Hidden Markov Models: lecture 6

Model selection and checking

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:

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

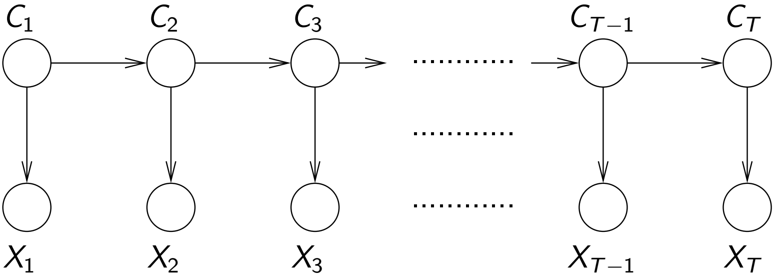
  - The emission probabilities $p_i(x)$ where i is a state and x is an emission:

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

- 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



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

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

Parameter estimation

- In the previous lectures, we discussed how to calculate the likelihood, how to estimate the hidden states, and how to estimate the parameters of a HMM
- This assumed that we know the (unparametrised) structure of the HMM to use, and in particular the number of hidden states
- In this lecture we are concerned with two closely related questions
- How can we **select** between two models? For example, a model with 2 states and a model with 3 states?
- How can we **check** that a model fits well with the data, without comparison with other models
- These are very common problems when modelling, not just for HMM, and even though we focus on HMM some of the solutions we will describe apply to more general models

Model selection

- The likelihood L is an indication of how well a model fits
- A HMM with more states m will always have a better likelihood L
- Two criteria are in common use to account for the complexity of models when comparing them
- The Akaike information criterion:

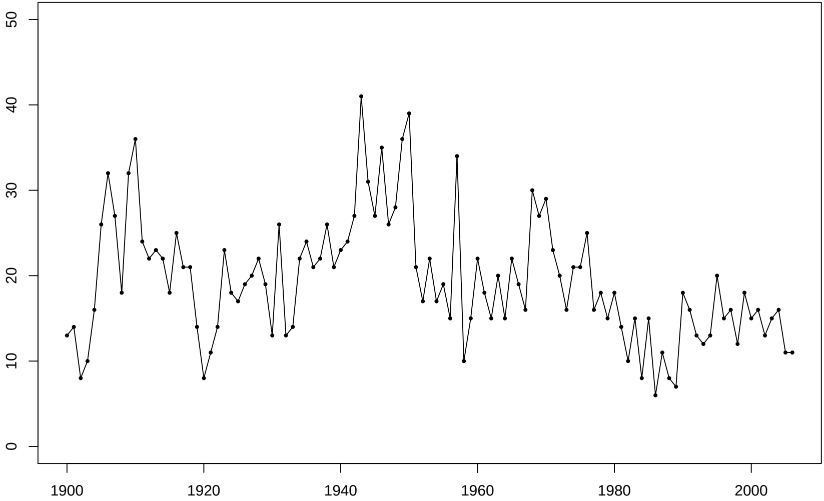
$$\mathrm{AIC} = -2\log(L) + 2p$$

- The Bayesian information criterion:

$$\mathrm{BIC} = -2\log(L) + p\log(T)$$

- p is the number of parameters of the model, and T the number of observations
- The model with the smallest AIC or BIC is selected
- In the case of a HMM with m states and Poisson emissions, we have $m^2 - m$ parameters in the transition matrix $\mathbf{\Gamma}$ and m parameters for each of the Poisson rates and so $p = m^2$

