

# Hidden Markov model

**Hidden Markov Model** (**HMM**) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. *hidden*) states.

The hidden Markov model can be represented as the simplest dynamic Bayesian network. The mathematics behind the HMM were developed by L. E. Baum and coworkers.<sup>[1][2][3][4][5]</sup> HMM is closely related to an earlier work on the optimal nonlinear filtering problem by Ruslan L. Stratonovich,<sup>[6]</sup> who was the first to describe the forward-backward procedure.

In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters, while in the hidden Markov model, the state is not directly visible, but the output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states.

The adjective *hidden* refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a hidden Markov model even if these parameters are known exactly.

Hidden Markov models are especially known for their application in reinforcement learning and temporal pattern recognition such as speech, handwriting, gesture recognition,<sup>[7]</sup> part-of-speech tagging, musical score following,<sup>[8]</sup> partial discharges<sup>[9]</sup> and bioinformatics.

A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other. Recently, hidden Markov models have been generalized to pairwise Markov models and triplet Markov models which allow consideration of more complex data structures<sup>[10][11]</sup> and the modeling of nonstationary data.<sup>[12][13]</sup>

## Contents

- Description in terms of urns**
- Architecture**
- Inference**
  - Probability of an observed sequence
  - Probability of the latent variables
    - Filtering
    - Smoothing
    - Most likely explanation
  - Statistical significance
- A concrete example**
- Learning**
- Mathematical description**
  - General description
  - Compared with a simple mixture model
  - Examples

6.4	A two-level Bayesian HMM
6.5	Poisson hidden Markov model
7	Applications
8	History
9	Types
10	Extensions
11	See also
12	References
13	External links
13.1	Concepts
13.2	Software

# Description in terms of urns

In its discrete form, a hidden Markov process can be visualized as a generalization of the Urn problem with replacement (where each item from the urn is returned to the original urn before the next step).<sup>[14]</sup> Consider this example: in a room that is not visible to an observer there is a genie. The room contains urns X1, X2, X3, ... each of which contains a known mix of balls, each ball labeled y1, y2, y3, ... . The genie chooses an urn in that room and randomly draws a ball from that urn. It then puts the ball onto a conveyor belt, where the observer can observe the sequence of the balls but not the sequence of urns from which they were drawn. The genie has some procedure to choose urns; the choice of the urn for the *n*-th ball depends only upon a random number and the choice of the urn for the (*n* – 1)-th ball. The choice of urn does not directly depend on the urns chosen before this single previous urn; therefore, this is called a Markov process. It can be described by the upper part of Figure 1.

The Markov process itself cannot be observed, only the sequence of labeled balls, thus this arrangement is called a "hidden Markov process". This is illustrated by the lower part of the diagram shown in Figure 1, where one can see that balls y1, y2, y3, y4 can be drawn at each state. Even if the observer knows the composition of the urns and has just observed a sequence of three balls, *e.g.* y1, y2 and y3 on the conveyor belt, the observer still cannot be *sure* which urn (*i.e.*, at which state) the genie has drawn the third ball from. However, the observer can work out other information, such as the likelihood that the third ball came from each of the urns.

# Architecture

The diagram below shows the general architecture of an instantiated HMM. Each oval shape represents a random variable that can adopt any of a number of values. The random variable *x(t)* is the hidden state at time *t* (with the model from the above diagram, *x(t)* ∈ { *x*<sub>1</sub>, *x*<sub>2</sub>, *x*<sub>3</sub> }). The random variable *y(t)* is the observation at time *t* (with

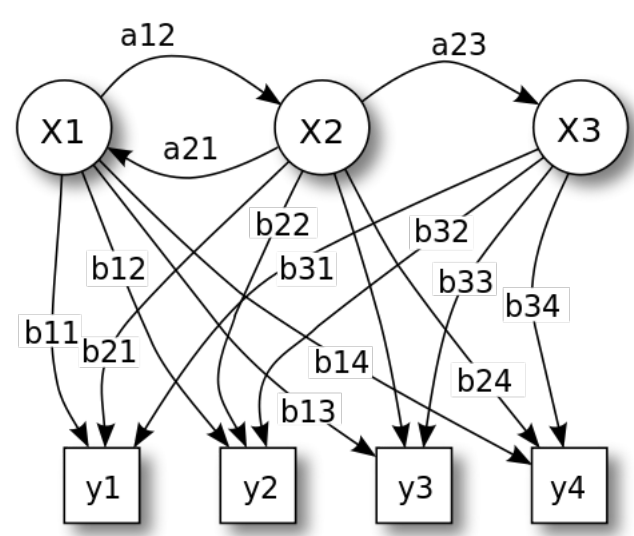


Figure 1. Probabilistic parameters of a hidden Markov model (example)  
*X* — states  
*y* — possible observations  
*a* — state transition probabilities  
*b* — output probabilities

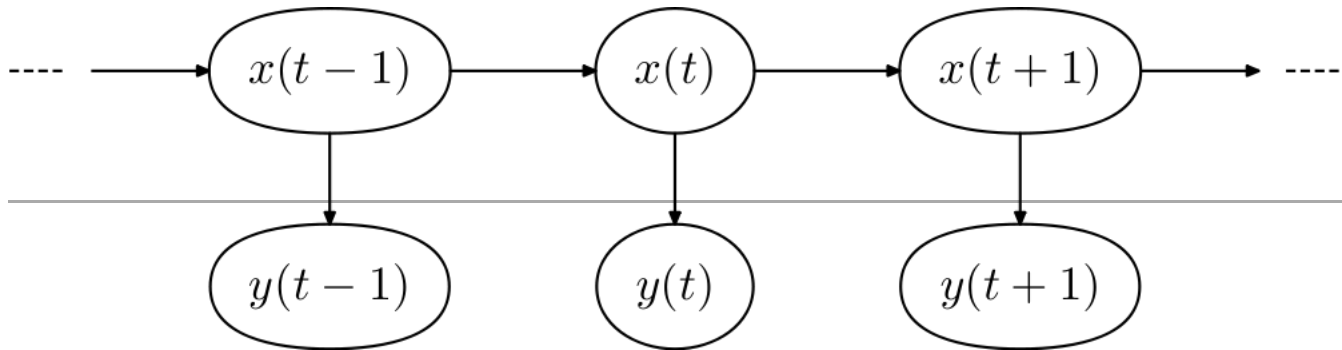
$y(t) \in \{y_1, y_2, y_3, y_4\}$ ). The arrows in the diagram (often called a trellis diagram) denote conditional dependencies.

