Hidden Markov Models: lecture 7

Bayesian analysis

Xavier Didelot

HMM definition

- A Hidden Markov Model (HMM) is a Markov chain in which the sequence of states $C_1, ..., C_T$ is not observed but hidden
- Instead of observing the sequence of states, we observe the emissions $X_1, ..., X_T$
- A HMM is defined by two quantities:
  - The transition matrix $\mathbf{\Gamma}$ of elements γ_{ij} where i and j are states:

$$\gamma_{ij} = p(C_t = j | C_{t-1} = i)$$

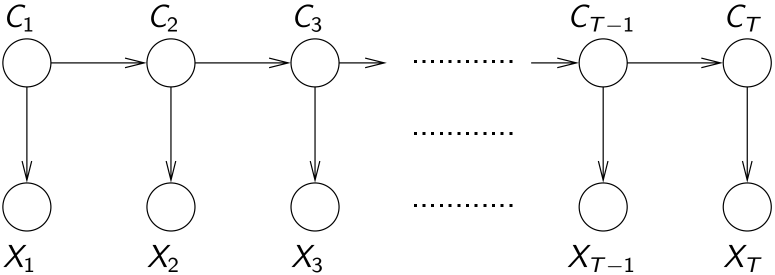
  - The emission probabilities $p_i(x)$ where i is a state and x is an emission:

$$p_i(x) = p(X_t = x | C_t = i)$$

- The unconditional distribution at t is denoted $\boldsymbol{u}(t)$ and the initial distribution is $\boldsymbol{u}(1)$

$$\boldsymbol{u}(t) = (p(C_t = 1), p(C_t = 2), ..., p(C_t = m))$$

Dependency graph of a hidden Markov model



$$p(\boldsymbol{X}^{(T)}, \boldsymbol{C}^{(T)}) = p(C_1) \prod_{k=2}^{T} p(C_k | C_{k-1}) \prod_{k=1}^{T} p(X_k | C_k)$$

$$p(\boldsymbol{x}^{(T)}, \boldsymbol{c}^{(T)}) = u_{c_1}(1) \prod_{k=2}^{T} \gamma_{c_{k-1} c_k} \prod_{k=1}^{T} p_{c_k}(x_k)$$

Bayesian inference

- Observed data x
- Parameter θ
- Likelihood function $p(x|\theta)$
- Prior distribution $p(\theta)$
- Posterior distribution $p(\theta|x)$
- Bayes Rule:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

Example

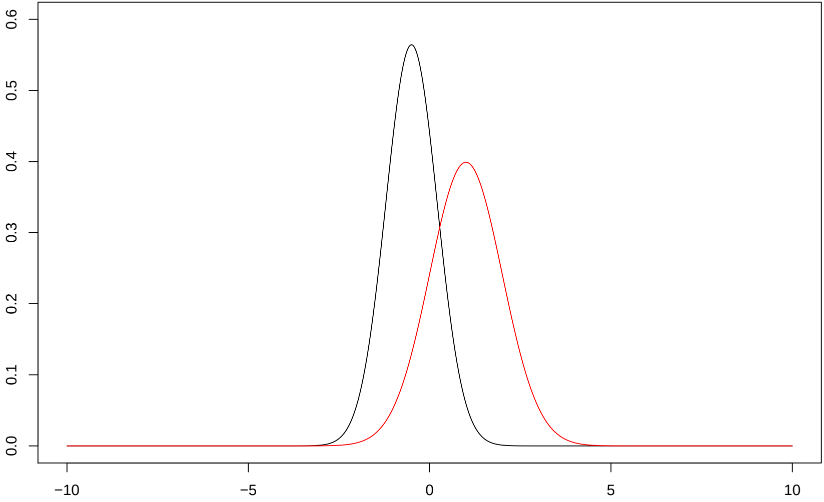
- Prior: $\theta \sim \mathrm{Normal}(1, 1)$
- Likelihood: $x \sim \mathrm{Normal}(\theta, 1)$
- Posterior:

$$
\begin{aligned}
p(\theta|x) &\propto p(\theta)p(x|\theta) \\
&\propto \exp\left(-\frac{(\theta-1)^2}{2}\right)\exp\left(-\frac{(x-\theta)^2}{2}\right) \\
&\propto \exp\left(-\frac{\theta^2+1-2\theta+x^2+\theta^2-2x\theta}{2}\right) \\
&\propto \exp\left(-\left(\theta-\frac{x+1}{2}\right)^2\right)
\end{aligned}
$$

