to environmental conditions using hidden Markov models". In: *Journal of Animal Ecology* 78.6, pp. 1113–1123.
Touron, A. (2019). "Consistency of the maximum likelihood estimator in seasonal hidden Markov models". In: *Statistics and Computing* 29.5, pp. 1055–1075.