From the diagram, it is clear that the conditional probability distribution of the hidden variable  $x(t)$  at time  $t$ , given the values of the hidden variable  $x$  at all times, depends *only* on the value of the hidden variable  $x(t - 1)$ ; the values at time  $t - 2$  and before have no influence. This is called the Markov property. Similarly, the value of the observed variable  $y(t)$  only depends on the value of the hidden variable  $x(t)$  (both at time  $t$ ).

In the standard type of hidden Markov model considered here, the state space of the hidden variables is discrete, while the observations themselves can either be discrete (typically generated from a categorical distribution) or continuous (typically from a Gaussian distribution). The parameters of a hidden Markov model are of two types, *transition probabilities* and *emission probabilities* (also known as *output probabilities*). The transition probabilities control the way the hidden state at time  $t$  is chosen given the hidden state at time  $t - 1$ .

The hidden state space is assumed to consist of one of  $N$  possible values, modeled as a categorical distribution. (See the section below on extensions for other possibilities.) This means that for each of the  $N$  possible states that a hidden variable at time  $t$  can be in, there is a transition probability from this state to each of the  $N$  possible states of the hidden variable at time  $t + 1$ , for a total of  $N^2$  transition probabilities. Note that the set of transition probabilities for transitions from any given state must sum to 1. Thus, the  $N \times N$  matrix of transition probabilities is a Markov matrix. Because any one transition probability can be determined once the others are known, there are a total of  $N(N - 1)$  transition parameters.

In addition, for each of the  $N$  possible states, there is a set of emission probabilities governing the distribution of the observed variable at a particular time given the state of the hidden variable at that time. The size of this set depends on the nature of the observed variable. For example, if the observed variable is discrete with  $M$  possible values, governed by a categorical distribution, there will be  $M - 1$  separate parameters, for a total of  $N(M - 1)$  emission parameters over all hidden states. On the other hand, if the observed variable is an  $M$ -dimensional vector distributed according to an arbitrary multivariate Gaussian distribution, there will be  $M$  parameters controlling the means and  $\frac{M(M + 1)}{2}$  parameters controlling the covariance matrix, for a total of  $N \left( M + \frac{M(M + 1)}{2} \right) = \frac{NM(M + 3)}{2} = O(NM^2)$  emission parameters. (In such a case, unless the value of  $M$  is small, it may be more practical to restrict the nature of the covariances between individual elements of the observation vector, e.g. by assuming that the elements are independent of each other, or less restrictively, are independent of all but a fixed number of adjacent elements.)



# Inference

Several inference problems are associated with hidden Markov models, as outlined below.

## Probability of an observed sequence



where the sum runs over all possible hidden-node sequences

$$X = x(0), x(1), \dots, x(L - 1).$$

Applying the principle of dynamic programming, this problem, too, can be handled efficiently using the forward algorithm.

## Probability of the latent variables

A number of related tasks ask about the probability of one or more of the latent variables, given the model's parameters and a sequence of observations  $y(1), \dots, y(t)$ .

### Filtering

The task is to compute, given the model's parameters and a sequence of observations, the distribution over hidden states of the last latent variable at the end of the sequence, i.e. to compute  $P(x(t) \mid y(1), \dots, y(t))$ . This task is normally used when the sequence of latent variables is thought of as the underlying states that a process moves through at a sequence of points of time, with corresponding observations at each point in time. Then, it is natural to ask about the state of the process at the end.

This problem can be handled efficiently using the forward algorithm.

### Smoothing

This is similar to filtering but asks about the distribution of a latent variable somewhere in the middle of a sequence, i.e. to compute  $P(x(k) \mid y(1), \dots, y(t))$  for some  $k < t$ . From the perspective described above, this can be thought of as the probability distribution over hidden states for a point in time  $k$  in the past, relative to time  $t$ .

The forward-backward algorithm is an efficient method for computing the smoothed values for all hidden state variables.

### Most likely explanation

The task, unlike the previous two, asks about the joint probability of the *entire* sequence of hidden states that generated a particular sequence of observations (see illustration on the right). This task is generally applicable when HMM's are applied to different sorts of problems from those for which the tasks of filtering and smoothing are applicable. An example is part-of-speech tagging, where the hidden states represent the underlying parts of speech corresponding to an observed sequence of words. In this case, what is of interest is the entire sequence of parts of speech, rather than simply the part of speech for a single word, as filtering or smoothing would compute.

This task requires finding a maximum over all possible state sequences, and can be solved efficiently by the Viterbi algorithm.

## Statistical significance

For some of the above problems, it may also be interesting to ask about statistical significance. What is the probability that a sequence drawn from some null distribution will have an HMM probability (in the case of the forward algorithm) or a maximum state sequence probability (in the case of the Viterbi algorithm) at least as large as that of a

particular output sequence?<sup>[15]</sup> When an HMM is used to evaluate the relevance of a hypothesis for a particular output sequence, the statistical significance indicates the false positive rate associated with failing to reject the hypothesis for the output sequence.

## A concrete example

Consider two friends, Alice and Bob, who live far apart from each other and who talk together daily over the telephone about what they did that day. Bob is only interested in three activities: walking in the park, shopping, and cleaning his apartment. The choice of what to do is determined exclusively by the weather on a given day. Alice has no definite information about the weather, but she knows general trends. Based on what Bob tells her he did each day, Alice tries to guess what the weather must have been like.

