Hidden Markov Models: lecture 10

Links and extensions

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:

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

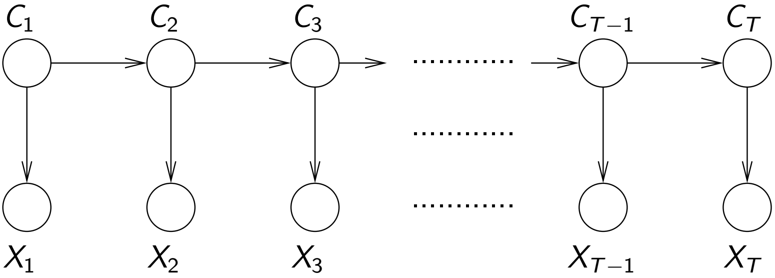
  - The emission probabilities $p_i(x)$ where i is a state and x is an emission:

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

- 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

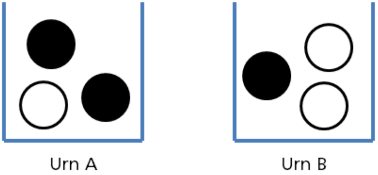


$$p(\boldsymbol{X}^{(T)}, \boldsymbol{C}^{(T)}) = p(C_1) \prod_{k=2}^{T} p(C_k | C_{k-1}) \prod_{k=1}^{T} p(X_k | C_k)$$

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

Definitions of HMMs as generalization of models

- HMM is an extension of an independent mixture model (cf lecture 1, earthquake example)
- HMM is an extension of Markov chain models (cf lecture 2, weather example)
- When the data is categorical (multinomial-HMM) the HMM is a generalization of the urn problem with replacement, in which the choice of urn is ruled by a Markov chain



Regular grammars

- Transformational grammars are important objects in computer science
- Grammars are defined by:
  - A set of terminal symbols, usually denoted with lower case: a, b, c,...
  - A set of non-terminal symbols, usually denoted with upper case: A, B, C,...
  - Production rules of the form $\alpha \rightarrow \beta$ where α and β are strings of (terminal or non-terminal) symbols
- The most fundamental type is the regular grammar, in which the only allowed rules are of the form $X \rightarrow a$ or $X \rightarrow aY$

Examples

- Let S denote the start and ϵ denote the end
- A regular grammar allowing any string of as and bs:

$$S \to aS, S \to bS, S \to \epsilon$$

- This is also written:

$$S \to aS|bS|\epsilon$$

- A regular grammar for strings of as and bs with an odd number of as:

$$S \to aT|bS, \qquad T \to aS|bT|\epsilon$$

- The grammar of palindroms however is not regular:

$$S \to aSa|bSb|\epsilon$$

Link between HMMs and grammars

- In a stochastic grammar, each rule is given a probability such that the probabilities of the production rules from any non-terminal symbol add up to one
- HMMs are equivalent to stochastic regular grammars
- More general classes of grammars also exist
- An important class is the stochastic context-free grammar (SCFG), which allows any rule of the form $X \rightarrow \beta$ (but not something like $aXb \rightarrow a\beta b$ which would be context dependent)
- The palindrom grammar in the previous slide is an example of context-free grammar
- SCFGs have their own dynamic programming algorithms (inside, outside and CYK equivalent to forward, backward and Viterbi, respectively) which are less computationally efficient but still usable
- More general grammars are very hard to work with

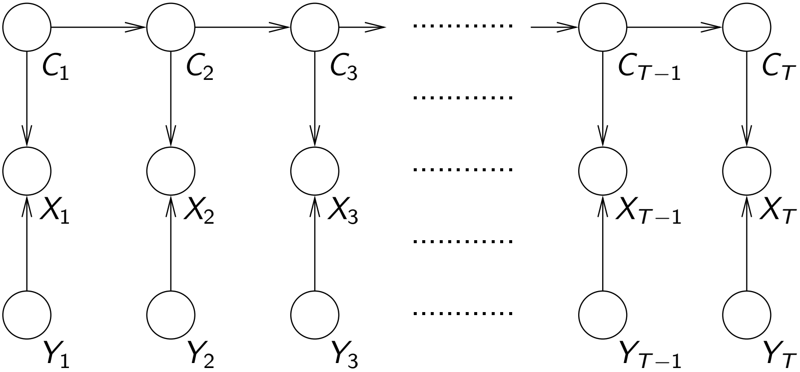
HMMs with various univariate state-dependent distributions

