

# An Introduction to Hidden Markov Models

17th IACHEC Meeting  
Osaka, Japan

Robert Zimmerman

Imperial College London

May 14, 2025

# Agenda

- 1 Introduction
- 2 Ingredients: Mixture Models and Markov Chains
- 3 Hidden Markov Models
- 4 Fitting Hidden Markov Models
- 5 Decoding the State Sequence
- 6 Extensions (Time Permitting)
- 7 Summary

# Agenda

- 1 Introduction
- 2 Ingredients: Mixture Models and Markov Chains
- 3 Hidden Markov Models
- 4 Fitting Hidden Markov Models
- 5 Decoding the State Sequence
- 6 Extensions (Time Permitting)
- 7 Summary

# Hidden Processes in Real Data

- Real-world time series often exhibit abrupt or gradual changes in behavior that are driven by unobserved states (i.e., latent variables)

 **Astronomy:** Flaring and quiescence in stellar X-ray light curves

- ▶ Latent variable: flare intensity or state

 **Ecology:** Animal movement switching between foraging and resting

- ▶ Latent variable: behavioral mode

 **Finance:** Stock returns alternating between volatility regimes

- ▶ Latent variable: market state

 **Bioinformatics:** Coding vs. non-coding DNA regions

- ▶ Latent variable: genomic structure

 **Speech:** Recognizing spoken units from acoustic signals

- ▶ Latent variable: spoken unit

# Enter Hidden Markov Models

- Hidden Markov models give us a structured way to model time-dependent processes whose behavior depends on a hidden state that evolves over time

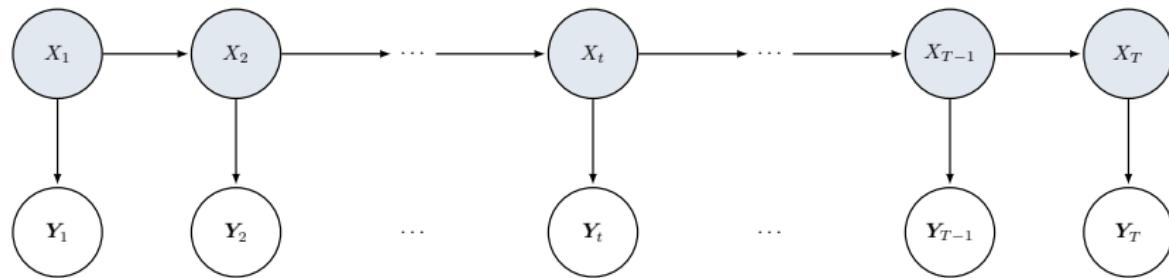
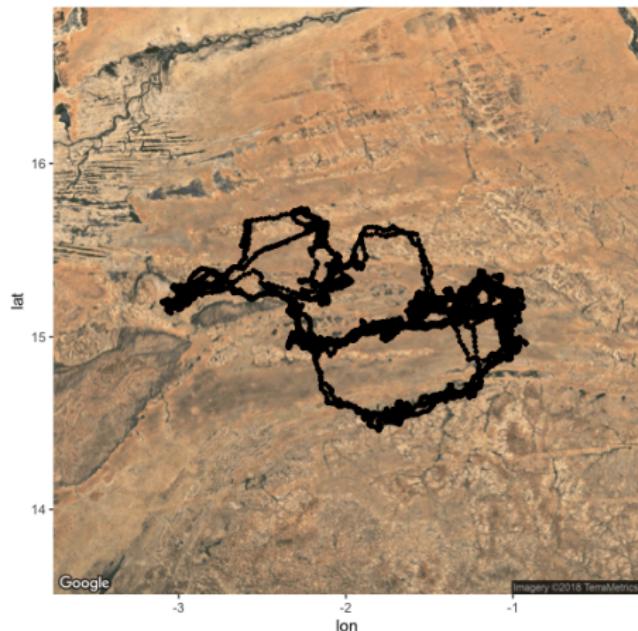


Figure: A graphical model of the standard discrete-time HMM dependence structure

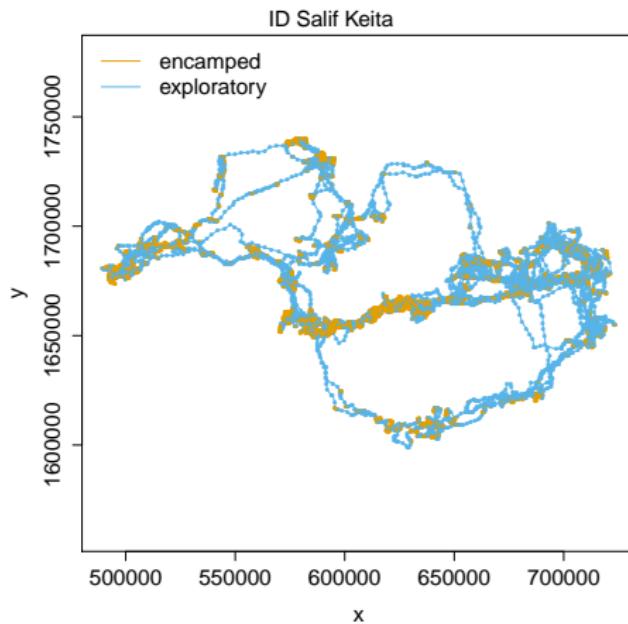
## Example: African Elephant Movement

- The figure below shows an African elephant's tracks in Mali over several days [Wall et al., 2014]
- It is believed that elephants typically spend time in either of two states: *encamped* and *exploratory*