Alice believes that the weather operates as a discrete Markov chain. There are two states, "Rainy" and "Sunny", but she cannot observe them directly, that is, they are *hidden* from her. On each day, there is a certain chance that Bob will perform one of the following activities, depending on the weather: "walk", "shop", or "clean". Since Bob tells Alice about his activities, those are the *observations*. The entire system is that of a hidden Markov model (HMM).

Alice knows the general weather trends in the area, and what Bob likes to do on average. In other words, the parameters of the HMM are known. They can be represented as follows in Python:

```
states = ( 'Rainy', 'Sunny' )

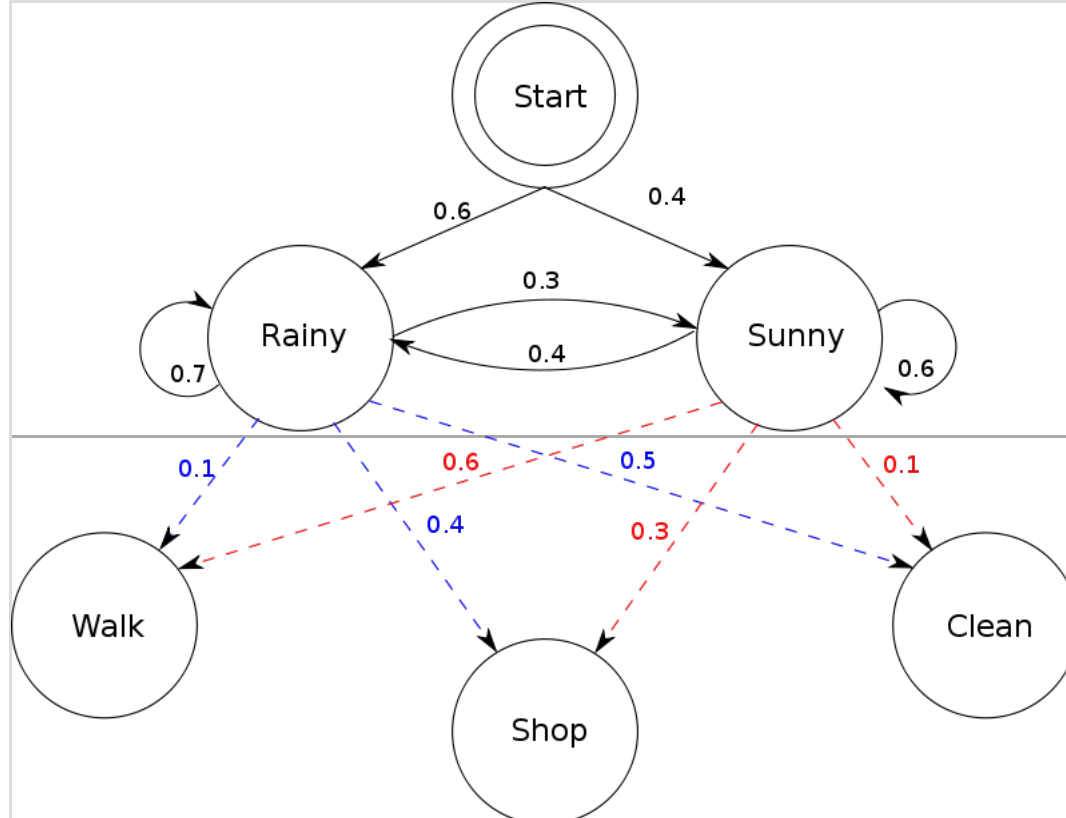
observations = ( 'walk', 'shop', 'clean' )

start_probability = { 'Rainy': 0.6, 'Sunny': 0.4 }

transition_probability = {
    'Rainy' : { 'Rainy': 0.7, 'Sunny': 0.3 },
    'Sunny' : { 'Rainy': 0.4, 'Sunny': 0.6 },
}

emission_probability = {
    'Rainy' : { 'walk': 0.1, 'shop': 0.4, 'clean': 0.5 },
    'Sunny' : { 'walk': 0.6, 'shop': 0.3, 'clean': 0.1 },
}
```

In this piece of code, `start_probability` represents Alice's belief about which state the HMM is in when Bob first calls her (all she knows is that it tends to be rainy on average). The particular probability distribution used here is not the equilibrium one, which is (given the transition probabilities) approximately `{ 'Rainy': 0.57, 'Sunny': 0.43 }`. The `transition_probability` represents the change of the weather in the underlying Markov chain. In this example, there is only a 30% chance that tomorrow will be sunny if today is rainy. The `emission_probability` represents how likely Bob is to perform a certain activity on each day. If it is rainy, there is a 50% chance that he is cleaning his apartment; if it is sunny, there is a 60% chance that he is outside for a walk.



A similar example is further elaborated in the [Viterbi algorithm](#) page.

# Learning

The parameter learning task in HMMs is to find, given an output sequence or a set of such sequences, the best set of state transition and emission probabilities. The task is usually to derive the maximum likelihood estimate of the parameters of the HMM given the set of output sequences. No tractable algorithm is known for solving this problem exactly, but a local maximum likelihood can be derived efficiently using the Baum–Welch algorithm or the Baldi–Chauvin algorithm. The Baum–Welch algorithm is a special case of the expectation-maximization algorithm. If the HMMs are used for time series prediction, more sophisticated Bayesian inference methods, like Markov chain Monte Carlo (MCMC) sampling are proven to be favorable over finding a single maximum likelihood model both in terms of accuracy and stability.<sup>[16]</sup> Since MCMC imposes significant computational burden, in cases where computational scalability is also of interest, one may alternatively resort to variational approximations to Bayesian inference, e.g.<sup>[17]</sup> Indeed, approximate variational inference offers computational efficiency comparable to expectation-maximization, while yielding an accuracy profile only slightly inferior to exact MCMC-type Bayesian inference.

