HMMs for Complex Biologging Data

CANSSI Collaborative Research Team Project #22
Day 1

Arturo Esquivel Robert Zimmerman

University of Toronto

November 8, 2022

Agenda

Day 1

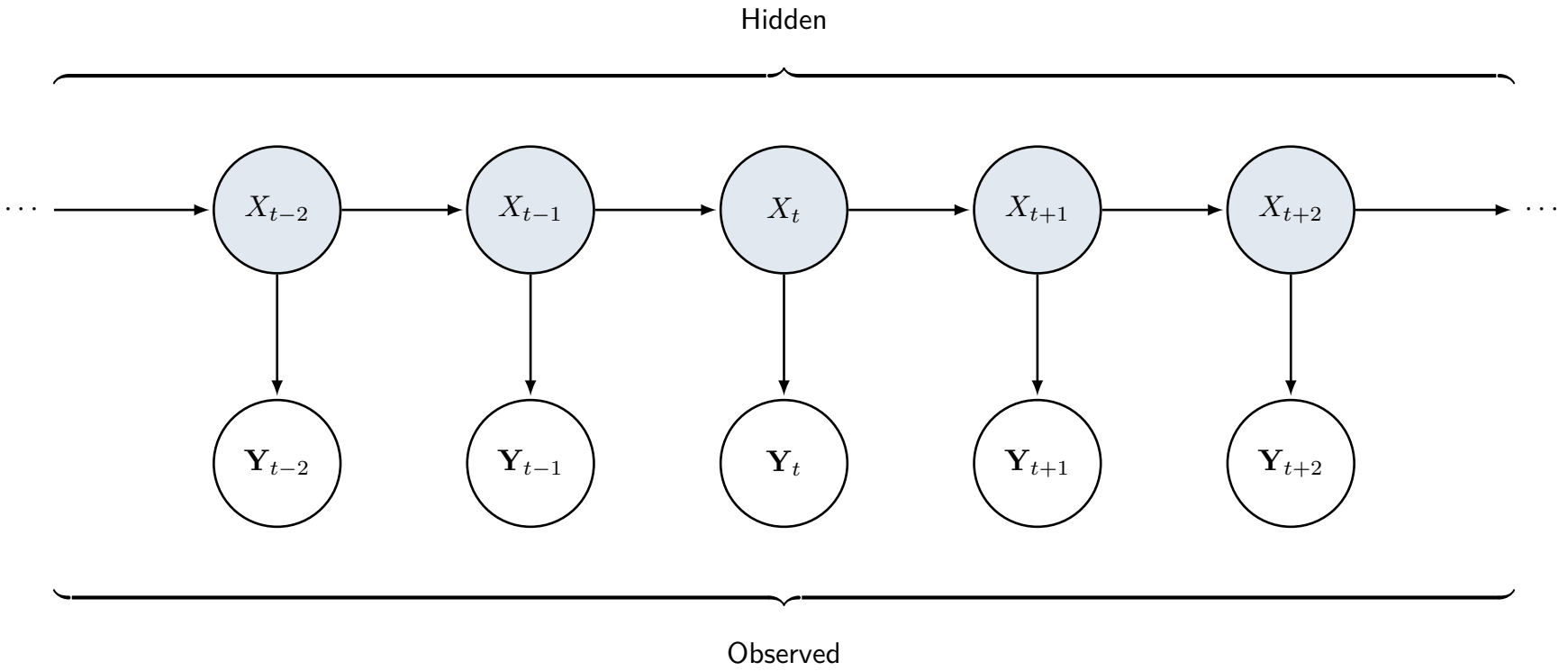
- Finite mixture models
- Markov chains
- Hidden Markov models
- Forward algorithm and likelihood computations
- Likelihood maximization

Day 2

- State decoding
- Including covariates
- Mixed HMMs and random effects
- Multivariate observations
- Bayesian inference

Hidden Markov Model (HMM) Overview

A hidden Markov model (HMM) is a classical time series model composed of two stochastic processes. An underlying, non-observable (hidden), process, the state process; and an observable process, the state-dependent process.



