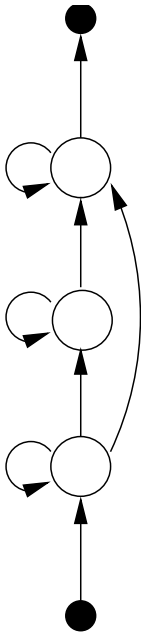
UniS Centre for Vision Speech & Signal Processing
University of Surrey, Guildford GU2 7XH.

HMM tutorial 1

by Dr Philip Jackson

- Fundamentals
- Markov models
- Hidden Markov models
  - Likelihood calculation
  - Optimum state sequence (Viterbi)
  - Re-estimation (Baum-Welch)
- Output probabilities
- Extensions and applications



http://www.ee.surrey.ac.uk/Personal/P.Jackson/tutorial/

Fundamentals

- Least-squares parameter estimation
- Likelihood equation and ML estimation
- Bayes' theorem and MAP
- Discrete vs. continuous observations
- Probability distribution functions

Scientific inference

Steps of scientific investigation:

1. develop experimental apparatus
2. perform measurements
3. analyse data

Levels of inference:

1. parameter estimation
2. classification
3. recognition

Estimation of deterministic parameters

Bias

The *expectation* of a parameter estimate can be written

$$E\left\{\hat{\lambda}(x)\right\} = \int \hat{\lambda}(x)\, p(x|\lambda)\, dx, \tag{1}$$

where $\hat{\lambda}$ is the estimate of parameter λ, and x is the feature space, which can lead to three kinds of result:

1. *Unbiased*: $E\left\{\hat{\lambda}(x)\right\} = \lambda$, for all values of λ. The average of the estimates tends towards the true value of the parameter.
2. *Known bias*: $E\left\{\hat{\lambda}(x)\right\} = \lambda + b$, where b is independent of λ. Hence, we can obtain an unbiased estimate by subtracting b from $\hat{\lambda}(x)$.
3. *Unknown bias*: $E\left\{\hat{\lambda}(x)\right\} = \lambda + b(\lambda)$, where b depends on λ.

Variance

The *variance* of the estimation error,

$$\text{var}\left[\hat{\lambda}(x) - \lambda\right] = E\left\{\left[\hat{\lambda}(x) - \lambda\right]^2\right\} - b^2(\lambda), \qquad (2)$$

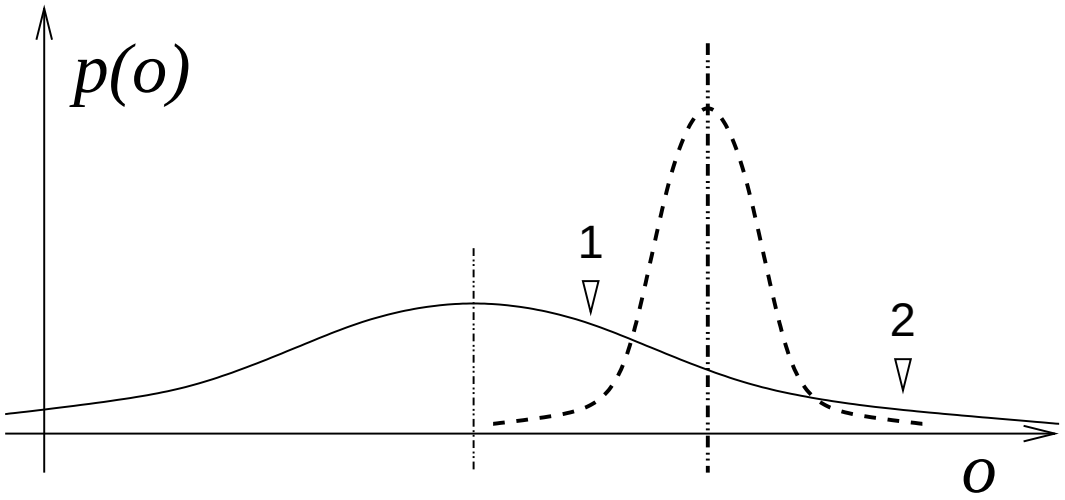
describes the spread of the error.

In general, we want unbiased estimates with minimum variance, but no simple procedure exists to find them. However, one approach to improving the quality of our estimates is to use *maximum likelihood*.