# Mathematical description

## General description

A basic hidden Markov model can be described as follows:

$N$	=	number of states
$T$	=	number of observations
$\theta_{i=1\dots N}$	=	emission parameter for an observation associated with state $i$
$\phi_{i=1\dots N,j=1\dots N}$	=	probability of transition from state $i$ to state $j$
$\phi_{i=1\dots N}$	=	$N$ -dimensional vector, composed of $\phi_{i,1\dots N}$ ; must sum to 1, the row of the matrix $\phi_{i=1\dots N,j=1\dots N}$
$x_{t=1\dots T}$	=	(hidden) state at time $t$
$y_{t=1\dots T}$	=	observation at time $t$

$$\begin{aligned}
\boldsymbol{F}(\boldsymbol{y}|\boldsymbol{\theta}) &= \text{probability distribution of an observation, parametrized on } \boldsymbol{\theta} \\
\boldsymbol{x}_{t=2\dots T} &\sim \text{Categorical}(\boldsymbol{\phi}_{\boldsymbol{x}_{t-1}}) \\
\boldsymbol{y}_{t=1\dots T} &\sim \boldsymbol{F}(\boldsymbol{\theta}_{\boldsymbol{x}_t})
\end{aligned}$$

Note that, in the above model (and also the one below), the prior distribution of the initial state  $\boldsymbol{x}_1$  is not specified. Typical learning models correspond to assuming a discrete uniform distribution over possible states (i.e. no particular prior distribution is assumed).

In a Bayesian setting, all parameters are associated with random variables, as follows:

$$\begin{aligned}
N, T &= \text{as above} \\
\boldsymbol{\theta}_{i=1\dots N}, \boldsymbol{\phi}_{i=1\dots N, j=1\dots N}, \boldsymbol{\phi}_{i=1\dots N} &= \text{as above} \\
\boldsymbol{x}_{t=1\dots T}, \boldsymbol{y}_{t=1\dots T}, \boldsymbol{F}(\boldsymbol{y}|\boldsymbol{\theta}) &= \text{as above} \\
\boldsymbol{\alpha} &= \text{shared hyperparameter for emission parameters} \\
\boldsymbol{\beta} &= \text{shared hyperparameter for transition parameters} \\
\boldsymbol{H}(\boldsymbol{\theta}|\boldsymbol{\alpha}) &= \text{prior probability distribution of emission parameters, parametrized on } \boldsymbol{\alpha} \\
\boldsymbol{\theta}_{i=1\dots N} &\sim \boldsymbol{H}(\boldsymbol{\alpha}) \\
\boldsymbol{\phi}_{i=1\dots N} &\sim \text{Symmetric-Dirichlet}_N(\boldsymbol{\beta}) \\
\boldsymbol{x}_{t=2\dots T} &\sim \text{Categorical}(\boldsymbol{\phi}_{\boldsymbol{x}_{t-1}}) \\
\boldsymbol{y}_{t=1\dots T} &\sim \boldsymbol{F}(\boldsymbol{\theta}_{\boldsymbol{x}_t})
\end{aligned}$$

These characterizations use  $\boldsymbol{F}$  and  $\boldsymbol{H}$  to describe arbitrary distributions over observations and parameters, respectively. Typically  $\boldsymbol{H}$  will be the conjugate prior of  $\boldsymbol{F}$ . The two most common choices of  $\boldsymbol{F}$  are Gaussian and categorical; see below.

## Compared with a simple mixture model

As mentioned above, the distribution of each observation in a hidden Markov model is a mixture density, with the states of the corresponding to mixture components. It is useful to compare the above characterizations for an HMM with the corresponding characterizations, of a mixture model, using the same notation.

A non-Bayesian mixture model:

$$\begin{aligned}
N &= \text{number of mixture components} \\
T &= \text{number of observations} \\
\boldsymbol{\theta}_{i=1\dots N} &= \text{parameter of distribution of observation associated with component } i \\
\boldsymbol{\phi}_{i=1\dots N} &= \text{mixture weight, i.e. prior probability of component } i \\
\boldsymbol{\phi} &= N\text{-dimensional vector, composed of } \boldsymbol{\phi}_{1\dots N}; \text{ must sum to } \mathbf{1} \\
\boldsymbol{x}_{t=1\dots T} &= \text{component of observation } t \\
\boldsymbol{y}_{t=1\dots T} &= \text{observation } t \\
\boldsymbol{F}(\boldsymbol{y}|\boldsymbol{\theta}) &= \text{probability distribution of an observation, parametrized on } \boldsymbol{\theta} \\
\boldsymbol{x}_{t=1\dots T} &\sim \text{Categorical}(\boldsymbol{\phi}) \\
\boldsymbol{y}_{t=1\dots T} &\sim \boldsymbol{F}(\boldsymbol{\theta}_{\boldsymbol{x}_t})
\end{aligned}$$



A Bayesian mixture model:

$N, T$	=	as above
$\theta_{i=1 \dots N}, \phi_{i=1 \dots N}, \phi$	=	as above
$x_{t=1 \dots T}, y_{t=1 \dots T}, F(y \theta)$	=	as above
$\alpha$	=	shared hyperparameter for component parameters
$\beta$	=	shared hyperparameter for mixture weights
$H(\theta \alpha)$	=	prior probability distribution of component parameters, parametrized on $\alpha$
$\theta_{i=1 \dots N}$	$\sim$	$H(\alpha)$
$\phi$	$\sim$	$\text{Symmetric-Dirichlet}_N(\beta)$
$x_{t=1 \dots T}$	$\sim$	$\text{Categorical}(\phi)$
$y_{t=1 \dots T}$	$\sim$	$F(\theta_{x_t})$

## Examples

The following mathematical descriptions are fully written out and explained, for ease of implementation.