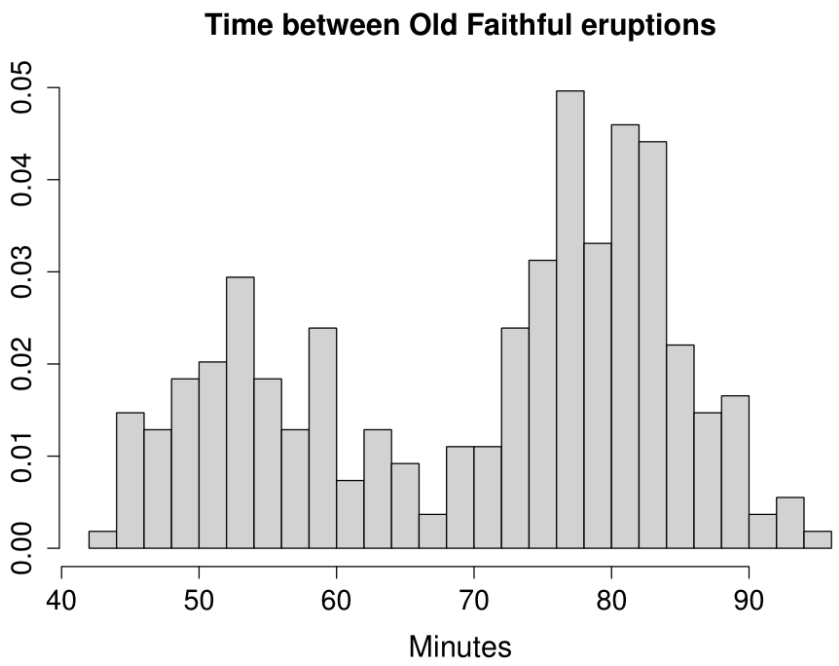
Mixture Models

- An HMM is a special type of dependent 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.

Example: Time between Old Faithful eruptions

- Waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA [Azzalini and Bowman, 1990].

- The observations seem to exhibit two patterns, arising from one of two possible distributions.



Independent Mixture Models

- Consists of a finite number of component distributions (continuous or discrete) and a random mechanism "mixing" them.
- Each realisation is assumed independent of the rest.
- A mixture comprised of K components is given by $f(y) = \sum_{j=1}^{K} \pi_j f_j(y)$
  - Where $f_j(y)$ corresponds to the j-th component distribution and π_j to the probability that distribution j is active.
  - $\pi_j \in [0, 1]$ and $\sum_{j=1}^{K} \pi_j = 1$.
- The model is fully characterized by the parameters for each of the component distributions $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and $\pi_1, \ldots, \pi_K$.

Independent Mixtures: MLE

- Maximum likelihood estimation can be used to estimate the parameters of a mixture.
- Given a sample with n observations, the likelihood function is given by

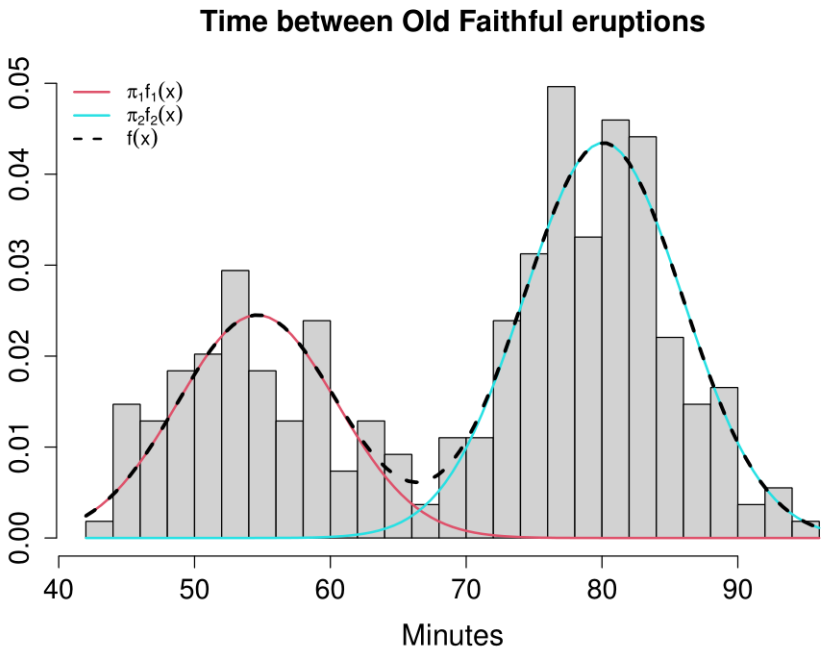
$$\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{K} \pi_j f_j(y_i) \right\}.$$

  - And the log-likelihood by $l(\boldsymbol{\Theta}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{K} \pi_j f_j(y_i) \right)$.