Maximum likelihood

Motivation for the most likely

Consider these two different probability distribution functions (pdfs):



Maximum likelihood estimation

We try to use as our estimate the value of λ that most likely caused a given value of o to occur. We denote the value obtained by using such a procedure as a *maximum likelihood* (ML) estimate, $\hat{\lambda}_{\mathrm{ML}}(o)$. The ML estimate is obtained by differentiating $\ln p(o|\lambda)$ with respect to λ and setting the result equal to zero:

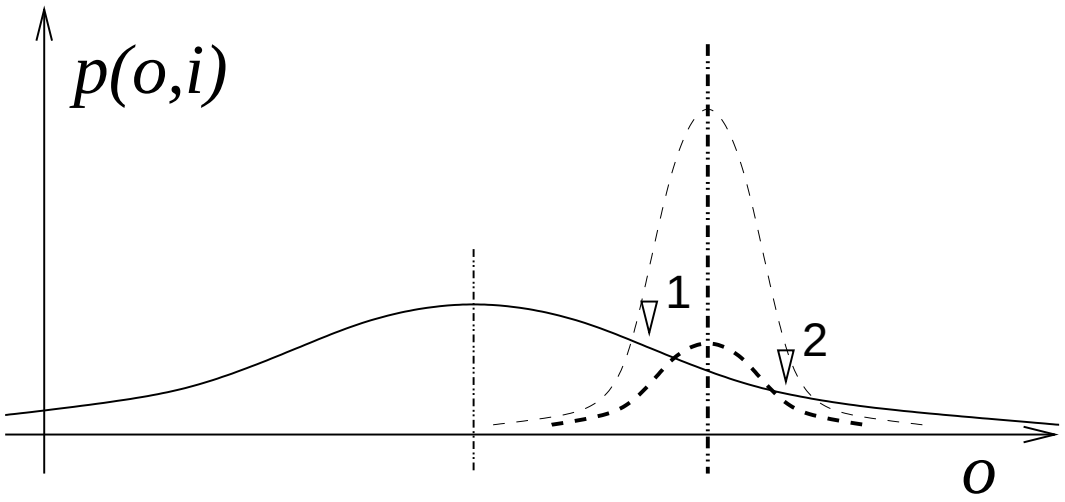
$$\frac{\partial L_o(\lambda)}{\partial \lambda} = \frac{\partial \ln p(o|\lambda)}{\partial \lambda} = 0. \tag{3}$$

This equation is called the *likelihood equation*.

Bayesian inference

Value of prior knowledge

Let us suppose there are two conditional pdfs, as follows:



Conditional probability

Now imagine two dependent events:

	Event A	
Event B	True	False
True	0.1	0.3
False	0.4	0.2

The probability of both events occurring can be expressed as

$$P(A, B) = P(A)\, P(B|A), \tag{4}$$

but also as

$$P(A, B) = P(A|B)\, P(B). \tag{5}$$

which leads us to the theorem proposed by Rev. Thomas Bayes (C.18th).

Bayes' theorem

Equating the RHS of eqs. 4 and 5 gives

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}, \tag{6}$$

which can be interpreted as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{normalisation factor}}. \tag{7}$$

Interpretation

Consider the occurrence of entities λ and $\mathcal{O}$,

$$p(\lambda|\mathcal{O}) = \frac{p(\mathcal{O}|\lambda)\, p(\lambda)}{p(\mathcal{O})}, \tag{8}$$

where $\mathcal{O}$ denotes a series of measured or observed data, and λ comprises a set of model parameters.

$p(\lambda|\mathcal{O})$ is the *posterior probability*

$p(\mathcal{O}|\lambda)$ is the *likelihood*

$p(\lambda)$ is the *prior probability*

$p(\mathcal{O})$ is the *evidence*

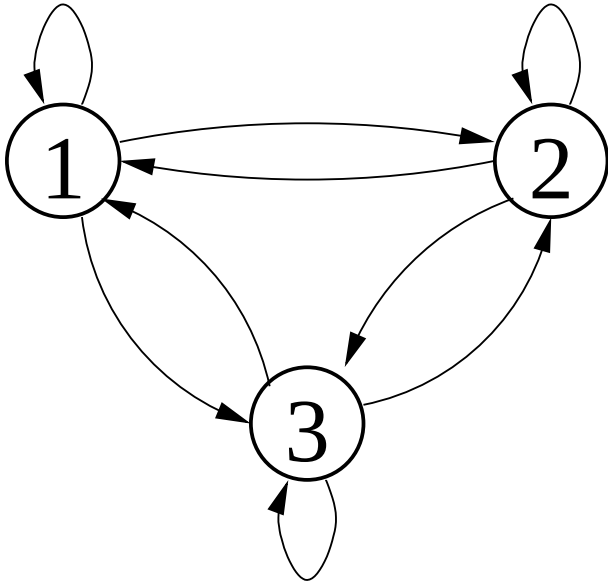
Discrete and continuous pdfs

Discrete probability distribution functions

Continuous probability distribution functions

Markov models

Ergodic model:

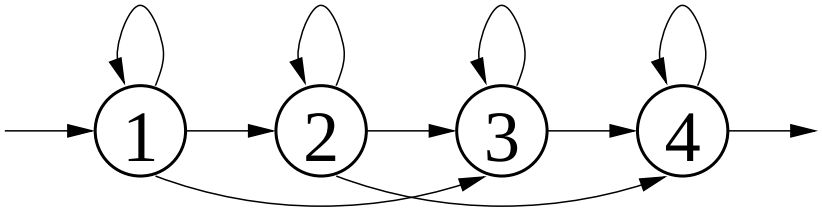


For a first-order discrete-time Markov chain, probability of state occupation depends only on the previous step:

$$P(x_t = j | x_{t-1} = i, x_{t-2} = h, \ldots) = P(x_t = j | x_{t-1} = i). \quad (9)$$

Modelling stochastic time series

Left-right Markov model:



If we assume that the RHS of eq. 9 is independent of time, we can express the state-transition probabilities,

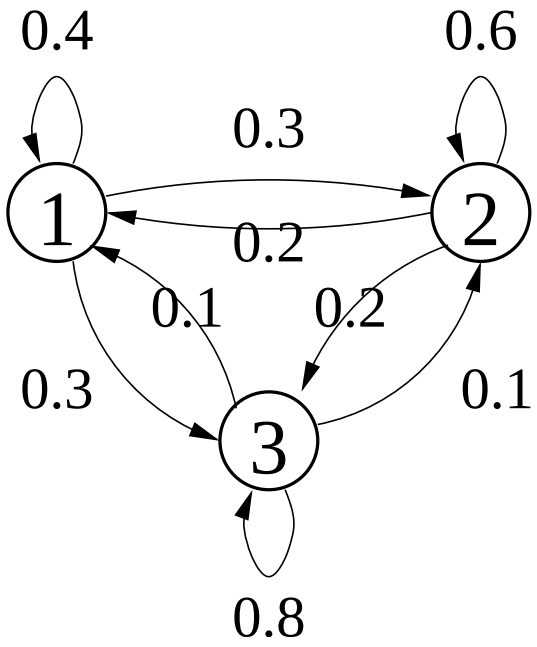
$$a_{ij} = P(x_t = j | x_{t-1} = i), \quad 1 \leq i, j \leq N, \tag{10}$$

with the properties

$$a_{ij} \geq 0, \qquad \text{and} \qquad \sum_{j=1}^{N} a_{ij} = 1. \tag{11}$$

Weather predictor example of a Markov model

State 1: rain
State 2: cloud
State 3: sun



State-transition probabilities,

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \qquad (12)$$

Weather predictor calculation

Given today is sunny (i.e., $x_1 = 3$), what is the probability of "sun-sun-rain-cloud-cloud-sun" with model $\mathcal{M}$?

$$\begin{aligned} P(X|\mathcal{M}) &= P(X = \{3,3,1,2,2,3\}|\mathcal{M}) \\ &= P(x_1 = 3)\, P(x_2 = 3|x_1 = 3) \\ &\quad P(x_3 = 1|x_2 = 3)\, P(x_4 = 2|x_3 = 1) \\ &\quad P(x_5 = 2|x_4 = 2)\, P(x_6 = 3|x_5 = 2) \\ &= \pi_3\, a_{33}\, a_{31}\, a_{12}\, a_{22}\, a_{23} \\ &= 1 \cdot (0.8)(0.1)(0.3)(0.6)(0.2) \\ &= 0.00288 \end{aligned}$$