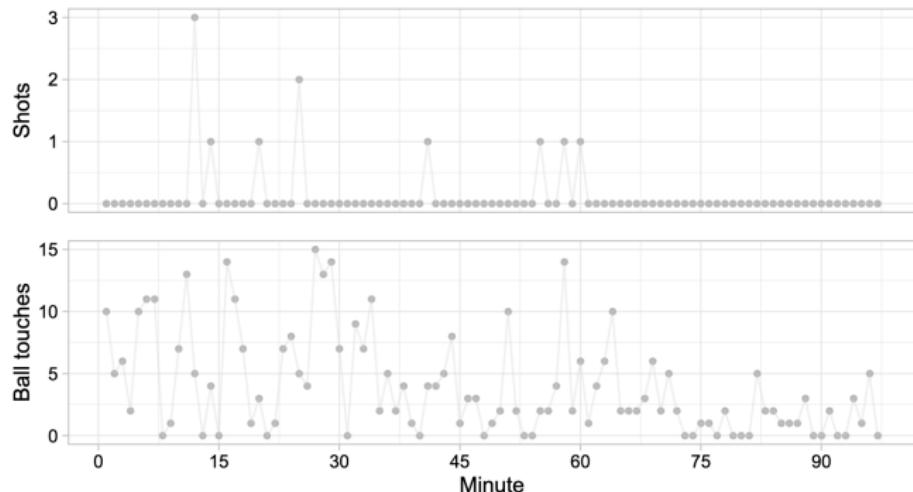
# Hidden States Revealed

- [McClintock and Michelot, 2018] fit a 2-state HMM to the observed data, allowing ecologists to classify the elephant's state at each time point and predict its future states



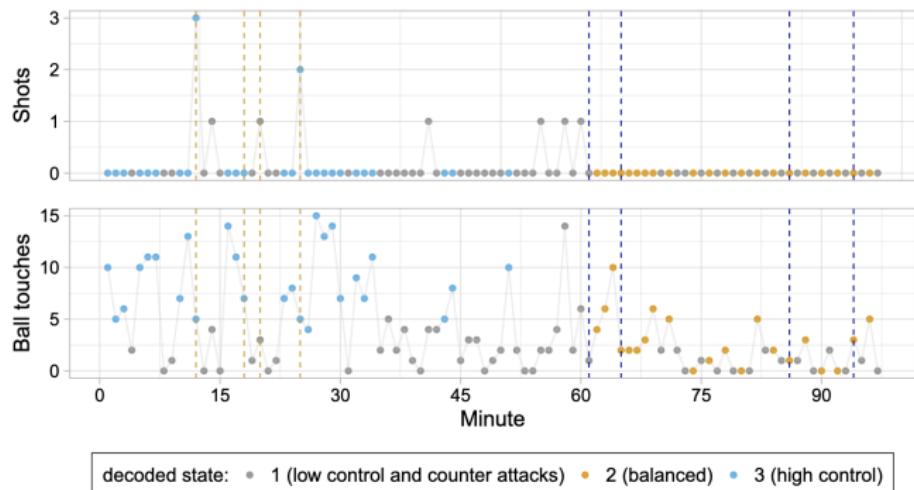
## Example: Momentum in Football (aka Soccer) Matches

- The figure below shows a bivariate time series of the number of shots on goal (top) and the ball touches (bottom) of Borussia Dortmund for a match vs. FC Schalke 04 [Ötting et al., 2023]
- We imagine three states for Borussia: *low control*, *balanced*, and *high control*



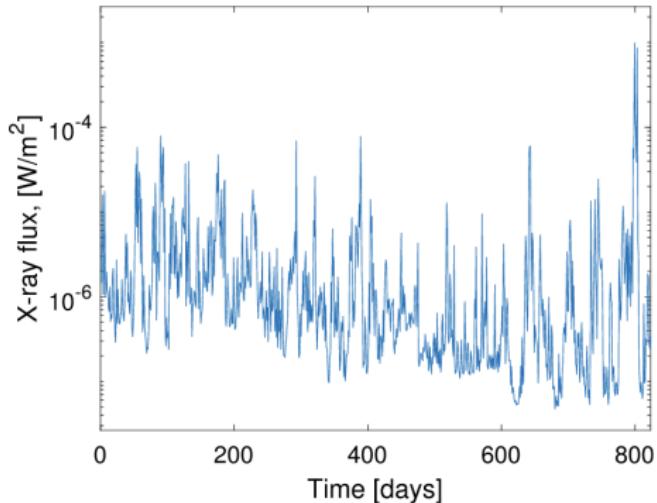
# Hidden States Revealed

- [Ötting et al., 2023] fit a 3-state HMM to the data, with state classifications shown below
- The vertical dashed lines show goals scored by Borussia (yellow lines) and Schalke 04 (blue lines)



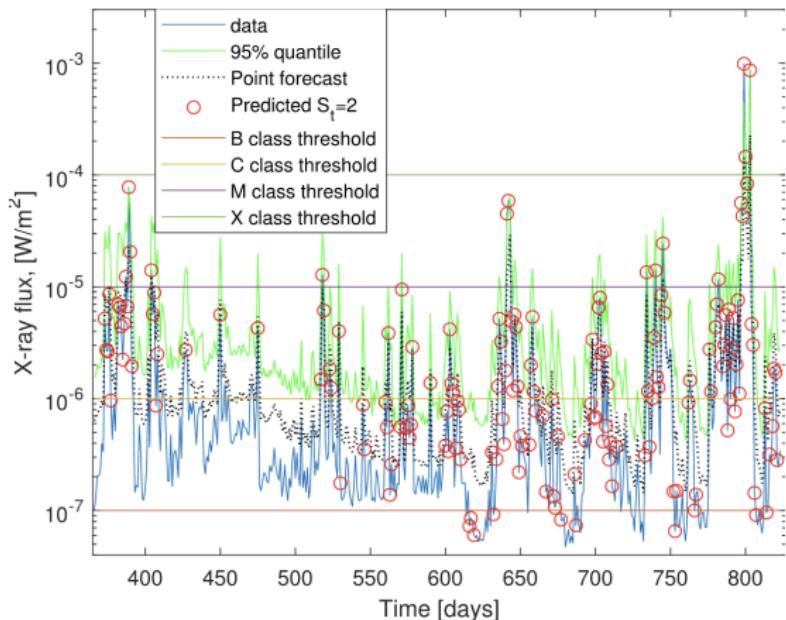
## Example: Solar Flare Activity

- The figure below shows solar X-ray log flux (from GOES data) in the period from 1 July 2015 to 30 September 2017 [[Stanislavsky et al., 2020](#)]
- They assume two states: *low activity* ("1") and *high activity* ("2")



# Hidden States Revealed... and Predicted!

- [Stanislavsky et al., 2020] fit a 2-state HMM to rolling 365 day windows of the data, and predict both the solar flux and the state for the following day