- Numerical maximization can be used to obtain the MLEs.
- It is good practice to somehow (e.g. imposing order constraints) identify the parameters of the model to prevent label switching.

Back to the Old Faithful

- Estimating a Gaussian mixture for the data (two Normal distributions).
  - $f_1(y)$ was estimated to be $\mathcal{N}(54.6, 5.9^2)$.
  - $f_2(y)$ was estimated to be $\mathcal{N}(80.1, 5.9^2)$.
  - π_1 was estimated to be 0.36 (π_2 0.64).



Serial Dependence

- Time series commonly show dependence between consecutive time steps.
- Dependent mixtures better accommodate system dynamics arising from serial correlation.
- In many cases it is reasonable to assume that the component distribution active at time t will more likely remain active at time $t + 1$.
- Markov chains are a natural selection to model such dependence.

Markov Chains

- A discrete time Markov chain is a stochastic process $\{X_t \in \{1, \ldots, K\}; t = 1, 2, \ldots\}$ that satisfies the Markov property

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t, \ldots, X_1 = x_1) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t).$$

  - I.e., the distribution of X_{t+1} is entirely determined by x_t.

- It is fully characterized by:
  - K, the number of states
  - $\boldsymbol{\pi}$, the initial state distribution
  - and $\gamma_{i,j}^{(t)} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$, the state transition probabilities.

Markov Chains: Transition Probabilities

- $\gamma_{i,j}^{(t)}$ is the probability that the chain enters state j at time $t+1$ given that it is in state i at time t.

- The chain is called homogeneous when

$$\gamma_{i,j}^{(t)} = \gamma_{i,j} \text{ for all } t.$$

- $\mathbf{\Gamma}$ is the transition probability matrix, given by

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

  - with $\gamma_{i,j} \in [0,1]$ for all i, j
  - and $\sum_{j=1}^{K} \gamma_{i,j} = 1$.