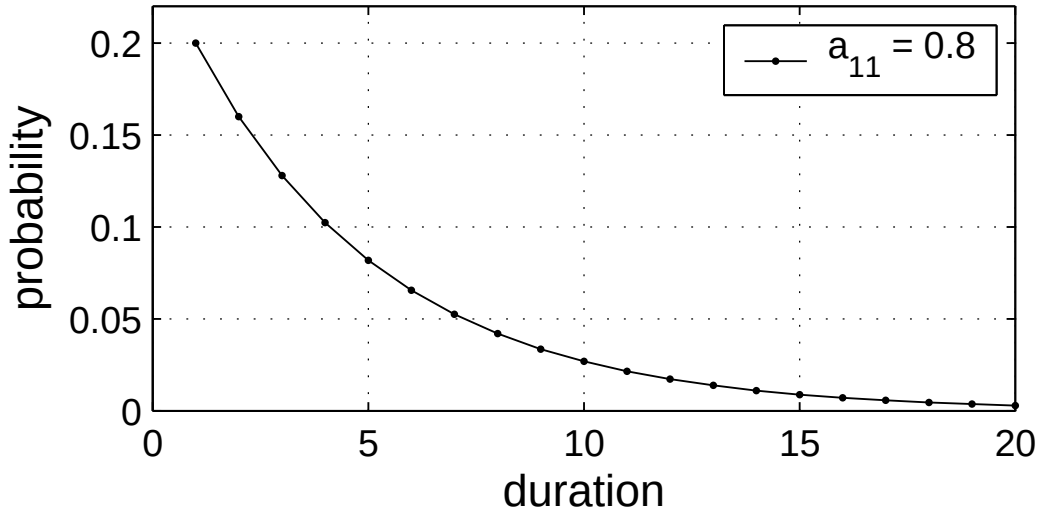
where the initial state probability for state i is

$$\pi_i = P(x_1 = i). \tag{13}$$

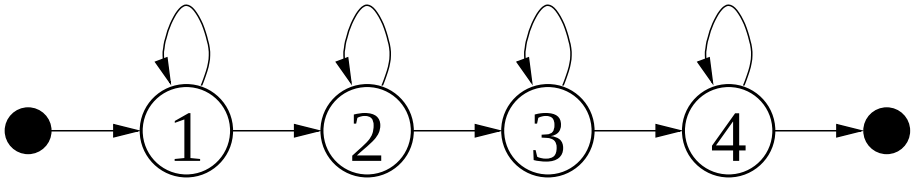
State duration probability

As a consequence of the first-order Markov model, the probability of occupying a state for a given duration, τ, is exponential:

$$p(X|\mathcal{M}, x_1 = i) = (a_{ii})^{\tau-1} (1 - a_{ii}) . \qquad (14)$$



Summary of Markov models



Transition probabilities:

$$A = \{a_{ij}\} = \begin{bmatrix} 0.6 & 0.4 & 0 & 0 \\ 0 & 0.9 & 0.1 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

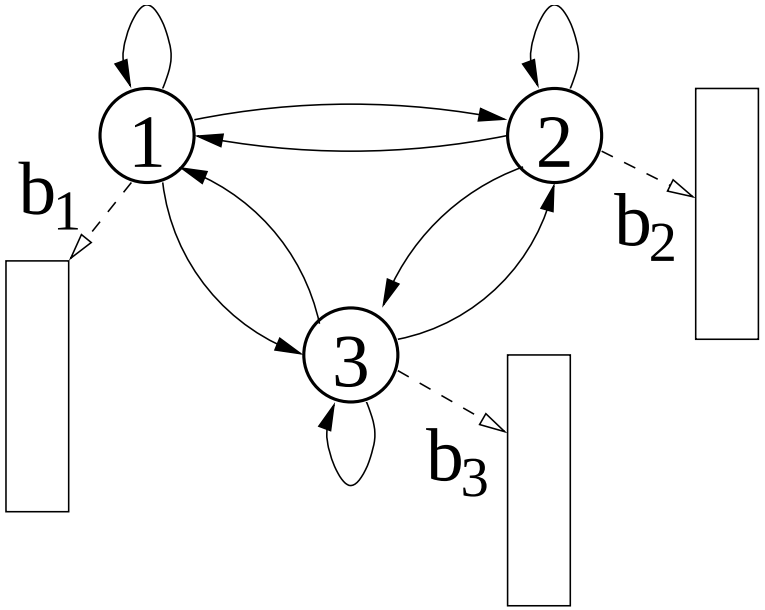
and $\pi = \{\pi_i\} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$.

Probability of a given state sequence X:

$$\begin{aligned} P(X|\mathcal{M}) &= \pi_{x_1} a_{x_1 x_2} a_{x_2 x_3} a_{x_3 x_4} \cdots \\ &= \pi_{x_1} \prod_{t=2}^{T} a_{x_{t-1} x_t}. \end{aligned} \tag{15}$$

Hidden Markov Models

Urns and balls example (Ferguson)



Probability of state i producing an observation o_t is:

$$b_i(o_t) = P(o_t|x_t = i), \tag{16}$$

which can be *discrete* or *continuous* in o.