A typical non-Bayesian HMM with Gaussian observations looks like this:

$N$	=	number of states
$T$	=	number of observations
$\phi_{i=1 \dots N, j=1 \dots N}$	=	probability of transition from state $i$ to state $j$
$\phi_{i=1 \dots N}$	=	$N$ -dimensional vector, composed of $\phi_{i, 1 \dots N}$ ; must sum to 1
$\mu_{i=1 \dots N}$	=	mean of observations associated with state $i$
$\sigma_{i=1 \dots N}^2$	=	variance of observations associated with state $i$
$x_{t=1 \dots T}$	=	state of observation at time $t$
$y_{t=1 \dots T}$	=	observation at time $t$
$x_{t=2 \dots T}$	$\sim$	$\text{Categorical}(\phi_{x_{t-1}})$
$y_{t=1 \dots T}$	$\sim$	$\mathcal{N}(\mu_{x_t}, \sigma_{x_t}^2)$

A typical Bayesian HMM with Gaussian observations looks like this:

$N$	=	number of states
$T$	=	number of observations
$\phi_{i=1 \dots N, j=1 \dots N}$	=	probability of transition from state $i$ to state $j$
$\phi_{i=1 \dots N}$	=	$N$ -dimensional vector, composed of $\phi_{i, 1 \dots N}$ ; must sum to 1
$\mu_{i=1 \dots N}$	=	mean of observations associated with state $i$
$\sigma_{i=1 \dots N}^2$	=	variance of observations associated with state $i$
$x_{t=1 \dots T}$	=	state of observation at time $t$
$y_{t=1 \dots T}$	=	observation at time $t$
$\beta$	=	concentration hyperparameter controlling the density of the transition matrix

$\mu_0, \lambda$	=	shared hyperparameters of the means for each state
$\nu, \sigma_0^2$	=	shared hyperparameters of the variances for each state
$\phi_{i=1\dots N}$	$\sim$	$\text{Symmetric-Dirichlet}_N(\beta)$
$x_{t=2\dots T}$	$\sim$	$\text{Categorical}(\phi_{x_{t-1}})$
$\mu_{i=1\dots N}$	$\sim$	$\mathcal{N}(\mu_0, \lambda\sigma_i^2)$
$\sigma_{i=1\dots N}^2$	$\sim$	$\text{Inverse-Gamma}(\nu, \sigma_0^2)$
$y_{t=1\dots T}$	$\sim$	$\mathcal{N}(\mu_{x_t}, \sigma_{x_t}^2)$

A typical non-Bayesian HMM with categorical observations looks like this:

$N$	=	number of states
$T$	=	number of observations
$\phi_{i=1\dots N, j=1\dots N}$	=	probability of transition from state $i$ to state $j$
$\phi_{i=1\dots N}$	=	$N$ -dimensional vector, composed of $\phi_{i,1\dots N}$ ; must sum to 1
$V$	=	dimension of categorical observations, e.g. size of word vocabulary
$\theta_{i=1\dots N, j=1\dots V}$	=	probability for state $i$ of observing the $j$ th item
$\theta_{i=1\dots N}$	=	$V$ -dimensional vector, composed of $\theta_{i,1\dots V}$ ; must sum to 1
$x_{t=1\dots T}$	=	state of observation at time $t$
$y_{t=1\dots T}$	=	observation at time $t$
$x_{t=2\dots T}$	$\sim$	$\text{Categorical}(\phi_{x_{t-1}})$
$y_{t=1\dots T}$	$\sim$	$\text{Categorical}(\theta_{x_t})$

A typical Bayesian HMM with categorical observations looks like this:

$N$	=	number of states
$T$	=	number of observations
$\phi_{i=1\dots N, j=1\dots N}$	=	probability of transition from state $i$ to state $j$
$\phi_{i=1\dots N}$	=	$N$ -dimensional vector, composed of $\phi_{i,1\dots N}$ ; must sum to 1
$V$	=	dimension of categorical observations, e.g. size of word vocabulary
$\theta_{i=1\dots N, j=1\dots V}$	=	probability for state $i$ of observing the $j$ th item
$\theta_{i=1\dots N}$	=	$V$ -dimensional vector, composed of $\theta_{i,1\dots V}$ ; must sum to 1
$x_{t=1\dots T}$	=	state of observation at time $t$
$y_{t=1\dots T}$	=	observation at time $t$
$\alpha$	=	shared concentration hyperparameter of $\theta$ for each state
$\beta$	=	concentration hyperparameter controlling the density of the transition matrix
$\phi_{i=1\dots N}$	$\sim$	$\text{Symmetric-Dirichlet}_N(\beta)$
$\theta_{1\dots V}$	$\sim$	$\text{Symmetric-Dirichlet}_V(\alpha)$
$x_{t=2\dots T}$	$\sim$	$\text{Categorical}(\phi_{x_{t-1}})$
$y_{t=1\dots T}$	$\sim$	$\text{Categorical}(\theta_{x_t})$

Note that in the above Bayesian characterizations,  $\beta$  (a concentration parameter) controls the density of the transition matrix. That is, with a high value of  $\beta$  (significantly above 1), the probabilities controlling the transition out of a particular state will all be similar, meaning there will be a significant probability of transitioning to any of the other states. In other words, the path followed by the Markov chain of hidden states will be highly random. With a low value of  $\beta$  (significantly below 1), only a small number of the possible transitions out of a given state will have significant probability, meaning that the path followed by the hidden states will be somewhat predictable.

## A two-level Bayesian HMM

An alternative for the above two Bayesian examples would be to add another level of prior parameters for the transition matrix. That is, replace the lines

$$\begin{aligned} \beta &= \text{concentration hyperparameter controlling the density of the transition matrix} \\ \phi_{i=1 \dots N} &\sim \text{Symmetric-Dirichlet}_N(\beta) \end{aligned}$$

with the following:

$$\begin{aligned} \gamma &= \text{concentration hyperparameter controlling how many states are intrinsically likely} \\ \beta &= \text{concentration hyperparameter controlling the density of the transition matrix} \\ \boldsymbol{\eta} &= N\text{-dimensional vector of probabilities, specifying the intrinsic probability of a given state} \\ \boldsymbol{\eta} &\sim \text{Symmetric-Dirichlet}_N(\gamma) \\ \phi_{i=1 \dots N} &\sim \text{Dirichlet}_N(\beta N \boldsymbol{\eta}) \end{aligned}$$

What this means is the following:

1.  $\boldsymbol{\eta}$  is a probability distribution over states, specifying which states are inherently likely. The greater the probability of a given state in this vector, the more likely is a transition to that state (regardless of the starting state).
2.  $\gamma$  controls the density of  $\boldsymbol{\eta}$ . Values significantly above 1 cause a dense vector where all states will have similar prior probabilities. Values significantly below 1 cause a sparse vector where only a few states are inherently likely (have prior probabilities significantly above 0).
3.  $\beta$  controls the density of the transition matrix, or more specifically, the density of the  $N$  different probability vectors  $\phi_{i=1 \dots N}$  specifying the probability of transitions out of state  $i$  to any other state.