- Unconditional probabilities are given by

$$\mathbb{P}(X_t = x_t) = \boldsymbol{\pi}\mathbf{\Gamma}^t.$$

Markov Chains: Stationary Distribution

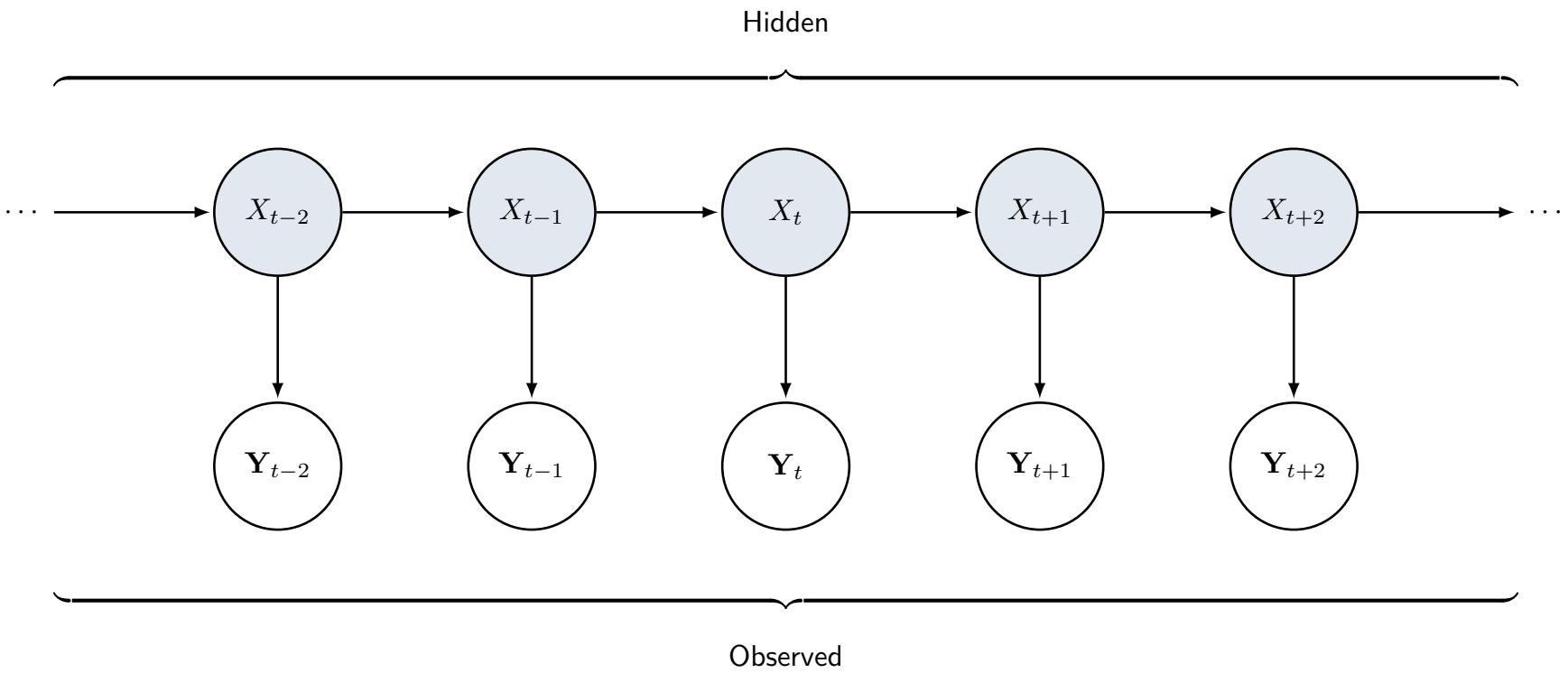
- An homogeneous Markov chain is said to have a stationary distribution $\boldsymbol{\delta}$ when

$$\boldsymbol{\delta}\boldsymbol{\Gamma} = \boldsymbol{\delta} \text{ subject to } \sum_{j=1}^{K} \delta_j = 1.$$

- Homogeneous, discrete-time, finite-space Markov chains have a unique, strictly positive, stationary distribution.

- If the chain is also aperiodic, it exists a unique limiting distribution given by the stationary distribution. [Zucchini et al., 2016]

Putting Things Together: The HMM



- The (non observable) state process is a Markov chain.
- The (observed) state-dependent process comprises a dependent mixture which random mechanism is given by the state process.

HMM: Basic Formulation

- In a K-state HMM $X_1, \ldots, X_T$ ($X_{1:T}$ for convenience) are assumed to take values across $\{1, \ldots, K\}$ and satisfy the Markov property.

- $Y_1, \ldots, Y_T$ ($Y_{1:T}$) are derived from K component distributions that become active in accordance with the state process.
  - They are assumed conditionally independent:

$$f(y_t \mid x_{1:t}, y_{1:t-1}) = f(y_t \mid x_t)$$

- Thus, an HMM can be fully characterised by:
  - K, the number of states
  - $\boldsymbol{\pi}$, the initial state distribution of the state process
  - $\boldsymbol{\Gamma}$, the transition probabilities matrix
  - and the state-dependent distributions

$$f_j(y_t) = f(y_t \mid X_t = j) = f(y_t; \boldsymbol{\theta}_j).$$

HMM: Likelihood Function

- Given a sample $y_{1:T}$ and a model consisting of an specific set of K state-dependent distributions, the likelihood of the HMM is given by
$$\mathcal{L}(\boldsymbol{\eta}) = \sum_{x_1=1}^{K} \cdots \sum_{x_T=1}^{K} f(y_{1:T}, x_{1:T}).$$
  - Where $\boldsymbol{\eta}$ includes $\boldsymbol{\pi}, \boldsymbol{\Gamma}$, and $\boldsymbol{\Theta}$.

- For any particular sequence of states $x_{1:T}$
$f(y_{1:T}, x_{1:T}) = \pi_{x_1} f_{x_1}(y_1) \gamma_{x_1,x_2} f_{x_2}(y_2) \cdots \gamma_{x_{T-1},x_T} f_{x_T}(y_T).$

- $\mathcal{L}(\boldsymbol{\eta})$ is sum of K^T terms, each of which is the product of $2T$ factors.