Earthquake example

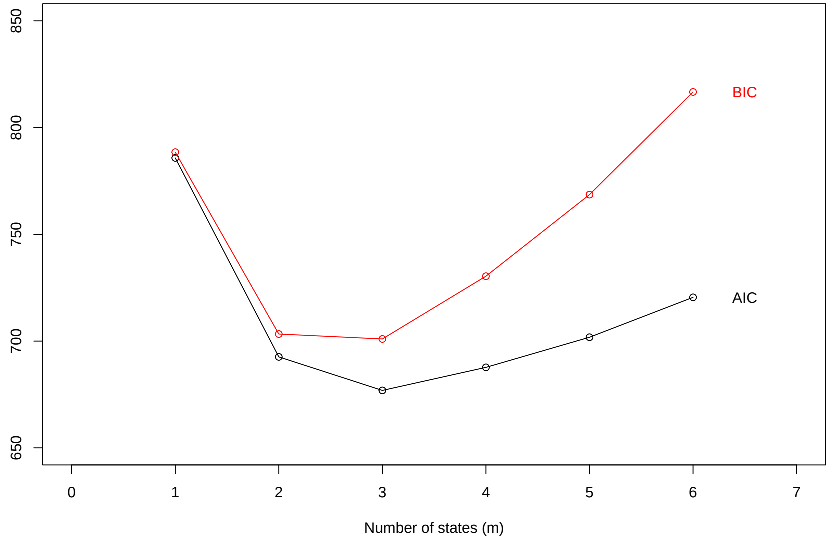


Application to earthquakes data

- HMM = hidden Markov model ($p = m^2$)
- IM = independent mixture ($p = 2m - 1$)

Model	m	p	logL	AIC	BIC
HMM	1	1	-391.9189	785.8	788.5
HMM	2	4	-342.3183	692.6	703.3
HMM	3	9	-329.4603	676.9	701.0
HMM	4	16	-327.8316	687.7	730.4
HMM	5	25	-325.9000	701.8	768.6
HMM	6	36	-324.2270	720.5	816.7
IM	1	1	-391.9189	785.8	788.5
IM	2	3	-360.3690	726.7	734.8
IM	3	5	-356.8489	723.7	737.1
IM	4	7	-356.7337	727.5	746.2

Application to earthquakes data



Application to earthquakes data

- Both AIC and BIC select the HMM with $m = 3$ states
- More generally, BIC and AIC do not always agree
- When $T > e^2$ (as is usually the case) the BIC penalizes larger models more than the AIC
- Independent mixture models are not as good as HMM for this dataset, despite the higher number of parameters in HMM

Model checking

- Even after selection of the best model, there remains the question of how good the model is in absolute (not relative) terms
- Need to assess the goodness of fit of the model
- This is something commonly done for simpler model and that we need to adapt for HMM
- For example, in a simple linear regression model we have:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

  with $\epsilon_i \sim \mathrm{Norm}(0, \sigma^2)$
- The residuals $y_i - \beta_0 - \beta_1 x_i$ are therefore expected to be independently and identically distributed as $\mathrm{Norm}(0, \sigma^2)$ and this can be used to check the model
- For a HMM, what is the residual for each observation X_t?

HMM pseudo-residuals

- We consider that X_t is continuous (similar results can be derived for the discrete case)
- If X_t is from a distribution with cumulative density F_{X_t} then we can consider the pseudo-residual:

  $$z_t = \Phi^{-1}(F_{X_t}(x_t))$$

  where Φ is the cumulative density of a Normal(0,1)
- Since X_t has cumulative density F_{X_t} we have that $F_{X_t}(x_t)$ should be distributed as Unif(0,1) and therefore z_t should be distributed as Normal(0,1)
- What is the distribution of X_t?

HMM pseudo-residuals

- What is the distribution of X_t?
- One approach is to use the conditional distribution given all other data:

$$f_{X_t}(x) = p(X_t = x | \boldsymbol{X}^{(-t)} = \boldsymbol{x}^{(-t)})$$

- Using our previous calculation of the likelihood, we find:

$$f_{X_t}(x) = \frac{\boldsymbol{u}(1)\boldsymbol{P}(x_1)\boldsymbol{\Gamma}...\boldsymbol{P}(x_{t-1})\boldsymbol{\Gamma}\boldsymbol{P}(x)\boldsymbol{\Gamma}\boldsymbol{P}(x_{t+1})...\boldsymbol{\Gamma}\boldsymbol{P}(x_T)\boldsymbol{1}'}{\boldsymbol{u}(1)\boldsymbol{P}(x_1)\boldsymbol{\Gamma}...\boldsymbol{P}(x_{t-1})\boldsymbol{\Gamma}\boldsymbol{\Gamma}\boldsymbol{P}(x_{t+1})...\boldsymbol{\Gamma}\boldsymbol{P}(x_T)\boldsymbol{1}'}$$