Imagine that the value of  $\beta$  is significantly above 1. Then the different  $\phi$  vectors will be dense, i.e. the probability mass will be spread out fairly evenly over all states. However, to the extent that this mass is unevenly spread,  $\boldsymbol{\eta}$  controls which states are likely to get more mass than others.

Now, imagine instead that  $\beta$  is significantly below 1. This will make the  $\phi$  vectors sparse, i.e. almost all the probability mass is distributed over a small number of states, and for the rest, a transition to that state will be very unlikely. Notice that there are different  $\phi$  vectors for each starting state, and so even if all the vectors are sparse, different vectors may distribute the mass to different ending states. However, for all of the vectors,  $\boldsymbol{\eta}$  controls which ending states are likely to get mass assigned to them. For example, if  $\beta$  is 0.1, then each  $\phi$  will be sparse and, for any given starting state  $i$ , the set of states  $\mathbf{J}_i$  to which transitions are likely to occur will be very small, typically having only one or two members. Now, if the probabilities in  $\boldsymbol{\eta}$  are all the same (or equivalently, one of the above models without  $\boldsymbol{\eta}$  is used), then for different  $i$ , there will be different states in the corresponding  $\mathbf{J}_i$ , so that all states are equally likely to occur in any given  $\mathbf{J}_i$ . On the other hand, if the values in  $\boldsymbol{\eta}$  are unbalanced, so that one state has a much higher probability than others, almost all  $\mathbf{J}_i$  will contain this state; hence, regardless of the starting state, transitions will nearly always occur to this given state.

Hence, a two-level model such as just described allows independent control over (1) the overall density of the transition matrix, and (2) the density of states to which transitions are likely (i.e. the density of the prior distribution of states in any particular hidden variable  $\boldsymbol{x}_i$ ). In both cases this is done while still assuming ignorance over which particular states are more likely than others. If it is desired to inject this information into the model, the probability vector  $\boldsymbol{\eta}$  can be directly specified; or, if there is less certainty about these relative probabilities, a non-symmetric Dirichlet distribution can be used as the prior distribution over  $\boldsymbol{\eta}$ . That is, instead of using a symmetric Dirichlet distribution with the single parameter  $\gamma$  (or equivalently, a general Dirichlet with a vector all of whose values are equal to  $\gamma$ ), use a general Dirichlet with values that are variously greater or less than  $\gamma$ , according to which state is more or less preferred.

## Poisson hidden Markov model

*Poisson hidden Markov models (PHMM)* are special cases of hidden Markov models where a Poisson process has a rate which varies in association with changes between the different states of a Markov model.<sup>[18]</sup> PHMMs are not necessarily Markovian processes themselves because the underlying Markov chain or Markov process cannot be observed and only the Poisson signal is observed.

# Applications

---

HMMs can be applied in many fields where the goal is to recover a data sequence that is not immediately observable (but other data that depend on the sequence are). Applications include:

- Computational finance<sup>[19][20]</sup>
- Single Molecule Kinetic analysis<sup>[21]</sup>
- Cryptanalysis
- Speech recognition
- Speech synthesis
- Part-of-speech tagging
- Document Separation in scanning solutions
- Machine translation
- Partial discharge
- Gene prediction
- Handwriting Recognition
- Alignment of bio-sequences
- Time Series Analysis
- Activity recognition
- Protein folding<sup>[22]</sup>
- Metamorphic Virus Detection<sup>[23]</sup>
- DNA Motif Discovery<sup>[24]</sup>

# History

---

The forward and backward recursions used in HMM as well as computations of marginal smoothing probabilities were first described by Ruslan L. Stratonovich in 1960<sup>[6]</sup> (pages 160—162) and in the late 1950s in his papers in Russian. The Hidden Markov Models were later described in a series of statistical papers by Leonard E. Baum and other authors in the second half of the 1960s. One of the first applications of HMMs was speech recognition, starting in the mid-1970s.<sup>[25][26][27][28]</sup>

In the second half of the 1980s, HMMs began to be applied to the analysis of biological sequences,<sup>[29]</sup> in particular DNA. Since then, they have become ubiquitous in the field of bioinformatics.<sup>[30]</sup>

## Types

---

Hidden Markov models can model complex Markov processes where the states emit the observations according to some probability distribution. One such example is the Gaussian distribution, in such a Hidden Markov Model the states output are represented by a Gaussian distribution.

Moreover, it could represent even more complex behavior when the output of the states is represented as mixture of two or more Gaussians, in which case the probability of generating an observation is the product of the probability of first selecting one of the Gaussians and the probability of generating that observation from that Gaussian. In cases of modeled data exhibiting artifacts such as outliers and skewness, one may resort to finite mixtures of heavier-tailed elliptical distributions, such as the multivariate Student's-t distribution, or appropriate non-elliptical distributions, such as the multivariate Normal Inverse-Gaussian.<sup>[31]</sup>

## Extensions

---

In the hidden Markov models considered above, the state space of the hidden variables is discrete, while the observations themselves can either be discrete (typically generated from a categorical distribution) or continuous (typically from a Gaussian distribution). Hidden Markov models can also be generalized to allow continuous state spaces. Examples of such models are those where the Markov process over hidden variables is a linear dynamical system, with a linear relationship among related variables and where all hidden and observed variables follow a Gaussian distribution. In simple cases, such as the linear dynamical system just mentioned, exact inference is tractable (in this case, using the Kalman filter); however, in general, exact inference in HMMs with continuous latent variables is infeasible, and approximate methods must be used, such as the extended Kalman filter or the particle filter.