- Optimization can prove to be rather costly as sample size grows.

- Taking advantage of the recursive nature of the process the likelihood can be computed using just $O(TK^2)$ operations.

Forward Variables

- Consider the diagonal matrix

$$\mathbf{P}(y_t) = \begin{pmatrix} f_1(y_t) & & 0 \\ & \ddots & \\ 0 & & f_K(y_t) \end{pmatrix}.$$

- And the vector

$$\boldsymbol{\alpha}_t = \boldsymbol{\pi}\mathbf{P}(y_1)\boldsymbol{\Gamma}\mathbf{P}(y_2)\ldots\boldsymbol{\Gamma}\mathbf{P}(y_t) = \boldsymbol{\pi}\mathbf{P}(y_1)\prod_{s=2}^{t}\boldsymbol{\Gamma}\mathbf{P}(y_t)$$

- Note that the j-th element of $\boldsymbol{\alpha}_t$,

$$\alpha_{j,t} = f(y_{1:t}, X_t = j).$$

- Embedded in $\boldsymbol{\alpha}_t$ there is information on both:
  - the likelihood of $y_1, \ldots, y_t$
  - and the probability for each of the states to be active at time t.

- One can easily compute $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(y_t)$.

Forward Variables: 2-state Example

$$\boldsymbol{\alpha}_1 = \boldsymbol{\pi}\mathbf{P}(y_1) = \begin{bmatrix} \alpha_{1,1} \\ \alpha_{2,1} \end{bmatrix} = \begin{bmatrix} \pi_1 f_1(y_1) \\ \pi_2 f_2(y_1) \end{bmatrix}$$

$$\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_1 \boldsymbol{\Gamma}\mathbf{P}(y_2) = \begin{bmatrix} \alpha_{1,2} \\ \alpha_{2,2} \end{bmatrix} = \begin{bmatrix} [\pi_1 f_1(y_1)\gamma_{1,1} + \pi_2 f_2(y_1)\gamma_{2,1}] f_1(y_2) \\ [\pi_2 f_2(y_1)\gamma_{2,2} + \pi_1 f_1(y_1)\gamma_{1,2}] f_2(y_2) \end{bmatrix}$$

$$\vdots$$

$$\boldsymbol{\alpha}_T = \boldsymbol{\alpha}_{T-1} \boldsymbol{\Gamma}\mathbf{P}(y_T) = \begin{bmatrix} \alpha_{1,T} \\ \alpha_{2,T} \end{bmatrix} = \begin{bmatrix} \sum_{x_1=1}^{K} \cdots \sum_{x_{T-1}=1}^{K} f(y_{1:T}, x_{1:T-1}, X_T = 1) f_1(y_T) \\ \sum_{x_1=1}^{K} \cdots \sum_{x_{T-1}=1}^{K} f(y_{1:T}, x_{1:T-1}, X_T = 2) f_2(y_T) \end{bmatrix}$$

- It follows that $\alpha_{1,T} + \alpha_{2,T} = \mathcal{L}(\boldsymbol{\eta})$.

Forward Algorithm

- For any K,
$$\mathcal{L}(\boldsymbol{\eta}) = \sum_{j=1}^{K} \alpha_{j,T}.$$
- The forward algorithm consists of computing $\boldsymbol{\alpha}_1$, use it to compute $\boldsymbol{\alpha}_2$ and so on until $\boldsymbol{\alpha}_T$ is reached.
- Each time t the new $\boldsymbol{\alpha}_t$ is calculated, O(K^2) operations are involved.
  - Each element in the new vector (K elements) is a sum of K products. Which factors are an element from the previous vector, a transition probability and a conditional density for observation y_t.
- Hence, this way the likelihood function can be evaluated with $O(TK^2)$ operations and increasing the sample size is not overly expensive.

Maximizing the Likelihood

- Having defined the likelihood of the model and determined a feasible, efficient, way to evaluate it. Numerical optimisation can be carried out for parameter estimation.
  - The EM algorithm is also a commonly used alternative.