# Agenda

- 1 Introduction
- 2 Ingredients: Mixture Models and Markov Chains
- 3 Hidden Markov Models
- 4 Fitting Hidden Markov Models
- 5 Decoding the State Sequence
- 6 Extensions (Time Permitting)
- 7 Summary

# Mixture Models

- Let  $X \in \mathcal{X}$  be a random variable with pdf  $\pi(x)$  or pmf  $\pi_x = \mathbb{P}(X = x)$
- Conditional on  $X = x$ , let  $Y \in \mathcal{Y}$  be a random variable with pdf/pmf  $f_x(y)$
- The *unconditional* pdf/pmf of  $Y$  is given by

$$f(y) = \int_{\mathcal{X}} \pi(x) \cdot f_x(y) dx \quad \text{or} \quad f(y) = \sum_{x \in \mathcal{X}} \pi_x \cdot f_x(y)$$

and  $Y$  is said to follow a **mixture model**

- Mixture models have a simple design that can accommodate unobserved heterogeneity in a population
- They are often used to handle multi-modal distributions

## Special Case: Finite Mixture Models

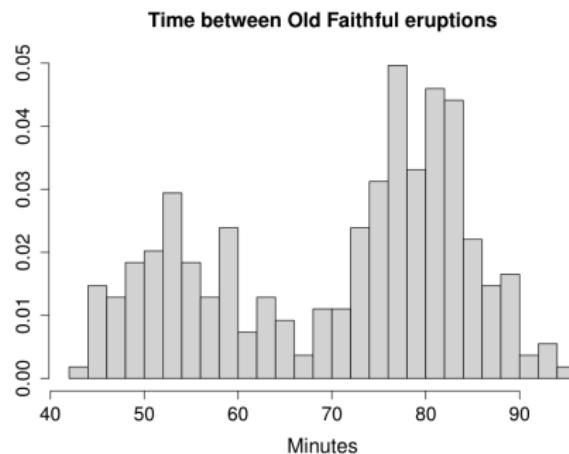
- When  $\mathcal{X} = \{1, 2, \dots, K\}$ , we have a  **$K$ -component finite mixture model** with pdf/pmf

$$f(y) = \sum_{k=1}^K \pi_k \cdot f_k(y)$$

- Note: in general, each  $f_x(y)$  can — and usually does — have an associated vector of parameters  $\theta_x$  that varies with  $x$
- We often write  $f_x(y; \theta_x)$  to emphasize dependence on the state-dependent parameter  $\theta_x$

## Example: Time Between Old Faithful Eruptions

- The figure below shows a histogram of time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA  
[Azzalini and Bowman, 1990]
- The observations seem to include two distinct components
- Histograms like this are highly characteristic of finite mixture models



# Maximum Likelihood for Finite Mixture Models

- Given an independent sample  $y_1, \dots, y_n \stackrel{iid}{\sim} f$ , the likelihood function is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\pi} \mid y_{1:n}) = \prod_{i=1}^n \left( \sum_{k=1}^K \pi_k \cdot f_k(y_i; \boldsymbol{\theta}_k) \right)$$

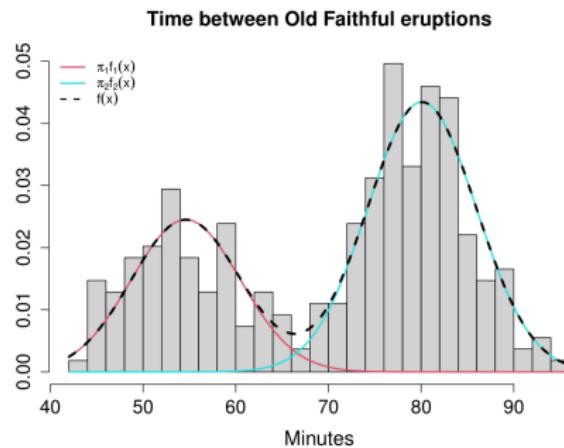
with  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$  and  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$

- ...and the log-likelihood by  $\ell(\boldsymbol{\theta}, \boldsymbol{\pi} \mid y_{1:n}) = \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k \cdot f_k(y_i; \boldsymbol{\theta}_k) \right)$

- Numerical maximization (or often the *EM algorithm*) can be used to obtain the MLEs of  $\boldsymbol{\theta}$  and  $\boldsymbol{\pi}$
- If some/all  $f_k$  are in the same parametric family, it is good practice to somehow (e.g., by imposing order constraints) identify the parameters of the model to prevent label switching

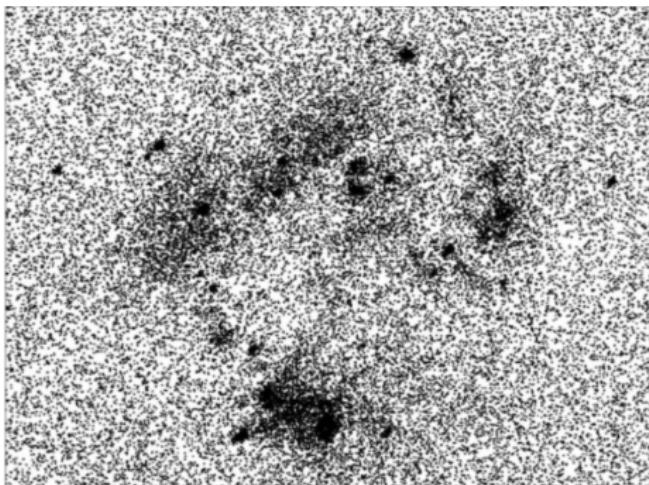
## Back to Old Faithful

- Suppose we assume a 2-component Gaussian mixture model (i.e.,  $K = 2$  and each  $f_k$  is a univariate Gaussian pdf)
- If we perform maximum likelihood estimation, we get that
  - $f_1(y)$  is estimated to be  $\mathcal{N}(54.6, 5.9^2)$
  - $f_2(y)$  is estimated to be  $\mathcal{N}(80.1, 5.9^2)$
  - $\pi_1$  is estimated to be 0.36 (thus  $\pi_2$  is estimated as 0.64).



