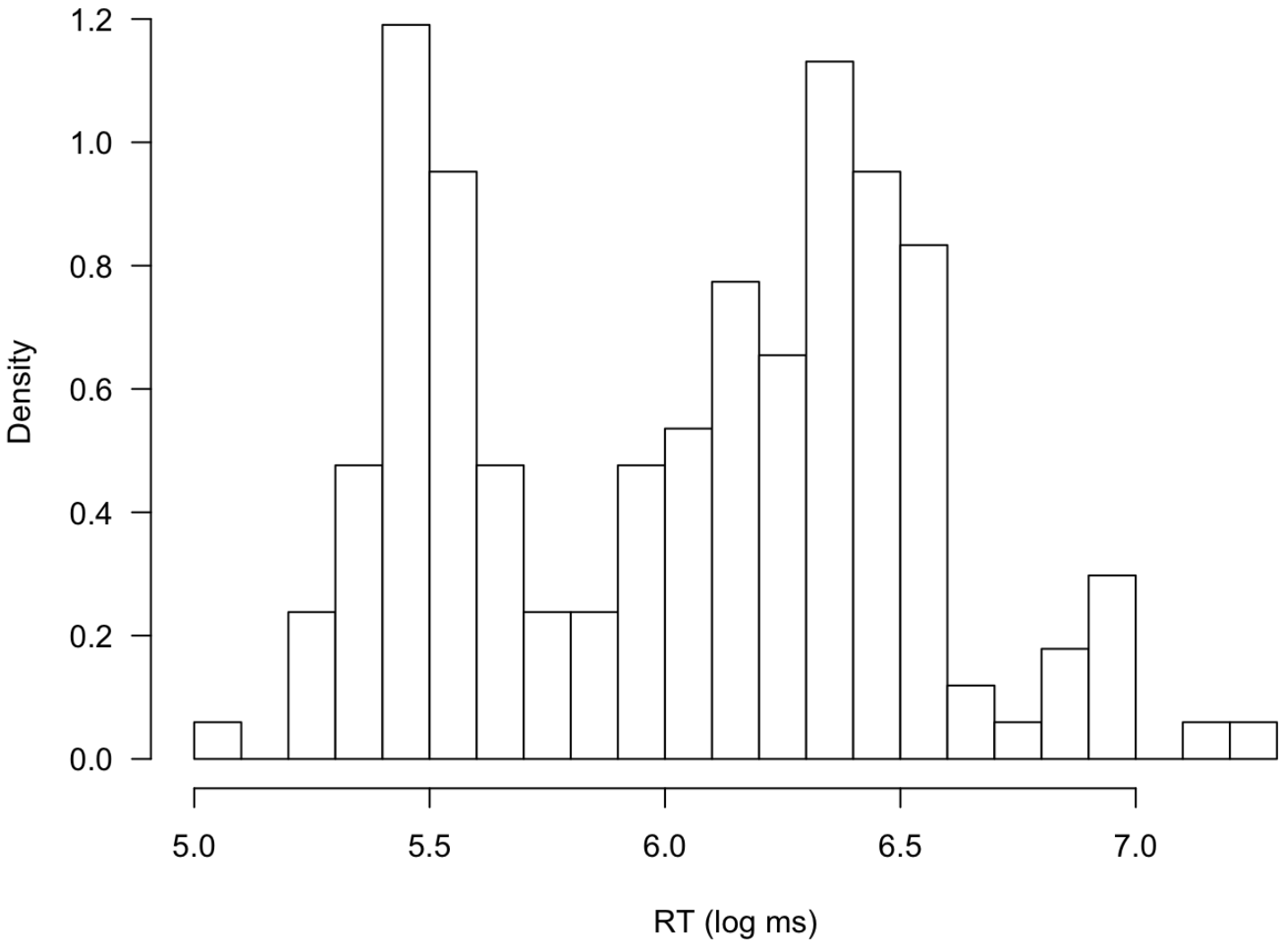
Hidden Markov models

Theory

Ingmar Visser

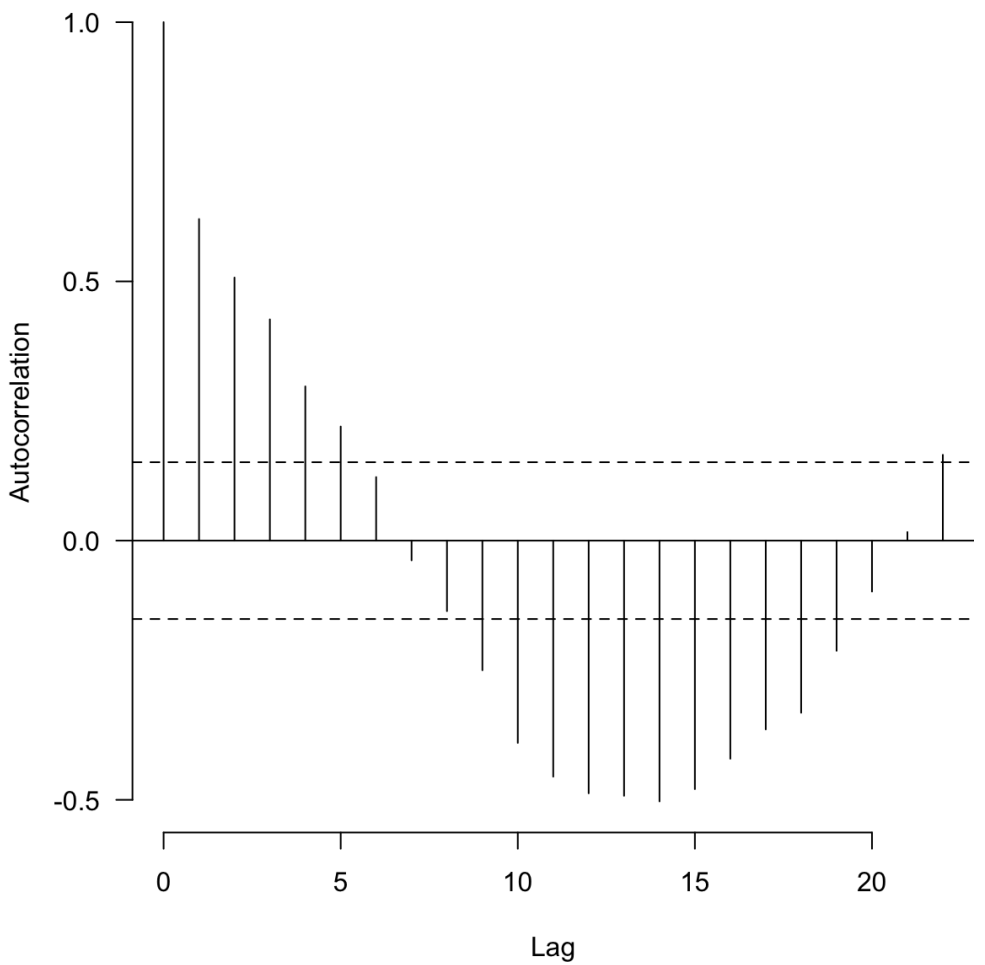
A motivating example

Consider the response times (in log ms) from 168 trials of a lexical decision experiment.



A motivating example

In mixture models, all the states S_t are independent. But if a set of observations $Y_{1:T}$ comes from a limited number of people, or is a time-series, this assumption is likely invalid.



A motivating example

To account for the sequential nature of observations, we can allow for dependence between the latent states.

A hidden Markov model can be viewed as an extension of mixture models in which we also account for the process according to which the latent states develop over time. In particular, we model the sequence of states $\mathrm{S}_{1:\mathrm{T}}$ as a first-order Markov process.

Markov models