- So that: $\theta|x \sim \mathrm{Normal}(\frac{x+1}{2}, \frac{1}{2})$

Example

- Prior: $\theta \sim \mathrm{Normal}(1, 1)$
- Likelihood: $x \sim \mathrm{Normal}(\theta, 1)$
- Observed value: $x = -2$



Monte-Carlo methods

- Computational algorithms that rely on random samples from the posterior to compute their results
- For example, to compute the expectation of the posterior distribution:

$$
\begin{aligned}
\hat{\theta} &= \int \theta p(\theta|x) d\theta \\
&\approx \frac{1}{n} \sum_{i=1}^{n} \theta_i \text{ with } \theta_i \sim p(\theta|x)
\end{aligned}
$$

- Also called a Monte-Carlo approximation or Monte-Carlo integration
- Pioneered by John von Neumann in the 1940s
- Became increasingly important as computer power increased

Markov Chain Monte Carlo

- The idea: we do not need the θ_i to be independently and identically distributed from $p(\theta|x)$
- Instead they could come from a Markov chain with stationary distribution $p(\theta|x)$
- During WW2, Metropolis and Ulam worked as part of the Manhattan project
- MCMC first published in Metropolis and Ulam (1949)
- MCMC was made popular by Gelfand and Smith (1990)

Metropolis-Hastings algorithm

- The Metropolis-Hastings (MH) algorithm was first described by Hastings (1970)
- It is a generalisation of the algorithm of Metropolis et al (1953)
- The MH algorithm produces a Markov chain $\theta_1, \theta_2, \ldots$
- At each step, the new value θ_{i+1} is generated from the previous value θ_i as follows:
  - Draw θ' from the proposal distribution $q(\theta'|\theta_i)$
  - Set $\theta_{i+1} = \theta'$ with probability $\min\left(1, \frac{p(\theta'|x)q(\theta_i|\theta')}{p(\theta_i|x)q(\theta'|\theta_i)}\right)$
  - Otherwise set $\theta_{i+1} = \theta_i$
- This Markov chain has for stationary distribution $p(\theta|x)$ (under some mild conditions...)

Detailed balance

- The MH algorithm creates a chain that satifies detailed balance, ie:

$$p(\theta_1|x)p(\theta_1 \to \theta_2) = p(\theta_2|x)p(\theta_2 \to \theta_1)$$

- Consider the case where:

$$\alpha = \frac{p(\theta_2|x)q(\theta_1|\theta_2)}{p(\theta_1|x)q(\theta_2|\theta_1)} > 1$$

- We have:

$$p(\theta_1 \to \theta_2) = q(\theta_2|\theta_1)\mathrm{min}(1, \alpha) = q(\theta_2|\theta_1) \text{ and}$$

$$p(\theta_2 \to \theta_1) = q(\theta_1|\theta_2)\mathrm{min}(1, \alpha^{-1}) = \frac{q(\theta_1|\theta_2)}{\alpha} = \frac{p(\theta_1|x)q(\theta_2|\theta_1)}{p(\theta_2|x)}$$

- Detailed balance is guaranteed, and likewise if $\alpha < 1$

Metropolis-Hastings algorithm applied to HMM

- To apply the MH algorithm, we need to calculate the ratio of posterior distributions which is equal to the ratio of likelihoods times prior:

$$\frac{p(\theta'|x)}{p(\theta|x)} = \frac{p(x|\theta')}{p(x|\theta)} \frac{p(\theta')}{p(\theta)}$$

- So we need to calculate the likelihood, which we can do for a HMM using the forward algorithm
- In particular, if the proposal distribution q is symmetric, then $q(\theta'|\theta) = q(\theta|\theta')$ and the acceptance ratio reduces to the posterior ratio
- However, such a strategy can lead to a high rejection rate since proposals are essentially random
- Such a MCMC is called sticky and will need to be run for many iterations before converging and fully exploring the posterior distribution
- There is a better strategy for HMM to avoid rejections. . .

