Chapter 2 - Definition and properties

02433 - Hidden Markov Models

Marco Zugno, Martin Wæver Pedersen, Henrik Madsen

Course week 2

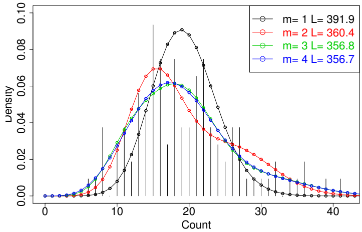
MWP, compiled June 7, 2011

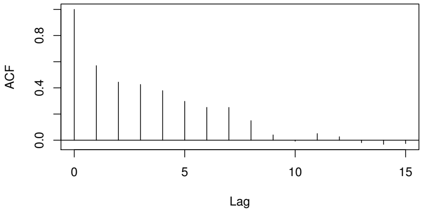
LUNDS UNIVERSITET
Lunds Tekniska Högskola

Interreg IVA
ÖRESUND – KATTEGAT – SKAGERRAK

DTU

Recall: US major earthquake count



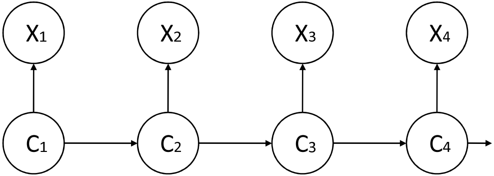


We observe:

- Left figure: Data show overdispersion, which can be captured by an independent mixture model.
- Right figure: Autocorrelation function for earth quake data shows serial dependence between observations. This dependence can be modelled with a Markov process.

An HMM is a dependent mixture model where the dependence between the mixtures is modelled by a Markov process.

Hidden Markov model: Definition



A hidden Markov model $\{X_t : t \in \mathbb{N}\}$ is a dependent mixture where

$$\Pr(C_t = i|\mathbf{C}^{(t-1)}) = \Pr(C_t = i|C_{t-1}),\ t = 2, 3, \ldots$$
$$\Pr(X_t = x|\mathbf{X}^{(t-1)}, \mathbf{C}^{(t)}) = \Pr(X_t = x|C_t),\ t \in \mathbb{N},$$

where $\{C_t : t = 1, 2, \ldots\}$ is the unobserved (hidden) parameter process, and $\{X_t : t = 1, 2, \ldots\}$ is the state-dependent process. When C_t is known the distribution of X_t only depends on C_t as shown in the directed graph above.

Hidden Markov model: Notation

$\{X_t\}$ is an m-state HMM if the Markov chain $\{C_t\}$ has m states.

When dealing with discrete observations, we define

$$p_i(x) = \Pr(X_t = x | C_t = i),\ i = 1, 2, \ldots, m,$$

as the state-dependent distributions, which is interpreted as the probability of the observation at time t conditional on the state C_t.

In the continuous case p_i is a probability density function instead of probability distribution function.

Also define

$$u_i(t) = \Pr(C_t = i),\ t = 1, \ldots, T,$$

which is simply the probability of the state being in i at time t.

Marginal distributions (I)

The marginal distribution $\Pr(X_t = x)$ of the observation X_t is often of interest. This can be calculated from the distribution of the hidden state and the state-dependent distribution:

$$\begin{aligned}\Pr(X_t = x) &= \sum_{i=1}^{m} \Pr(C_t = i)\Pr(X_t = x | C_t = i) = \sum_{i=1}^{m} u_i(t)p_i(x) \\ &= (u_1(t), \ldots, u_m(t)) \begin{pmatrix} p_1(x) & & 0 \\ & \ddots & \\ 0 & & p_m(x) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ &= \mathbf{u}(t)\mathbf{P}(x)\mathbf{1}'\end{aligned}$$

The result is written in a compact matrix-vector notation which will be used heavily throughout the course. Still $\mathbf{u}(t)$ is the distribution of the hidden state, and $\mathbf{P}(x)$ is the state-dependent distribution of the observation.

Marginal distributions (II)

Using the properties of the probability transition matrix $\boldsymbol{\Gamma}$ we have $\mathbf{u}(t) = \mathbf{u}(1)\boldsymbol{\Gamma}^{t-1}$. It therefore follows that

$$\Pr(X_t = x) = \mathbf{u}(1)\boldsymbol{\Gamma}^{t-1}\mathbf{P}(x)\mathbf{1}',$$

which holds if the Markov chain is homogeneous, but not necessarily stationary.

For a stationary Markov chain with stationary distribution $\boldsymbol{\delta}$ we get

$$\Pr(X_t = x) = \boldsymbol{\delta}\mathbf{P}(x)\mathbf{1}',$$

since $\boldsymbol{\delta}\boldsymbol{\Gamma}^{t-1} = \boldsymbol{\delta}$ for all $t \in \ltimes$.

Parents and Conditional independence

In any directed graphical model the joint distribution of a set of random variables V_i is

$$\Pr(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} \Pr(V_i | pa(V_i)),$$

where $pa(V_i)$ are the "parents" of V_i. For the directed graph on page 3 we have e.g. that $pa(C_2) = \{C_1\}$ and that $pa(X_2) = \{C_2\}$. Thus, that parents of a random variable A, say, are the variables that influence the distribution of A.