In a *first-order* Markov process, the current state S_t depends on the history of the process only through the previous state, i.e.

$$P(S_t | S_{1:(t-1)}) = P(S_t | S_{t-1})$$

In a *homogeneous* first-order Markov process, the state transition probabilities are independent of the time index, i.e.

$$P(S_t = i | S_{t-1} = j) = P(S_{t-1} = i | S_{t-2} = j)$$

for all i, j, t.

Markov models

A homogeneous first-order Markov process is completely defined by the initial state probabilities

$$\pi = (P(S_1 = 1), \ldots, P(S_1 = N))$$

and the transition matrix

$$\mathbf{A} = \begin{pmatrix} P(S_t = 1|S_{t-1} = 1) & P(S_t = 2|S_{t-1} = 1) & \cdots & P(S_t = N|S_{t-1} = 1) \\ P(S_t = 1|S_{t-1} = 2) & P(S_t = 2|S_{t-1} = 2) & \cdots & P(S_t = N|S_{t-1} = 2) \\ \vdots & \vdots & \ddots & \vdots \\ P(S_t = 1|S_{t-1} = N) & P(S_t = 2|S_{t-1} = N) & \cdots & P(S_t = N|S_{t-1} = N) \end{pmatrix}$$

(note: in the transition matrix, rows are the current state, and columns the next state)