# Finite Mixture Models in Astronomy: Stellar Populations

- Astronomical populations often consist of overlapping groups (e.g., stars in different evolutionary phases)
- Finite mixture models help disentangle these subpopulations using photometric data [Fan et al., 2023]



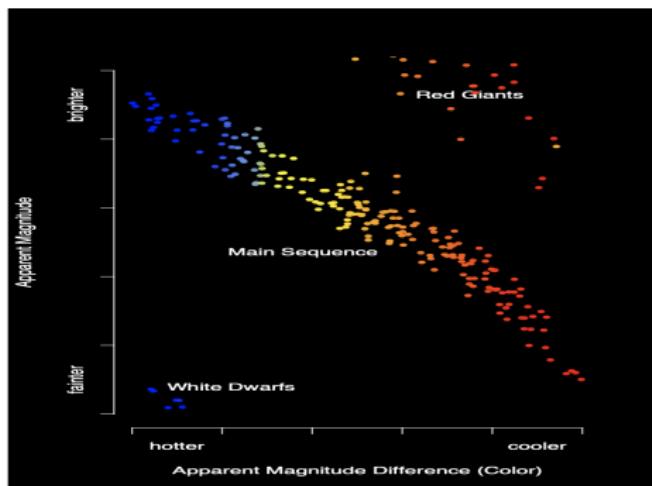
Chandra X-ray observations of colliding Antennae galaxies; the source appears over a diffuse background



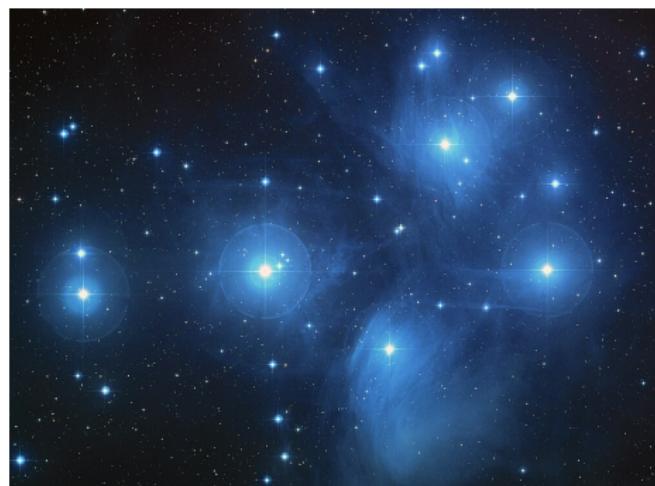
Hubble optical image of colliding Antennae galaxies; emission sources are spatially structured (image credit: [NASA, ESA, and the Hubble Heritage Team](#))

# Finite Mixture Models in Astronomy: Source Separation

- Finite mixture models group spatial or photometric patterns
- We will see that HMMs extend this idea to sequences, where latent group membership evolves over time



Stylized color-magnitude diagram; mixture components reflect stellar evolution stages



Hubble optical image of the Pleiades cluster; mixture models separate cluster members from the background (image credit: [NASA, ESA and AURA/Caltech](#))

# Markov Chains

- A **discrete time Markov chain** on  $\mathcal{X}$  is an  $\mathcal{X}$ -valued stochastic process  $\{X_t\}^1$  that satisfies the Markov property:

$$\mathbb{P}(X_{t+1} \in A \mid X_t = x_t, \dots, X_1 = x_1) = \mathbb{P}(X_{t+1} \in A \mid X_t = x_t)$$

for  $A \subseteq \mathcal{X}$  and  $t \geq 0$

- ▶ i.e., the distribution of  $X_{t+1}$  is entirely determined by  $X_t$

- A discrete time Markov chain on  $\mathcal{X}$  is fully characterized by
  - ① An initial pdf  $\delta(x)$  or pmf  $\delta_x = \mathbb{P}(X_0 = x)$  that determines the distribution of  $X_0$
  - ② A transition pdf  $\gamma^{(t)}(x_{t-1}, x)$  or pmf  $\gamma_{x_{t-1}, x}^{(t)} = \mathbb{P}(X_t = x \mid X_{t-1} = x_{t-1})$  that determines the conditional distribution of  $X_t$  given  $X_{t-1} = x_{t-1}$
- If the transition pdf/pmf does not depend on  $t$ , then the chain is said to be **time-homogeneous**

---

<sup>1</sup>Notation:  $\{X_t\}$  means the (possibly infinite) sequence  $X_0, X_1, X_2, \dots$

## Finite Space Markov Chains: Transition Probabilities

- An important special case is a time-homogeneous Markov chain on  $\mathcal{X} = \{1, 2, \dots, K\}$
- Here, the transition probability  $\gamma_{i,j}$  (no superscript!) is the probability that the chain enters state  $j$  at time  $t+1$  given that it is in state  $i$  at time  $t$
- We can collect the  $K^2$  transition probabilities into a **transition probability matrix**

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{1,1} & \cdots & \gamma_{1,K} \\ \vdots & \ddots & \vdots \\ \gamma_{K,1} & \cdots & \gamma_{K,K} \end{pmatrix}$$

- One can show that unconditional probability  $\mathbb{P}(X_t = k)$  is given by the  $k$ th entry of  $\boldsymbol{\delta}\boldsymbol{\Gamma}^t$ , where  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_K)$

# Markov Chains: Stationary and Limiting Distributions

- A Markov chain has a **limiting distribution** if the distribution of  $X_t$  (starting from any initial distribution) exists as  $t \rightarrow \infty$
- A time-homogeneous Markov chain is said to have a **stationary distribution** if there exists a pdf  $s(x)$  or a pmf  $s_x$  which satisfies

$$\int_{\mathcal{X}} s(x) \cdot \gamma(x, x') dx = s(x') \quad \text{or} \quad \sum_{x \in \mathcal{X}} s_x \cdot \gamma_{x,x'} = s_{x'}$$

- ▶ In the finite space case, if  $s = (s_1, \dots, s_K)$ , then the first statement is equivalent to  $s\Gamma = s$
- A stationary distribution exists under mild conditions, and when it does it is *equal to the limiting distribution* (and hence unique)