Hidden Markov models are generative models, in which the joint distribution of observations and hidden states, or equivalently both the prior distribution of hidden states (the *transition probabilities*) and conditional distribution of observations given states (the *emission probabilities*), is modeled. The above algorithms implicitly assume a uniform prior distribution over the transition probabilities. However, it is also possible to create hidden Markov models with other types of prior distributions. An obvious candidate, given the categorical distribution of the transition probabilities, is the Dirichlet distribution, which is the conjugate prior distribution of the categorical distribution. Typically, a symmetric Dirichlet distribution is chosen, reflecting ignorance about which states are inherently more likely than others. The single parameter of this distribution (termed the *concentration parameter*) controls the relative density or sparseness of the resulting transition matrix. A choice of 1 yields a uniform distribution. Values greater than 1 produce a dense matrix, in which the transition probabilities between pairs of states are likely to be nearly equal. Values less than 1 result in a sparse matrix in which, for each given source state, only a small number of destination states have non-negligible transition probabilities. It is also possible to use a two-level prior Dirichlet distribution, in which one Dirichlet distribution (the upper distribution) governs the parameters of another Dirichlet distribution (the lower distribution), which in turn governs the transition probabilities. The upper distribution governs the overall distribution of states, determining how likely each state is to occur; its concentration parameter determines the density or sparseness of states. Such a two-level prior distribution, where both concentration parameters are set to produce sparse distributions, might be useful for example in unsupervised part-of-speech tagging, where some parts of speech occur much more commonly than others; learning algorithms that assume a

uniform prior distribution generally perform poorly on this task. The parameters of models of this sort, with non-uniform prior distributions, can be learned using Gibbs sampling or extended versions of the expectation-maximization algorithm.

An extension of the previously described hidden Markov models with Dirichlet priors uses a Dirichlet process in place of a Dirichlet distribution. This type of model allows for an unknown and potentially infinite number of states. It is common to use a two-level Dirichlet process, similar to the previously described model with two levels of Dirichlet distributions. Such a model is called a *hierarchical Dirichlet process hidden Markov model*, or *HDP-HMM* for short. It was originally described under the name "Infinite Hidden Markov Model"<sup>[4]</sup> and was further formalized in<sup>[5]</sup>.

A different type of extension uses a discriminative model in place of the generative model of standard HMMs. This type of model directly models the conditional distribution of the hidden states given the observations, rather than modeling the joint distribution. An example of this model is the so-called maximum entropy Markov model (MEMM), which models the conditional distribution of the states using logistic regression (also known as a "maximum entropy model"). The advantage of this type of model is that arbitrary features (i.e. functions) of the observations can be modeled, allowing domain-specific knowledge of the problem at hand to be injected into the model. Models of this sort are not limited to modeling direct dependencies between a hidden state and its associated observation; rather, features of nearby observations, of combinations of the associated observation and nearby observations, or in fact of arbitrary observations at any distance from a given hidden state can be included in the process used to determine the value of a hidden state. Furthermore, there is no need for these features to be statistically independent of each other, as would be the case if such features were used in a generative model. Finally, arbitrary features over pairs of adjacent hidden states can be used rather than simple transition probabilities. The disadvantages of such models are: (1) The types of prior distributions that can be placed on hidden states are severely limited; (2) It is not possible to predict the probability of seeing an arbitrary observation. This second limitation is often not an issue in practice, since many common usages of HMM's do not require such predictive probabilities.

A variant of the previously described discriminative model is the linear-chain conditional random field. This uses an undirected graphical model (aka Markov random field) rather than the directed graphical models of MEMM's and similar models. The advantage of this type of model is that it does not suffer from the so-called *label bias* problem of MEMM's, and thus may make more accurate predictions. The disadvantage is that training can be slower than for MEMM's.

Yet another variant is the *factorial hidden Markov model*, which allows for a single observation to be conditioned on the corresponding hidden variables of a set of  $K$  independent Markov chains, rather than a single Markov chain. It is equivalent to a single HMM, with  $N^K$  states (assuming there are  $N$  states for each chain), and therefore, learning in such a model is difficult: for a sequence of length  $T$ , a straightforward Viterbi algorithm has complexity  $O(N^{2K} T)$ . To find an exact solution, a junction tree algorithm could be used, but it results in an  $O(N^{K+1} K T)$  complexity. In practice, approximate techniques, such as variational approaches, could be used.<sup>[32]</sup>

All of the above models can be extended to allow for more distant dependencies among hidden states, e.g. allowing for a given state to be dependent on the previous two or three states rather than a single previous state; i.e. the transition probabilities are extended to encompass sets of three or four adjacent states (or in general  $K$  adjacent states). The disadvantage of such models is that dynamic-programming algorithms for training them have an  $O(N^K T)$  running time, for  $K$  adjacent states and  $T$  total observations (i.e. a length- $T$  Markov chain).

Another recent extension is the *triplet Markov model*,<sup>[33]</sup> in which an auxiliary underlying process is added to model some data specificities. Many variants of this model have been proposed. One should also mention the interesting link that has been established between the *theory of evidence* and the *triplet Markov models* <sup>[10]</sup> and which allows to fuse

data in Markovian context<sup>[11]</sup> and to model nonstationary data.<sup>[12][13]</sup> Note that alternative multi-stream data fusion strategies have also been proposed in the recent literature, e.g.<sup>[34]</sup>

Finally, a different rationale towards addressing the problem of modeling nonstationary data by means of hidden Markov models was suggested in.<sup>[35]</sup> It consists in employing a small recurrent neural network (RNN), specifically a reservoir network,<sup>[36]</sup> to capture the evolution of the temporal dynamics in the observed data. This information, encoded in the form of a high-dimensional vector, is used as a conditioning variable of the HMM state transition probabilities. Under such a setup, we eventually obtain a nonstationary HMM the transition probabilities of which evolve over time in a manner that is inferred from the data itself, as opposed to some unrealistic ad-hoc model of temporal evolution.