Conditional independence
Definition: If for a random variable A it holds that $\Pr(A = a | B, C) = \Pr(A = a | B)$, then A is said to be conditional independent of C given B. This is an important property influencing the random variables in a directed graph. For example in the figure on page 3 it holds that X_t are conditional independent of $X_1, \ldots, X_{t-1}, X_{t+1}, \ldots$ given C_t for all t. For the hidden state it holds that C_{t+1} is conditional independent of $C_1, \ldots, C_{t-1}$ given C_t. So, by conditioning on the parents of a random variable it is independent of everything else.

Marginal multivariate distributions (I)

Consider the four random variables X_t, X_{t+k}, C_t, and C_{t+k} with dependency relations as specified by the directed graph on page 3. Using conditional independence we have

$$\begin{aligned}
\Pr(X_t = v, X_{t+k} = w) &= \sum_{i=1}^{m}\sum_{j=1}^{m} \Pr(X_t = v, X_{t+k} = w, C_t = i, C_{t+k} = j) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \Pr(C_t = i) p_i(v) \Pr(C_{t+k} = j | C_t = i) p_j(w) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} u_i(t) p_i(v) \gamma_{ij}(k) p_j(w),
\end{aligned}$$

recall that $\gamma_{ij}(k)$ are the elements of the transition probability matrix $\mathbf{\Gamma}(k)$.

Marginal multivariate distributions (II)

The expression on the previous page can be formulated using matrix notation

$$\Pr(X_t = v, X_{t+k} = w) = \mathbf{u}(t)\mathbf{P}(v)\mathbf{\Gamma}^k\mathbf{P}(w)\mathbf{1}'.$$

If the chain is stationary, this reduces to

$$\Pr(X_t = v, X_{t+k} = w) = \boldsymbol{\delta}\mathbf{P}(v)\mathbf{\Gamma}^k\mathbf{P}(w)\mathbf{1}'.$$

The above can be generalised to higher order distributions. For example a trivariate distribution

$$\Pr(X_t = v, X_{t+k} = w, X_{t+k+l} = z) = \boldsymbol{\delta}\mathbf{P}(v)\mathbf{\Gamma}^k\mathbf{P}(w)\mathbf{\Gamma}^l\mathbf{P}(z)\mathbf{1}'.$$

This generalisation is an important property of HMMs.

The likelihood in general

The likelihood of a set of observations $\mathbf{X}^{(T)} = (X_1, X_2, \ldots, X_T)$ given the parameters of an HMM is

$$L_T = \Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)} | \text{model})$$
$$= \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'.$$

The proof is shown on page 37-38 in Zucchini09. The expression is derived using the conditional independence of the HMM, which allows a recursive calculation of the joint probability of the observations.

Note that since this is a joint probability calculated by multiplying a (possibly) large number of terms, there is a risk that the likelihood value will lead to numerical over- or underflow (a remedy for this is considered in chapter 3).

Recursive algorithm for the likelihood

Define the forward probability vector

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_t).$$

We will elaborate further on the function of $\boldsymbol{\alpha}_t$ in chapter 4.

Using $\boldsymbol{\alpha}_t$ the likelihood can be calculated recursively:

$$\begin{aligned}
\boldsymbol{\alpha}_1 &= \boldsymbol{\delta}\mathbf{P}(x_1) \\
\boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t) \quad \text{for} \quad t = 2, 3, \ldots, T \\
L_T &= \boldsymbol{\alpha}_T \mathbf{1}'
\end{aligned}$$

The complexity of the calculation is Tm^2 since it consists of T vector-matrix multiplications each having complexity m^2.

Likelihood with missing data

In practice it is often the case that data do not arrive at uniform intervals in time. This case we say we have missing data. For example, the dataset $(x_1, x_2, x_4, x_7, x_8, \ldots, x_T)$ has data missing at times $t = 3, 5, 6$.

Fortunately it is straightforward to compute the likelihood under missing data. For example consider the dataset (x_1, x_3) where the second observation is missing:

$$\Pr(X_1 = x_1, X_3 = x_3) = \sum \delta_{C_1} p_{C_1}(x_1) \gamma_{C_1, C_3}(2) p_{C_3}(x_3),$$

where $\gamma_{ij}(k)$ is a k-step transition probability, and the sum is taken over c_1 and c_3.

In the other case where x_3, x_5 and x_6 are missing the likelihood is in matrix form written as

$$L_T^{-(3,5,6)} = \boldsymbol{\delta} \mathbf{P}(x_1) \boldsymbol{\Gamma} \mathbf{P}(x_2) \boldsymbol{\Gamma}^2 \mathbf{P}(x_4) \boldsymbol{\Gamma}^3 \mathbf{P}(x_7) \ldots \boldsymbol{\Gamma} \mathbf{P}(x_T) \mathbf{1}'.$$

Exercises

1,2,6,9,11