# Agenda

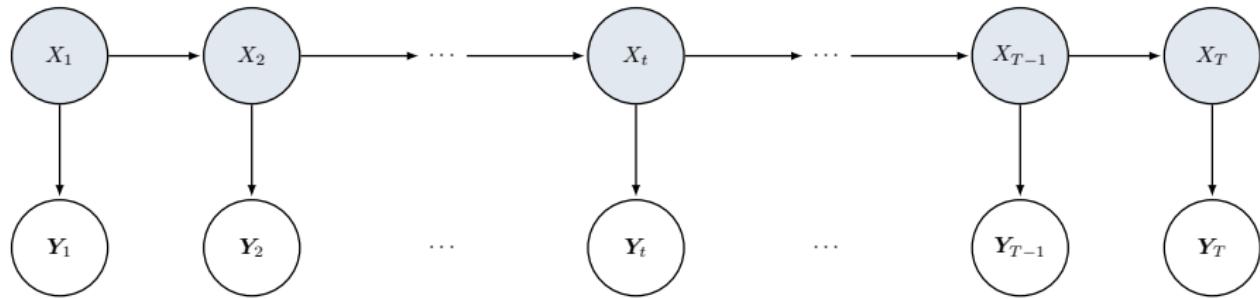
- 1 Introduction
- 2 Ingredients: Mixture Models and Markov Chains
- 3 Hidden Markov Models
- 4 Fitting Hidden Markov Models
- 5 Decoding the State Sequence
- 6 Extensions (Time Permitting)
- 7 Summary

# Serial Dependence

- We now consider an observed time series  $\{Y_t\}$
- Such time series commonly exhibit dependence between consecutive observations — a phenomenon known as **serial dependence**
- But sometimes, this serial dependence reflects a deeper structure: what if the behavior of  $Y_t$  is driven by an unobserved **state process**  $\{X_t\}$ ?
- In particular, what if...
  - ▶  $\{X_t\}$  evolves as a Markov chain, and
  - ▶ The distribution of  $Y_t$  depends on the current state  $X_t$ ? That is, the statistical properties of the observed process change over time depending on the hidden state

# Putting Things Together: the HMM

- This generative structure informally defines a hidden Markov model



- The unobserved state process  $\{X_t\}$  (shaded nodes) is a Markov chain
- The observed process  $\{Y_t\}$  (clear nodes) is conditionally independent given the states: each  $Y_t$  depends only on  $X_t$

# The HMM: General Definition

- A **discrete time hidden Markov model (HMM)** consists of...
  - ① A latent process  $\{X_t\}$  evolving as a Markov chain on some state space  $\mathcal{X}$ 
    - ★ Initial pdf/pmf  $\delta(x)$
    - ★ A transition pdf/pmf  $\gamma^{(t)}(x_{t-1}, x)$
  - ② An observation process  $\{Y_t\}$  on a space  $\mathcal{Y}$  which is conditionally independent given the states.<sup>2</sup>

$$\mathbb{P}(Y_t \in A \mid X_{1:T}, Y_{1:(t-1)}) = \mathbb{P}(Y_t \in A \mid X_t)$$

- ③ A state-dependent distribution model:

$$Y_t \mid X_t = x \sim f_x$$

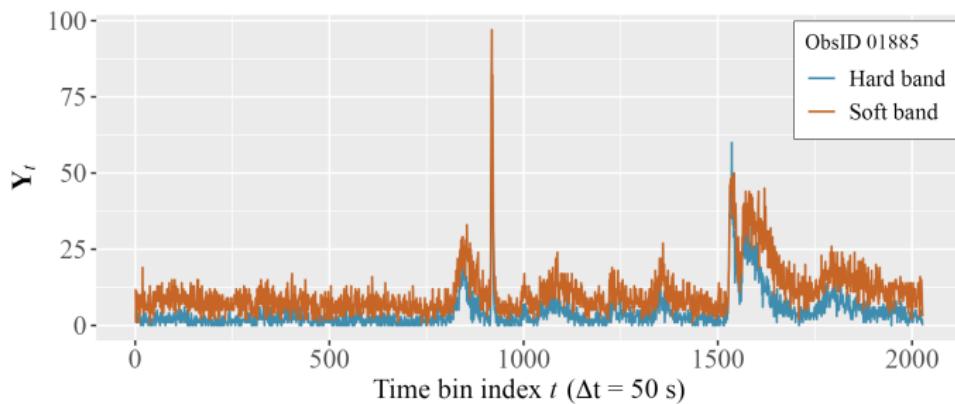
- Such an HMM is fully characterized by
  - ① The initial pdf  $\delta(x)$  or pmf  $(\delta_x)_{x \in \mathcal{X}}$
  - ② The transition pdf  $\gamma^{(t)}(x_{t-1}, x)$  or pmf  $(\gamma_{x_{t-1}, x}^{(t)})_{x \in \mathcal{X}}$
  - ③ The state-dependent pdfs/pmf  $\{f_x : x \in \mathcal{X}\}$

---

<sup>2</sup>Notation:  $X_{1:t}$  means  $(X_1, X_2, \dots, X_t)$  and similarly for  $Y_{1:t}$

## Example: Flaring Behaviour of EV Lac

- [Zimmerman et al., 2024] study X-ray light curves of the red dwarf star EV Lac
- The figure below shows photon counts in soft and hard bands for EV Lac over several days



## Example: Flaring Behaviour on EV Lac

- [Zimmerman et al., 2024] use a univariate Poisson state-space model to the capture flaring behaviour
- The latent Markov chain  $\{X_t\}$  evolves as an AR(1) process:

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

- The observed process  $\{\mathbf{Y}_t = (Y_{t,1}, Y_{t,2})\}$  is a 2-tuple of soft and hard band X-ray photon counts:

$$Y_{t,h} \mid X_t = x_t \sim \text{Poisson}(w \cdot \beta_h \cdot e^{x_t}), \quad h = 1, 2$$

- Smooth transitions in  $\{X_t\}$  capture variability in flaring activity as manifested in  $\{\mathbf{Y}_t\}$

# Agenda

- 1 Introduction
- 2 Ingredients: Mixture Models and Markov Chains
- 3 Hidden Markov Models
- 4 Fitting Hidden Markov Models
- 5 Decoding the State Sequence
- 6 Extensions (Time Permitting)
- 7 Summary