Gibbs sampler

- The Gibbs sampler is a special case of MH, first described by Geman and Geman (1984)
- Consider that the parameter is made of two components: $\theta = \{\alpha, \beta\}$
- To update θ we can propose to update α while keeping β fixed, and then update β while keeping α fixed
- In a Gibbs sampler, we use as proposals the conditional distributions given data and other parameters, ie: $q(\alpha) = p(\alpha|\beta, x)$ and $q(\beta) = p(\beta|\alpha, x)$
- In this case, both moves are accepted with probability one, for example for the α move:

$$
\begin{aligned}
\frac{p(\theta'|x)q(\theta|\theta')}{p(\theta|x)q(\theta'|\theta)} &= \frac{p(\alpha', \beta|x)p(\alpha|\beta, x)}{p(\alpha, \beta|x)p(\alpha'|\beta, x)} \\
&= \frac{p(\beta|x)p(\alpha'|\beta, x)p(\alpha|\beta, x)}{p(\beta|x)p(\alpha|\beta, x)p(\alpha'|\beta, x)} = 1
\end{aligned}
$$

Gibbs sampler applied to HMM

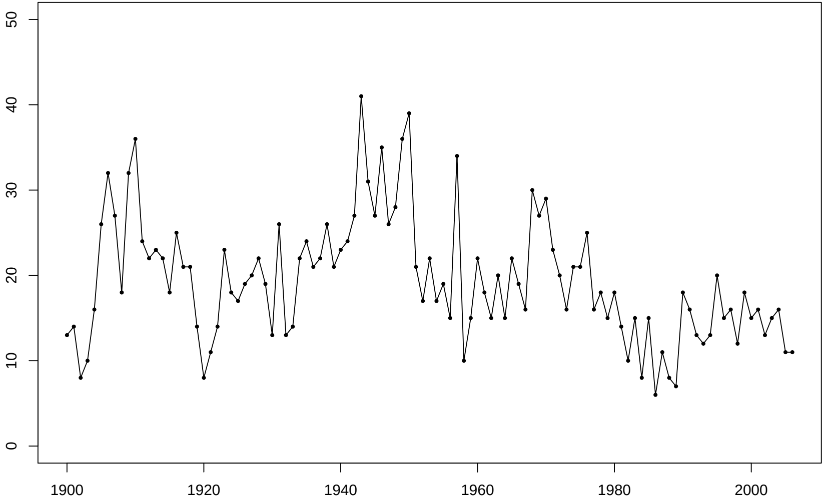
- To apply the Gibbs sampler to HMM, we alternate between two steps
- Step 1. Updating the parameters θ of the HMM given the observed emissions $\boldsymbol{x}^{(T)}$ and a sample of the hidden path $\boldsymbol{c}^{(T)}$
- Step 2. Updating the sample of the hidden path $\boldsymbol{x}^{(T)}$ given the parameters θ of the HMM and the observed emissions $\boldsymbol{x}^{(T)}$
- Here θ represents the transition probabilities contained in $\boldsymbol{\Gamma}$ as well as emission parameters contained in $\boldsymbol{P}$
- This is the Bayesian equivalent to the Baum-Welch algorithm (cf lecture 5)
- Step 1 is relatively easy and not specific to HMM since the path is known
- Step 2 can be done using a modified version of the forward-backward algorithm