- Many forms of the emission distribution can be used and all the results and algorithms we have described remain applicable
- We have seen several types of HMMs depending on the form of the emission probabilities:
  - Poisson-HMM (lecture 1, earthquake example)
  - Bernouilli-HMM (lecture 2, rain example)
  - Multinomial-HMM (lecture 3, casino example)
- If the data are real numbers, we could use for example a Normal-HMM
- If the data are proportions, we could use for example a Beta-HMM

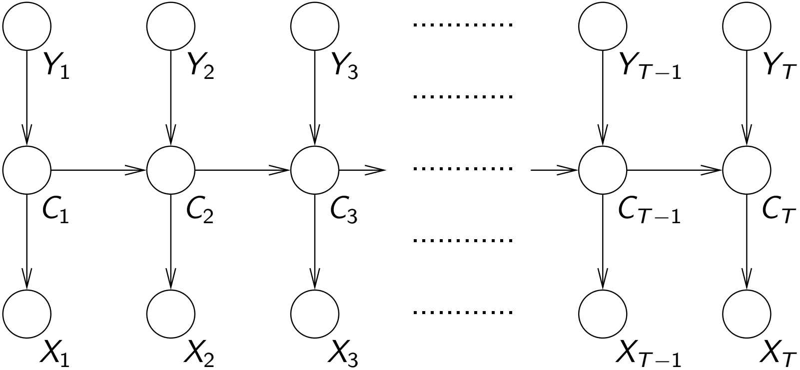
HMMs for multivariate data

- HMMs can also be used if the data is multivariate with q variables
- Need to have longitudinal conditional independence, ie the vectors observed at each time point are mutually independent given the hidden states
- Are the individual variables also mutually independent at each time point given the state?
- If not, need to specify a multivariate distribution for each state, for example a multivariate Normal distribution with location vector of length q and covariance matrix of dimension $q \times q$
- If yes, then we have contemporaneous conditional independence, and can specify q separate univarite distributions, for example q separate Normal distributions

Covariates in the state-dependent distribution

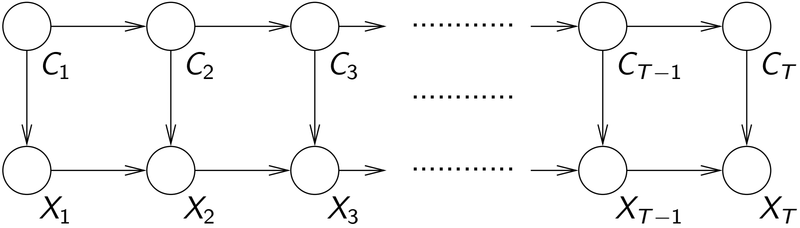


Covariates in the transition probabilities



- This structure is called a non-homogenous HMM

Markov-switching model



- Markov-switching AR(1) model
- Can likewise have AR(2), AR(3),...

State space models

- Instead of considering a discrete number of states, it may be useful to have a continuous-valued state space
- For example, in the earthquake dataset, we could have:

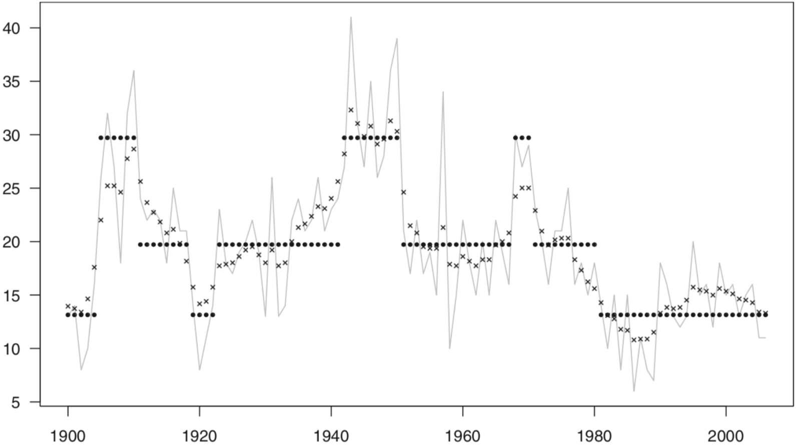
$$C_t = \phi C_{t-1} + \sigma \nu_t \text{ and } X_t \sim \text{Poisson}(\beta \text{exp}(C_t))$$

  with $0 < \phi < 1$ and $\beta > 0$ and $\sigma > 0$ and $\nu_t \sim \text{Normal}(0, 1)$
- The dependencies are the same as in a HMM
- But the state process $\{C_t\}$ determines the rate of occurrence of major earthquakes
- $\{C_t\}$ is an autoregressive process of order 1, with values fluctuating around zero
- This is a state-space model (SSM) with continuous-valued state process
- A HMM is the special case of a SSM when the state space is discrete
- What can we do if the state space is continuous as above?