# Likelihood Functions for HMMs

- The vector of parameters  $\theta$  in an HMM include those associated with the initial pdf/pmf, the transition pdf/pmf, and the state-dependent distributions
- Suppose we observe data  $y_{1:T}$  arising from an HMM
- When  $\mathcal{X} = \{1, \dots, K\}$ , the likelihood is a sum over all possible state paths:

$$L(\theta | y_{1:T}) = \sum_{x_1=1}^K \cdots \sum_{x_T=1}^K \delta_{x_1} \cdot f_{x_1}(y_1) \prod_{t=2}^T \gamma_{x_{t-1}, x_t}^{(t)} \cdot f_{x_t}(y_t)$$

- When  $\mathcal{X} = \mathbb{R}^d$ , the sums are replaced by integrals:

$$L(\theta | y_{1:T}) = \int \cdots \int \delta(x_1) \cdot f_{x_1}(y_1) \prod_{t=2}^T \gamma^{(t)}(x_{t-1}, x_t) \cdot f_{x_t}(y_t) dx_{1:T}$$

# Fitting HMMs via Likelihood Maximization

- Once an HMM has been specified, it can be fit by maximizing the likelihood:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta \mid y_{1:T})$$

- For discrete-state HMMs, the likelihood can be computed efficiently via the **forward algorithm** in  $O(TK^2)$  time
- For continuous-space HMMs, the likelihood must be approximated numerically (e.g., via state-space discretization [[Zimmerman et al., 2024](#)] or particle methods)
- In practice, we optimize the likelihood using numerical methods (e.g., L-BFGS)
  - Transformations ensure parameters stay within valid domains (e.g., log or  $\tanh^{-1}$ )

## Model Assessment: Pseudo-Residuals

- To assess how well the fitted HMM explains observed *univariate* data, we use **pseudo-residuals**
- These are constructed from the one-step-ahead forecast distribution:

$$r_t = \Phi^{-1}(\mathbb{P}(Y_t \leq y \mid Y_{1:t-1})), \quad t = 2, 3, \dots, T$$

The cdf above can either be computed exactly or estimated, depending on the type of HMM

- Under a well-specified model,  $r_2, \dots, r_T$  should be approximately  $\mathcal{N}(0, 1)$ 
  - ▶ Deviations reveal distributional misfit or unmodeled serial dependence

# Agenda

- 1 Introduction
- 2 Ingredients: Mixture Models and Markov Chains
- 3 Hidden Markov Models
- 4 Fitting Hidden Markov Models
- 5 Decoding the State Sequence
- 6 Extensions (Time Permitting)
- 7 Summary

# State Decoding: Inferring the Latent Process

- Once we've fit the HMM using an estimator  $\hat{\theta}$ , we can recover information about the hidden states  $\{X_t\}$  using one of two common approaches:
- For discrete-space HMMs: **local decoding**

$$\hat{X}_t = \operatorname{argmax}_{x \in \mathcal{X}} \mathbb{P}_{\hat{\theta}}(X_t = x \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}), \quad t = 1, \dots, T$$

or **global decoding**

$$\hat{X}_{1:T} = \operatorname{argmax}_{x_{1:T} \in \mathcal{X}^T} \mathbb{P}_{\hat{\theta}}(X_{1:T} = x_{1:T} \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$$

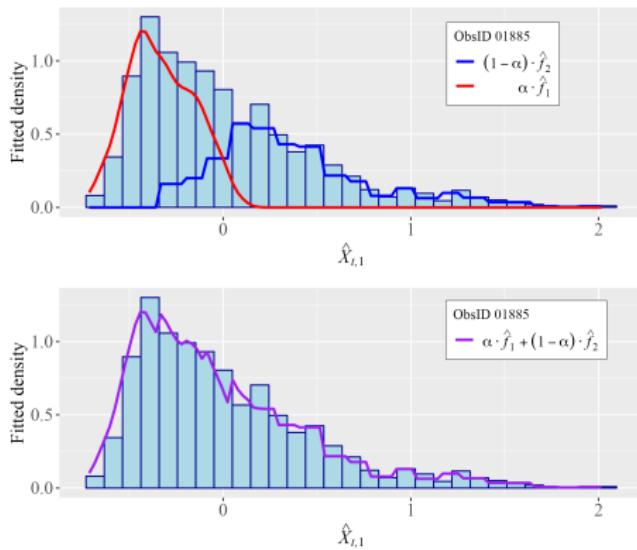
These both require the filtered state probabilities  $\mathbb{P}_{\hat{\theta}}(X_t = x \mid \mathbf{Y}_{1:t} = \mathbf{y}_{1:t})$ , which can be computed efficiently using the forward algorithm

- For continuous-space HMMs: **posterior expectation**

$$\hat{X}_t = \mathbb{E}_{\hat{\theta}}[X_t \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}], \quad t = 1, \dots, T$$

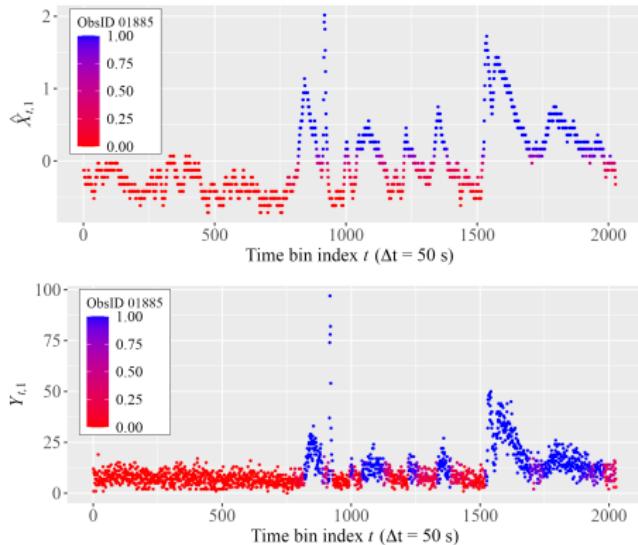
## Back to EV Lac

- In the EV Lac model, we compute the smoothed posterior means  $\{\hat{X}_t\}$  to estimate the underlying flare intensity at each time point
- We then fit a 2-component mixture model to the distribution of  $\{\hat{X}_t\}$ :