Generating a sample path of the HMM

- First run the forward algorithm to compute the forward probabilities:
$$\alpha_t(i) = p(\boldsymbol{x}^{(t)}, C_t = i)$$
- Sample C_T, the state of the HMM at the final position T from
$$p(C_T = i|\boldsymbol{x}^{(T)}) \propto \alpha_T(i)$$
- Then simulate all previous states by going backwards from $T-1$ to 1, each time sampling from:

$$\begin{aligned}
p(C_t = i|\boldsymbol{x}^{(T)}, C_{t+1} = j) &= p(C_t = i|\boldsymbol{x}^{(t)}, C_{t+1} = j) \\
&\propto p(C_t = i, \boldsymbol{x}^{(t)}, C_{t+1} = j) \\
&\propto p(C_t = i, \boldsymbol{x}^{(t)})p(C_{t+1} = j|C_t = i) \\
&\propto \alpha_t(i)\gamma_{ij}
\end{aligned}$$

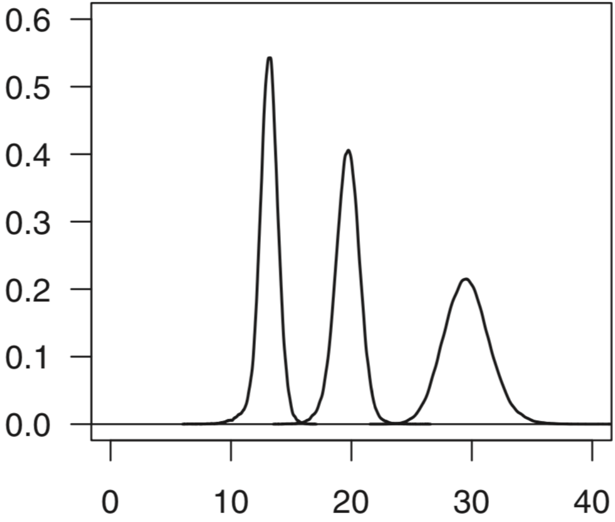
Earthquake example



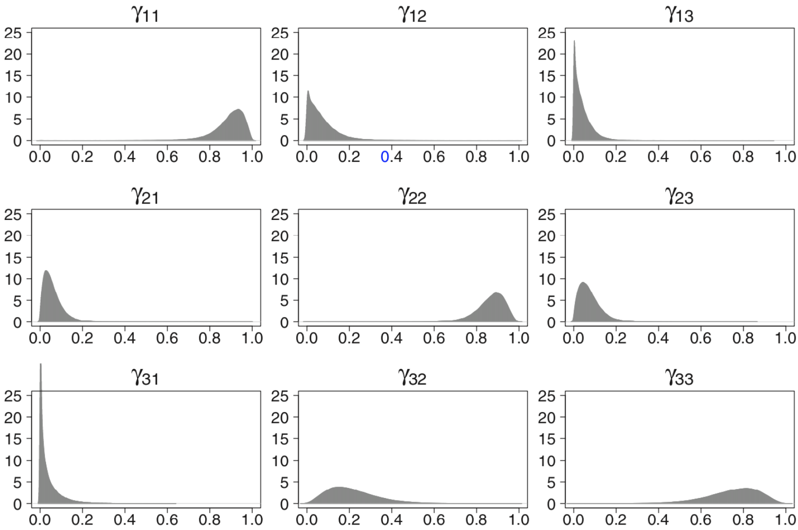
Earthquake example ($m = 3$)

Parameter	Min	Q1	Mode	Median	Mean	Q3	Max
λ_1	6.21	12.62	13.12	13.15	13.12	13.68	16.85
λ_2	13.53	19.05	19.79	19.74	19.71	20.42	27.12
λ_3	22.08	28.33	29.87	29.59	29.64	30.88	43.88
γ_{11}	0.001	0.803	0.882	0.861	0.843	0.905	0.998
γ_{12}	0.000	0.047	0.056	0.085	0.104	0.139	0.964
γ_{13}	0.000	0.020	0.011	0.042	0.053	0.075	0.848
γ_{21}	0.000	0.043	0.050	0.070	0.083	0.108	0.979
γ_{22}	0.009	0.784	0.858	0.837	0.824	0.880	0.992
γ_{23}	0.000	0.052	0.060	0.082	0.093	0.122	0.943
γ_{31}	0.000	0.021	0.011	0.049	0.068	0.096	0.758
γ_{32}	0.000	0.144	0.180	0.213	0.229	0.296	0.918
γ_{33}	0.010	0.627	0.757	0.718	0.703	0.795	0.986