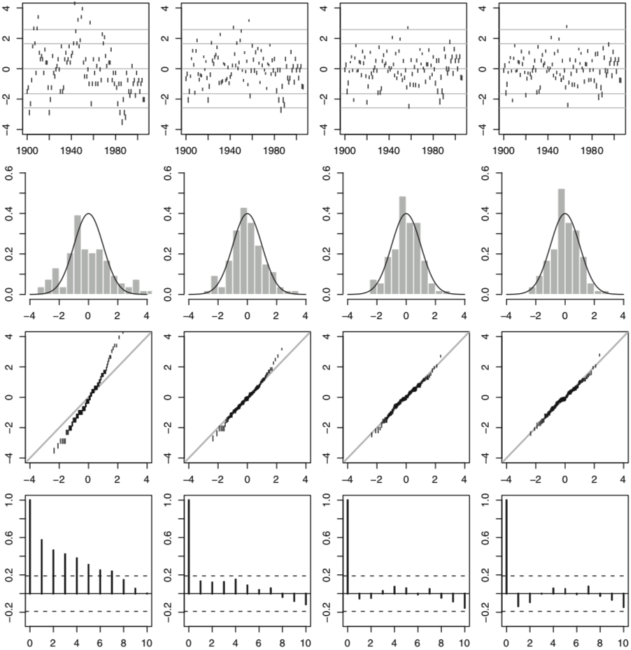
- Using the definitions of the forward and backward vectors:

$$f_{X_t}(x) \propto \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\boldsymbol{P}(x)\beta_t'$$

HMM pseudo-residuals

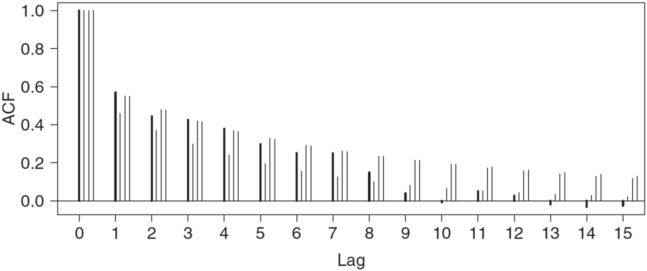
- Use forward-backward algorithm to calculate vectors $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$
- Compute distribution of X_t given all other observations
- Compute pseudo-residual z_t for each x_t
- This can be used to detect outliers
- Can also be used to check validity of a model
- Plot distribution of pseudo-residuals vs Normal distribution
- Q-Q plot of observed (y-axis) vs expected (x-axis)
- ACF of pseudo-residuals

Model checking for earthquakes dataset



Observed vs expected ACF

- The observed ACF in the data can be compared to the ACF expected under the HMM
- $\mathrm{Corr}(X_t, X_{t+k})$ can be computed analytically or simulated
- In the earthquake example, ACF of real data (bold) vs HMM with $m = 1, 2, 3$ states:



Conclusions

- Model selection is used to choose the best model amongst a set of candidates
- Comparing likelihoods directly is unfair since models with more parameters (eg HMM states) will always have better likelihoods
- AIC and BIC criteria penalize more complex models to make comparison more fair
- If all candidates are bad, the least bad is selected but this might still be very bad
- Model checking assesses the goodness of fit of a model (without comparison to other models)
- We can compute pseudo-residuals for a HMM and use them to do model checking
- So far we have considered HMM only within a classical statistics framework
- In our next lecture we will see how to use HMM in a Bayesian framework