# EV Lac: Flaring and Quiescence

- The fitted mixture model above allows us to estimate the “probability” of flaring for each observation:



# Forecasting States Ahead in Time

- Consider the discrete-space HMM and suppose we've made state predictions by computing the filtered state probabilities  $\mathbb{P}_{\hat{\theta}}(X_t = x \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$
- We can forecast future states conditional on the observed data  $\mathbf{Y}_{1:T}$  practically for free:

$$\hat{X}_{T+t} = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \sum_{k=1}^K \mathbb{P}_{\hat{\theta}}(X_t = k \mid \mathbf{Y}_{1:T} = \mathbf{y}_{1:T}) \cdot [\hat{\Gamma}^t]_{k,x}, \quad t = 1, 2, \dots$$

where  $\hat{\Gamma}$  is the fitted transition probability matrix

- **BUT**: as  $t \rightarrow \infty$ , the forecast distribution converges to the stationary distribution (regardless of history)
  - ▶ So predictive uncertainty increases with  $t$ : farther-out forecasts are more diffuse and less informative

# Agenda

- 1 Introduction
- 2 Ingredients: Mixture Models and Markov Chains
- 3 Hidden Markov Models
- 4 Fitting Hidden Markov Models
- 5 Decoding the State Sequence
- 6 Extensions (Time Permitting)
- 7 Summary

## Covariates in State Dependent Distributions

- The basic HMM may be too simplistic a model for certain applications
- Occasionally, we might want certain parameters in the model to depend on covariates (for example, an animal's sex, weight, age, etc.)
- For example, the state-dependent mean  $\theta_x$  might depend linearly on some fixed vector  $\mathbf{z} \in \mathbb{R}^p$ , perhaps through some link function  $g$  :

$$g(\theta_x) = g(\mathbb{E}[Y_t | X_t = x]) = \boldsymbol{\beta}_x^\top \mathbf{z},$$

where  $\boldsymbol{\beta}_x^\top = (\beta_{x,1}, \dots, \beta_{x,p})$  is a vector of regression coefficients

- In other words, each state-dependent distribution carries its own generalized linear model

## Mixed HMMs

- We may have *multiple* time series — say  $S$  of them — available for inference
- When the time series are believed to be iid, they can be pooled together in a straightforward manner
- More realistically, the  $S$  time series are not iid, but still arise from HMMs with common features (such as the same underlying set of states  $\mathcal{X}$ )
- When the time series arise from the same parametric model (but with series-specific parameters), there can be up to  $S \cdot \text{length}(\theta)$  parameters to estimate, which is cumbersome
- For example, there would be  $S$  state-dependent parameters for state  $j$ :  
 $\theta_{j,1}, \dots, \theta_{j,S}$

## Random Effects

- Instead, one could regard the  $\theta_{j,s}$  as continuous random variables:  
 $\theta_{j,1}, \dots, \theta_{j,S} \stackrel{iid}{\sim} g_{\eta_j}$
- That is, each  $\theta_{j,s}$  is a *random effect* with distribution  $g_{\eta_j}$
- Each inclusion of such a random effect in the model reduces the number of parameters to estimate by  $S - \text{length}(\sigma_j)$
- The drawback, however, is that the  $\theta_{j,s}$  must be integrated out of the likelihood:

$$L(\dots, \eta_j \mid \mathbf{y}_{1:T}) = \int \cdots \int L(\dots, \theta_{j,1}, \dots, \theta_{j,S} \mid \mathbf{y}_{1:T}) \prod_{s=1}^S (g_{\eta_j}(\theta_{j,s}) d\theta_{j,s})$$

## Discrete Random Effects

- Even for the simplest distributions  $g_{\eta_j}$ , such integrals are never available in closed form and must be computed numerically (which is difficult in high dimensions)
- Alternatively, one can assume the  $\theta_{j,s}$  to be *discrete* random variables on a finite sample space  $\mathcal{M}$
- This makes for a simpler likelihood computation:

$$L(\dots, \boldsymbol{\eta}_j \mid \mathbf{y}_{1:T}) = \sum_{s=1}^S \sum_{m \in \mathcal{M}} L(\dots, \theta_{j,1}, \dots, \theta_{j,S} \mid \mathbf{y}_{1:T}) \cdot \mathbb{P}_{\boldsymbol{\eta}_j}(\theta_{j,s} = m)$$

- However, the applicability of such models may be limited
- The same ideas can be extended to dependent random effects, in which two or more parameters in the model follow a joint distribution

# Covariates in Transition Probabilities

- Alternatively, we may incorporate covariates into the transition pdf/pmf
- In the discrete-state case, this is typically accomplished by applying a multinomial logistic regression model to each row of the transition matrix:

$$\gamma_{j,x} = \mathbb{P}(X_t = x \mid X_{t-1} = j) = \frac{e^{\beta_{x|j}^\top \mathbf{z}}}{1 + \sum_{k=1}^{K-1} e^{\beta_{k|j}^\top \mathbf{z}}}, \quad x, j \in \mathcal{X}$$

with  $\beta_{K|j} = \mathbf{0}$  for all  $j \in \mathcal{X}$

## More on Covariates

- In either case, the  $\beta_x$  and/or  $\beta_{x|j}$  are incorporated into the likelihood function and inference proceeds as usual
- We might also want to include covariates  $\mathbf{z}_t$  that depend on time (for example,  $\mathbf{z}_t$  could include the number of hours an animal has been awake at time  $t$ )
- In this case, inference proceeds in a similar fashion; however...
- Including time-varying covariates in the transition probabilities  $\gamma_{j,x}$  destroys the assumption of time-homogeneity, so the initial pmf  $\delta_x = \mathbb{P}(X_0 = x)$  must also be estimated

# Bayesian Inference

- One can also perform Bayesian inference on HMMs
- To do so, one must choose an appropriate prior distribution  $\pi(\theta)$  for the unknown parameters of the model
- In the discrete-space case, the rows of the transition matrix  $\Gamma$  and the initial distribution vector  $\delta$  are traditionally assigned Dirichlet priors (which are conjugate to the multinomial distribution)
- Priors for the parameters  $\theta_x$  of the state-dependent distributions are chosen on a case-by-case basis