State space models

- Linear Gaussian SSMs can be fitted using the Kalman filter
- Otherwise (as in the model above) need to use complex approximate method such as extended Kalman filter or particle filter (aka sequential Monte-Carlo, SMC)
- Alternatively, can use a fine (eg $m = 50$ to $m = 200$) discretization of the state space to make the HMM methodology applicable to the SSM
- Can vary the number of states m and range of values to make sure the approximation is stable
- For example, compare HMM with 3 states with SSM from previous slide with discretization $m = 200$
- This technique works well for one-dimensional state spaces but is problematic for high-dimensional state spaces

Earthquake example



Higher order of HMM

- For example, a second-order HMM has transition probabilities:

$$\gamma(i, j, k) = p(C_t = k | C_{t-1} = j, C_{t-2} = i)$$

- This implies a stationary bivariate distribution:

$$u(j, k) = p(C_{t-1} = j, C_t = k)$$

- A second-order HMM with m states can be defined as a first-order HMM with m^2 states and only m^3 possible transitions (because a state (i, j) can only transit to a state (j, k))
- The time to compute the likelihood is therefore $O(Tm^3)$
- The number of free parameters in the transition matrix is $m^2(m-1)$

Hidden semi-Markov models

- Considering higher orders of HMM is computationally demanding
- Hidden semi-Markov models (HSMMs) are an alternative way of relaxing the Markov assumption
- The number of steps where the model stays in a state is called the dwell time
- For a HMM, the dwell times are geometrically distributed
- A HSMM allows us to use different dwell times distributions
- A HSMM with m states consists of:
  - A transition matrix $\boldsymbol{\Omega}$ with zeros on the diagonal (so that staying in the same state is not allowed)
  - A dwell time distribution $d_i(r)$ for each state i
  - An emission distribution $p_i(x)$ for each state i
- Note that if all $d_i(r)$ are Geometric, the HSMM is a HMM

HSMM exactly represented as a HMM

- There are forward and backward recursions for HSMM but they are much more complex and computationally demanding than for HMM
- However, it is often possible to rewrite a HSMM as a HMM
- This can be done exactly if all dwelling distribution are either geometric or of finite support
- For example, consider a two-state Poisson-HSMM with $\lambda_1 = 5$ and $\lambda_2 = 10$ and $d_1(1) = 1/4$, $d_1(2) = 3/8$, $d_1(3) = 1/4$, $d_1(4) = 1/8$ and $d_2(r) = 0.1 \times 0.9^{r-1}$ for $r = 1, 2, 3, \ldots$
- This is equivalent to a 5-states HMM with $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 5$, $\lambda_5 = 10$ and:

$$\mathbf{\Gamma} = \begin{pmatrix} 0 & 3/4 & 0 & 0 & 1/4 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/3 & 2/3 \\ 0 & 0 & 0 & 0 & 1 \\ 1/10 & 0 & 0 & 0 & 9/10 \end{pmatrix}$$

HSMM approximately represented as a HMM

- If one of the dwelling time distribution is not geometric and with infinite support (for example a Poisson distribution shifted up by one) then there is no exact equivalent HMM
- But we can build an approximately equivalent HMM
- Each state i with a non-geometric infinite support is decomposed into m_i consecutive states, such that the cumulative distribution function associated with $d_i(r)$ is the same up to m_i
- The m_i-th state uses a geometric distribution which is approximate but a good approximation when m_i is large

Conclusions

- Hidden Markov Models can be formulated as:
  - An extension of an independent mixture model
  - An extension of Markov chain models
  - A type of urn problem with replacement
  - A stochastic regular grammar
- HMMs can accomodate multivariate data
- Covarites can be included within both emission or transition processes
- Hidden Markov Models can be extended into:
  - A state space model (SSM) by considering a continuous state space
  - A higher-order HMM
  - A hidden-semi Markov Model (HSMM)
- These extensions are harder to work with, but can often be approximated by HMMs with high number of states
- The powerful dynamic programming algorithms available for HMMs is what makes them so popular