Markov models

As an example, consider a Markov process with initial state probabilities

$$\pi = (0.5, 0.5)$$

and transition matrix

$$\mathbf{A} = \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix}$$

We can then compute the probability $P(S_{t=2} = 2)$ as

$$\begin{aligned} P(S_{t=2} = 2) &= \sum_{i=1}^{N} P(S_{t=2} = 2 | S_{t=1} = i) P(S_{t=1} = i) \\ &= .2 \times .5 + .6 \times .5 = .4 \end{aligned}$$

or, using a little matrix algebra

$$\begin{aligned}\mathbf{p}_2 &= \pi \cdot \mathbf{A} \\ &= (.6, .4)\end{aligned}$$

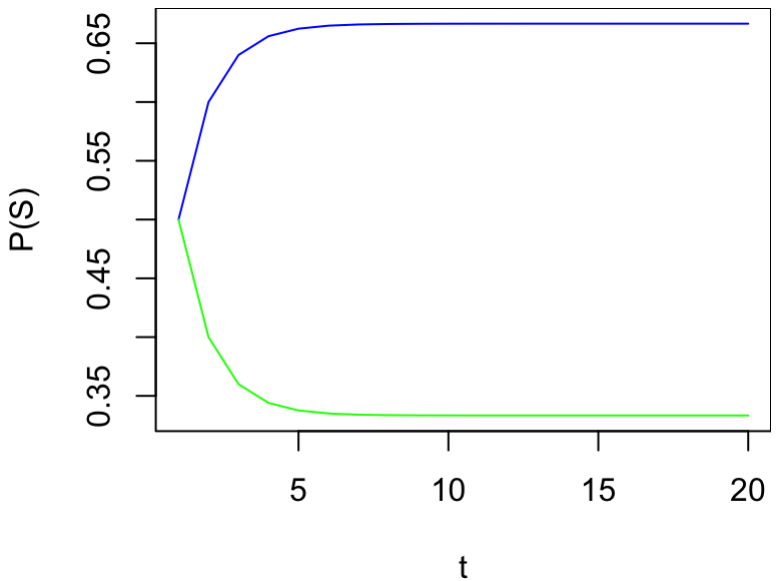
Markov models

This computation is recursive, e.g.:

$$\begin{aligned}\mathbf{p}_3 &= \mathbf{p}_2 \cdot \mathbf{A} \\ &= \pi \cdot \mathbf{A} \cdot \mathbf{A}\end{aligned}$$

and more generally

$$\mathbf{p}_t = \pi \cdot \mathbf{A}^{t-1}$$



For many models, the matrix power $\mathbf{A}^{t-1}$ will converge (as $t \to \infty$) to a matrix with identical rows. As a result, $\mathbf{p}_t$ will also converge to a value $\mathbf{p}_{stat}$ called the **stationary distribution**, defined as

$$\mathbf{p}_{stat} = \mathbf{p}_{stat} \cdot \mathbf{A}$$

For the example, the stationary distribution is:

$$\mathbf{p}_{stat} = (2/3, 1/3)$$

Hidden Markov models

As in a mixture model, in a *hidden* Markov model, the states are not directly observable (latent). Hidden Markov models usually model the state sequence as a homogenous first-order Markov process.

Such a hidden Markov model is defined by the

- observation distributions/densities: $p(Y_t | S_t)$
- state-transition distributions: $P(S_t | S_{t-1})$
- initial state distribution $P(S_1)$

This means that in addition to estimating the parameters of the component distributions and prior probabilities $P(S_1)$, we also need to determine the transition probabilities $P(S_t | S_{t-1})$.

Estimation and inference

The HMM likelihood

Let $\theta = (\theta_{\text{resp}}, \theta_{\text{trans}}, \theta_{\text{prior}})$ denote a general vector of parameters with in addition to those for MMs also elements θ_{trans} for the transition probabilities. The complete-data likelihood can be written as

$$\begin{aligned} L(\theta | y_{1:T}, s_{1:T}) &= p(y_{1:T}, s_{1:T} | \theta) \\ &= P(s_1 | \theta_{\text{prior}}) p(y_1 | s_1, \theta_{\text{resp}}) \times \prod_{t=2}^{T} P(s_t | s_{t-1}, \theta_{\text{trans}}) P(y_t | s_t, \theta_{\text{resp}}) \end{aligned}$$

To obtain the observed data likelihood function $L(\theta | y_{1:T})$ we could sum over all possible state sequences, but that is computationally infeasible. The number of possible state sequences is N^T; for a 2-state model with $T = 20$, this is already $2^{20} = 1048576$. The likelihood can be efficiently computed via the *Forward-Backward* algorithm.