# Bayesian Inference

- The posterior distribution

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}_{1:T}) \propto \pi(\boldsymbol{\theta}) \cdot L(\boldsymbol{\theta} \mid \mathbf{y}_{1:T})$$

is never available in closed form and is impossible to sample from directly

- Thus, Markov chain Monte Carlo (MCMC) methods are typically required to sample from it
- A popular choice of MCMC method for HMMs is Hamiltonian Monte Carlo (or variants thereof), as implemented in the Stan programming language
- Although written in C++, Stan has an R interface which is accessed through the `rstan` library and a Python interface accessed through the `PyStan` library

# Quantifying Uncertainty

- As in all statistical inference, it is always of interest to quantify uncertainty in estimates of unknown parameters
- For frequentist inference, asymptotic normality of the MLE has been proven under mild regularity conditions [[Bickel et al., 1998](#)]
- The observed information matrix — which itself is a consistent estimator of the Fisher information — can be approximated numerically, and this yields standard errors and confidence intervals for parameter estimates
- In the Bayesian setup, credible intervals can be obtained from posterior distributions using standard techniques

# Agenda

- 1 Introduction
- 2 Ingredients: Mixture Models and Markov Chains
- 3 Hidden Markov Models
- 4 Fitting Hidden Markov Models
- 5 Decoding the State Sequence
- 6 Extensions (Time Permitting)
- 7 Summary

# When Are HMMs a Good Choice?

- Use an HMM when...
  - ▶ You suspect that observed temporal patterns are driven by an unobserved process with temporal structure
  - ▶ Your observed data are conditionally independent, given the hidden state
  - ▶ You want to classify, decode, or predict based on latent regimes or behaviors
- In astronomy, HMMs are useful for
  - ▶ Identifying flaring vs. quiescent periods in light curves
  - ▶ Separating source vs. background states in high-energy data
  - ▶ Modeling transitions between different emission regimes
  - ▶ [Stanislavsky et al., 2020, Zimmerman et al., 2024, Esquivel et al., 2025]
- They can be applied to counts, images, spectra, or multivariate signals
- They can be flexibly extended (e.g., to hierarchical or switching models)

# Further Resources

- Introductory and advanced textbooks:
  - ▶ [Zucchini et al., 2016]: accessible, example-driven introduction (R-based)
  - ▶ [Cappé et al., 2005]: rigorous treatment, theory-heavy (math/stats focused)
- Software for fitting HMMs:
  - ▶ In R:
    - ★ `depmixS4`: Discrete-state HMMs with Gaussian, Poisson, multinomial state-dependent distributions
    - ★ `momentuHMM`: Geared toward animal movement, but widely used in practice
    - ★ `hmmTMB`: Flexible HMMs with random effects
    - ★ `nimble`: For custom Bayesian state-space/HMM models with full MCMC
  - ▶ In Python:
    - ★ `hmmlearn`: Standard library for discrete-state HMMs (scikit-learn-like)
    - ★ `pomegranate`: Modular, faster implementation for HMMs and other probabilistic models
    - ★ `tensorflow probability`: For building custom probabilistic models (Bayesian HMMs, etc.)

Thank you!

## Download the Slides

You can download this presentation at [https://rob-zimmerman.github.io/files/presentations/HMM\\_Tutorial\\_\\_IACHEC\\_2025\\_.pdf](https://rob-zimmerman.github.io/files/presentations/HMM_Tutorial__IACHEC_2025_.pdf)



# References

- [1] Azzalini, A. and Bowman, A. W. (1990).  
A Look at Some Data on the Old Faithful Geyser.  
*Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(3):357–365.
- [2] Bickel, P. J., Ritov, Y., and Rydén, T. (1998).  
Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models.  
*The Annals of Statistics*, 26(4):1614–1635.
- [3] Cappé, O., Moulines, E., and Ryden, T. (2005).  
*Inference in Hidden Markov Models*.  
Springer Series in Statistics. Springer, New York.
- [4] Esquivel, J. A., Shen, Y., Leos-Barajas, V., Eadie, G., Speagle, J. S., Craiu, R. V., Medina, A., and Davenport, J. R. A. (2025).  
Detecting Stellar Flares in Photometric Data Using Hidden Markov Models.  
*The Astrophysical Journal*, 979(2):141.  
Publisher: The American Astronomical Society.
- [5] Fan, M., Wang, J., Kashyap, V. L., Lee, T. C. M., van Dyk, D. A., and Zezas, A. (2023).  
Identifying Diffuse Spatial Structures in High-energy Photon Lists.  
*The Astronomical Journal*, 165(2):66.  
Publisher: The American Astronomical Society.
- [6] McClintock, B. T. and Michelot, T. (2018).  
momentuHMM: R package for generalized hidden Markov models of animal movement.  
*Methods in Ecology and Evolution*, 9(6):1518–1530.  
.eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12995>.
- [7] Stanislavsky, A., Nitka, W., Malek, M., Burnecki, K., and Janczura, J. (2020).  
Prediction performance of Hidden Markov modelling for solar flares.  
*Journal of Atmospheric and Solar-Terrestrial Physics*, 208:105407.
- [8] Wall, J., Wittemyer, G., LeMay, V., Douglas-Hamilton, I., and Klinkenberg, B. (2014).  
Elliptical Time-Density model to estimate wildlife utilization distributions.  
*Methods in Ecology and Evolution*, 5(8):780–790.  
.eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12218>.
- [9] Zimmerman, R., van Dyk, D. A., Kashyap, V. L., and Siemiginowska, A. (2024).  
Separating states in astronomical sources using hidden Markov models: With a case study of flaring and quiescence on EV Lac.  
*Monthly Notices of the Royal Astronomical Society*, 534(3):2142–2167.
- [10] Zucchini, W., MacDonald, I. L., and Langrock, R. (2016).  
*Hidden Markov Models for Time Series: An Introduction Using R, Second Edition*.  
Chapman and Hall/CRC, New York, 2nd edition.
- [11] Ötting, M., Langrock, R., and Maruotti, A. (2023).  
A copula-based multivariate hidden Markov model for modelling momentum in football.  
*AStA Advances in Statistical Analysis*, 107(1):9–27.