- There are some considerations that have to be taken when performing direct maximization of the likelihood:
  - missing data
  - numerical underflow (overflow)
  - parameter constraints
  - multiple maxima.

Dealing With Missing Data

- When it comes to missing data, HMMs work with a very simple adjustment of the likelihood.
- Suppose that for T time-steps, observations between times t and k are missing.
- It is still known that some sequence of states took the state process from x_{t-1} to x_{k+1} during $k+2-t$ transitions.
- All the possible transitions and their probabilities are contained in $\mathbf{\Gamma}^{k+2-t}$.
- Assuming the missingness is ignorable (at random) $\boldsymbol{\alpha}_T$ is given by

$$\boldsymbol{\alpha}_T = \boldsymbol{\pi}\mathbf{P}(y_1)\mathbf{\Gamma}\mathbf{P}(y_2)\dots\mathbf{\Gamma}\mathbf{P}(y_{t-1})\mathbf{\Gamma}^{k+2-t}\mathbf{P}(y_{k+1})\dots\mathbf{\Gamma}\mathbf{P}(y_T).$$

- Where $\mathbf{P}(y_t), \mathbf{P}(y_{t+1}), \dots, \mathbf{P}(y_k)$ are replaced by the identity matrix.

Underflow: Scaling the Likelihood

- Since the likelihood comprises the product of O(T) terms, the risk for it to become progressively smaller, or larger (overflow), as T grows larger should be addressed.
- Since it is a product of matrices, directly taking the log is not possible.
- An alternative is to use the scaled vector $\boldsymbol{\phi_t} = \boldsymbol{\alpha_t}/w_t$
  - with $w_t = \sum_{j=1}^{K} \alpha_{j,t}$.
- Note that $\mathcal{L}(\boldsymbol{\eta}) = \sum_{j=1}^{K} \alpha_{j,T} = w_T$ and that $w_t \boldsymbol{\phi_t} = w_{t-1} \boldsymbol{\phi}_{t-1} \boldsymbol{\Gamma} \mathbf{P}(y_t)$.
- $w_T = \prod_{t=1}^{T} (w_t / w_{t-1})$, with $w_0 = 1$.
- And the log-likelihood is given by

$$l(\boldsymbol{\eta}) = \sum_{t=1}^{T} \log (w_t / w_{t-1}).$$

Parameter Constraints

- The elements of $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ are subject to constraints.
- In particular, row sums of $\mathbf{\Gamma}$ must equal 1 and all elements must be non-negative. Constraints for $\mathbf{\Theta}$ will depend on the state-dependent distributions considered in the model (e.g. non-negative for Poisson).
- A constrained optimizer can be used. However, constrained optimization can sometimes be slow.
- An alternative is to impose the constraints by maximizing the likelihood with respect to unconstrained transformations of the parameters. Then back-transform the estimated parameters.
- E.g. $\log(\lambda)$ for Poisson and logit function for transition probabilities.

Multiple Maxima

- The likelihood functions is a complicated function of the parameters and often has multiple local maxima.
- Conditional on the starting values, maximization may reach a local maximum instead of a global one.
- A common strategy is to start from multiple, random, values and see whether the same maximum is reached.
- A maximum reached more often is more likely to be a global maximum.

Model Assessment: Pseudo-residuals

- Given that $y_{1:T}$ come from different distributions, the probability integral transformation can be used to take all of them into the same scale.
- Consider $F_{y_t}(y_t) = \mathbb{P}(Y_t \leq y_t \mid Y_{1:T} = y_{1:T})$, the probability of, under the estimated model, obtaining an observation less than or equal to y_t.
- If the model is correct, $F_{y_t}(y_t) \sim U(0, 1)$.
- Visually assessing the distribution and qq-plot of $F_{y_t}(y_t)$ can help determine if the models is valid.
- Additionally, the normal pseudo-residuals $z_t = \Phi^{-1}(F_{y_t}(y_t))$ are distributed standard normal around the median (with Φ the standard normal distribution function).
- Assessing the distribution of z_t offers further insight on the validity of the model allowing for extreme observations identification.

Thank you!

References


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.


Zucchini, W., MacDonald, I. L., and Langrock, R. (2016).
Hidden Markov Models for Time Series An Introduction Using R.
Chapman and Hall/CRC.