Earthquake example ($m = 3$)



Earthquake example ($m = 3$)



Bayesian estimation of the number of states

- In the Bayesian framework, how can we deal with model selection?
- For example estimating the number of states m
- When selecting between two competing models m_1 and m_2 we can form the posterior odds:

$$\frac{p(m_2|x)}{p(m_1|x)} = \frac{p(m_2)}{p(m_1)} \frac{p(x|m_2)}{p(x|m_1)}$$

- The posterior odds is equal to the prior odds times the Bayes Factor $\frac{p(x|m_2)}{p(x|m_1)}$
- If the prior odds is 1 (ie $p(m_1) = p(m_2)$) then the posterior odds is equal to the Bayes Factor

Estimating the Bayes Factor

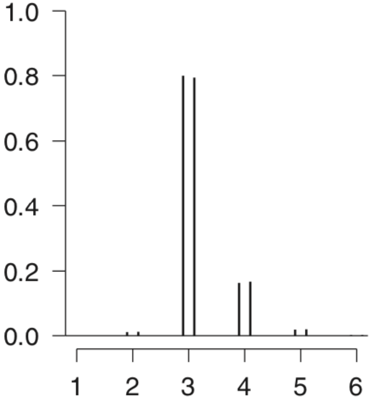
- How to calculate the Bayes Factor?
- One approach is to calculate separately the marginal likelihood of each model:

$$p(x|m) = \int p(x|m,\theta)p(\theta|m)\mathrm{d}\theta$$

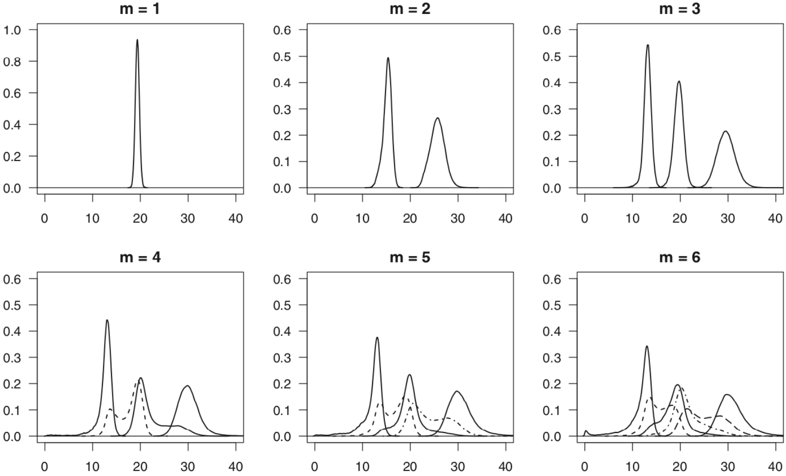
- Not easy but there are several methods to get an estimate (Newton and Raftery 1994)
- Alternatively, we can explore both models (or more) jointly using reversible-jump MCMC (Green 1995)
- These are general methods, not specific to HMM models, so for more details see a course on Bayesian model selection

Earthquake example

- Prior on the number of states m: uniform from 1 to 6
- Posterior probability distribution in two separate runs:



Earthquake example



Conclusions

- HMM can be analysed using Bayesian statistics and Monte-Carlo methods
- The forward-backward algorithm can be modified to return samples of the hidden states path
- This can be used within a Gibbs sampler to sample both the hidden path and the HMM parameters
- Model selection can be performed by computing Bayes Factors