Elements of a discrete HMM, λ

1. Number of states N, $x \in \{1, \ldots, N\}$;

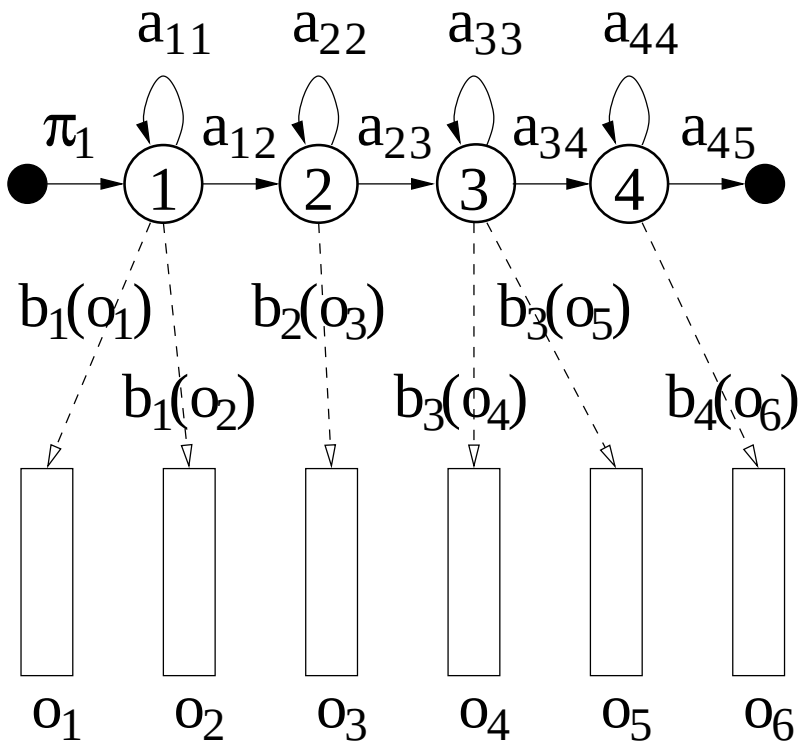
2. Number of events K, $k \in \{1, \ldots, K\}$;

3. Initial-state probabilities,
   $\pi = \{\pi_i\} = \{P(x_1 = i)\}$ for $1 \leq i \leq N$;

4. State-transition probabilities,
   $A = \{a_{ij}\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \leq i, j \leq N$;

5. Discrete output probabilities,
   $B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\}$ for $1 \leq i \leq N$ and $1 \leq k \leq K$.

Hidden Markov model example



with state sequence $X = \{1, 1, 2, 3, 3, 4\}$,

$$\begin{aligned} P(\mathcal{O}|X, \lambda) &= b_1(o_1)\, b_1(o_2)\, b_2(o_3)\, b_3(o_4)\, b_3(o_5)\, b_4(o_6) \\ P(X|\lambda) &= \pi_1\, a_{11}\, a_{12}\, a_{23}\, a_{33}\, a_{34} \end{aligned} \tag{17}$$

$$P(\mathcal{O}, X|\lambda) = \pi_1 b_1(o_1)\, a_{11} b_1(o_2)\, a_{12} b_2(o_3) \ldots \tag{18}$$

Continuous output probabilities

For a Gaussian pdf, the output probability of an emitting state, $x_t = i$, is

$$b_i(\mathbf{o}_t) = \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{19}$$

where $\mathcal{N}(\cdot)$ is a multivariate Gaussian pdf with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, evaluated at $\mathbf{o}_t$,

$$b_i(\mathbf{o}_t) = \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}_i|}} \exp\left(-\frac{1}{2}(\mathbf{o} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{o} - \boldsymbol{\mu}_i)\right) \tag{20}$$

where M is the dimensionality of the observed data $\mathbf{o}$.

HMM as observation generator

1. Initialise $t = 1$;

2. If $t = 1$, choose state x_1 using π_i;
   Else, transit to x_t according to a_{ij};

3. Choose $o_t = k$ according to $b_j(k)$;

4. Increment t, and repeat from 2 until $t > T$.

Today's summary

- Fundamentals:
  - Least-squares and ML estimation
  - Bayes' theorem and MAP
  - Discrete vs. continuous pdfs
- Markov models
- Hidden Markov models

Three HMM problems for next time

1. Compute $P(\mathcal{O}|\lambda)$;

2. Find best X;

3. Re-estimate models $\Lambda = \{\lambda\}$.

Further reading

L. R. Rabiner. *A tutorial on HMM and selected applications in speech recognition*. In *Proc. IEEE*, Vol. 77, No. 2, pp. 257–286, 1989.