Obtaining MLEs is conceptually similar to estimation for mixture models, and proceeds either through numerical optimization or the EM algorithm.

EM for HMM

As for MMs, the EM algorithm for HMMs works with the expected complete-data log likelihood

$$Q(\theta, \theta') = E_{\theta'} \left[\log P(S_1 | \theta_{prior}) + \log p(y_1 | S_1, \theta_{resp}) + \sum_{t=2}^{T} \log P(S_t | S_{t-1}, \theta_{trans}) + \sum_{t=2}^{T} \log P(y_t | S_t, \theta_{resp}) \right]$$

EM for HMM

This can be written as

$$
\begin{aligned}
Q(\theta, \theta') = & \sum_{j=1}^{N} \gamma_1(j) \log P(S_1 = j | \theta_{prior}) \\
& + \sum_{t=2}^{T} \sum_{j=1}^{N} \sum_{k=1}^{N} \xi_t(j,k) \log P(S_t = k | S_{t-1} = j, \theta_{trans}) \\
& + \sum_{t=1}^{T} \sum_{j=1}^{N} \gamma_t(j) \log p(y_t | S_t = j, \theta_{resp})
\end{aligned}
$$

where $\gamma_t(i) = P(S_t = i | y_{1:T}, \theta')$ as before and

$$
\xi_t(j,k) = P(S_t = k, S_{t-1} = j | y_{1:T}, \theta')
$$

Posterior state distribution

The posterior distribution of the state sequence, $S_{1:t}$, conditional upon observations $y_{1:t}$, can be defined recursively as

$$P(S_{1:t}|Y_{1:t}) = P(S_{1:(t-1)}|Y_{1:(t-1)}) \frac{P(S_t|S_{t-1})p(Y_t|S_t)}{p(Y_t|Y_{1:(t-1)})}$$

where * $P(S_{1:(t-1)}|Y_{1:(t-1)})$ is the posterior distribution at the previous time point * $P(S_t|S_{t-1})$ is the state transition distribution * $p(Y_t|S_t)$ is the conditional density of the observation * $p(Y_t|Y_{1:(t-1)})$ is the observation prediction distribution

Forward-Backward algorithm

The posterior distribution $P(S_t | y_{1:T})$ of the state at time t conditional upon all observed data $y_{1:T}$ is called the smoothing distribution. Smoothing distributions can be effectively computed by the **Forward-Backward algorithm**. This algorithm first makes a forwards pass through the data to compute the forward variables

$$\alpha_t(i) = p(y_{1:t}, S_t = i)$$

for $t = 1, \ldots, T$, and then makes a backwards pass through the data to compute the backwards variables

$$\beta_t(i) = p(y_{(t+1):T} | S_t = i)$$

for $t = T - 1, \ldots, 1$. Smoothing probabilities are then easily computed as

$$P(S_t = i | y_{1:T}) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}$$

Computing the likelihood

Using the forward variables, the likelihood is easily computed as

$$
\begin{aligned}
L(\theta|y_{1:T}) &= p(y_{1:t}|\theta) \\
&= \sum_{i=1}^{N} p(y_{1:T}, S_T = i|\theta) \\
&= \sum_{i=1}^{N} \alpha_T
\end{aligned}
$$

Estimating latent states: Viterbi

We can estimate the value of a state S_t as the state with the maximum smoothing probability $P(S_t | y_{1:T})$. However, the resulting sequence of estimated states may not be identical to the **maximum a posteriori** (MAP) state sequence:

$$s^*_{1:T} = \arg\max_{s_{1:t}} P(S_{1:T} = s_{1:T} | Y_{1:T})$$

Rather than searching over all possible state sequences, the MAP state sequence can be efficiently determined by the so-called Viterbi algorithm. In words, this algorithm uses ideas from dynamical programming to first determine the final element s^*_T in $s^*_{1:T}$, and passes backwards to determine which preceding state makes that one most probable.

Model selection

Model selection is similar to model selection in MMs.

- As for MMs, while a k-state HMM is theoretically nested under a $(\mathrm{k}+\mathrm{l})$-state HMM, the nesting relation is not unique and also involves fixing parameters to the bound of the parameter space. Hence, standard likelihood-ratio tests are not directly applicable, although approximate p-values can be obtained from a *parametric bootstrap*.
- Alternatively, model selection criteria such as the AIC and BIC can be used.