## See also


---

- Andrey Markov
- Baum–Welch algorithm
- Bayesian inference
- Bayesian programming
- Conditional random field
- Estimation theory
- HHpred / HHsearch free server and software for protein sequence searching
- HMMER, a free hidden Markov model program for protein sequence analysis
- Hidden Bernoulli model
- Hidden semi-Markov model
- Hierarchical hidden Markov model
- Layered hidden Markov model
- Sequential dynamical system
- Stochastic context-free grammar
- Time Series Analysis
- Variable-order Markov model
- Viterbi algorithm



## References

---

1. Baum, L. E.; Petrie, T. (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains" ([http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?handle=euclid.aoms/1177699147&view=body&content-type=pdf\\_1](http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?handle=euclid.aoms/1177699147&view=body&content-type=pdf_1)). *The Annals of Mathematical Statistics*. **37** (6): 1554–1563. doi:[10.1214/aoms/1177699147](https://doi.org/10.1214/aoms/1177699147) (<https://doi.org/10.1214%2Faoms%2F1177699147>). Retrieved 28 November 2011.
2. Baum, L. E.; Eagon, J. A. (1967). "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology" (<http://projecteuclid.org/euclid.bams/1183528841>). *Bulletin of the American Mathematical Society*. **73** (3): 360. doi:[10.1090/S0002-9904-1967-11751-8](https://doi.org/10.1090/S0002-9904-1967-11751-8) (<https://doi.org/10.1090%2FS0002-9904-1967-11751-8>). Zbl [0157.11101](https://zbmath.org/?format=complete&q=an:0157.11101) (<https://zbmath.org/?format=complete&q=an:0157.11101>).
3. Baum, L. E.; Sell, G. R. (1968). "Growth transformations for functions on manifolds" (<https://www.scribd.com/doc/6369908/Growth-Functions-for-Transformations-on-Manifolds>). *Pacific Journal of Mathematics*. **27** (2): 211–227. doi:[10.2140/pjm.1968.27.211](https://doi.org/10.2140/pjm.1968.27.211) (<https://doi.org/10.2140%2Fpjm.1968.27.211>). Retrieved 28 November 2011.
4. Baum, L. E.; Petrie, T.; Soules, G.; Weiss, N. (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains" (<http://projecteuclid.org/euclid.aoms/1177697196>). *The Annals of Mathematical Statistics*. **41**: 164. doi:[10.1214/aoms/1177697196](https://doi.org/10.1214/aoms/1177697196) (<https://doi.org/10.1214%2Faoms%2F1177697196>). JSTOR [2239727](https://www.jstor.org/stable/2239727) (<https://www.jstor.org/stable/2239727>). MR [0287613](https://www.ams.org/mathscinet-getitem?mr=0287613) (<https://www.ams.org/mathscinet-getitem?mr=0287613>). Zbl [0188.49603](https://zbmath.org/?format=complete&q=an:0188.49603) (<https://zbmath.org/?format=complete&q=an:0188.49603>).
5. Baum, L. E. (1973). "An Inequality and Associated Maximization Technique in Statistical Estimation of

- Baum, L.E. (1972). "An inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process". *Inequalities*. **3**: 1–8.
6. Stratonovich, R.L. (1960). "Conditional Markov Processes". *Theory of Probability and its Applications*. **5** (2): 156–178. doi:10.1137/1105015 (<https://doi.org/10.1137%2F1105015>).
7. Thad Starner, Alex Pentland. Real-Time American Sign Language Visual Recognition From Video Using Hidden Markov Models ([http://www.cc.gatech.edu/~thad/p/031\\_10\\_SL/real-time-asl-recognition-from%20video-using-hmm-ISCV95.pdf](http://www.cc.gatech.edu/~thad/p/031_10_SL/real-time-asl-recognition-from%20video-using-hmm-ISCV95.pdf)). Master's Thesis, MIT, Feb 1995, Program in Media Arts
8. B. Pardo and W. Birmingham. Modeling Form for On-line Following of Musical Performances (<http://www.cs.northwestern.edu/~pardo/publications/pardo-birmingham-aaai-05.pdf>). AAAI-05 Proc., July 2005.
9. Satish L, Gururaj BI (April 2003). "Use of hidden Markov models for partial discharge pattern classification ([http://ieeexplore.ieee.org/xpl/freeabs\\_all.jsp?arnumber=212242](http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=212242))". *IEEE Transactions on Dielectrics and Electrical Insulation*.
10. Pieczynski, Wojciech (2007). "Multisensor triplet Markov chains and theory of evidence" (<http://www.sciencedirect.com/science/article/pii/S0888613X06000375>). *International Journal of Approximate Reasoning*. **45**: 1–16. doi:10.1016/j.ijar.2006.05.001 (<https://doi.org/10.1016%2Fj.ijar.2006.05.001>).
11. Boudaren et al. (<http://asp.eurasipjournals.com/content/pdf/1687-6180-2012-134.pdf>), M. Y. Boudaren, E. Monfrini, W. Pieczynski, and A. Aissani, Dempster-Shafer fusion of multisensor signals in nonstationary Markovian context, EURASIP Journal on Advances in Signal Processing, No. 134, 2012.
12. Lanchantin et al. ([http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1468502&contentType=Journals+%26+Magazines&searchField%3DSearch\\_All%26queryText%3DLanchantin+pieczynski](http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=1468502&contentType=Journals+%26+Magazines&searchField%3DSearch_All%26queryText%3DLanchantin+pieczynski)), P. Lanchantin and W. Pieczynski, Unsupervised restoration of hidden non stationary Markov chain using evidential priors, IEEE Trans. on Signal Processing, Vol. 53, No. 8, pp. 3091-3098, 2005.
13. Boudaren et al. ([http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6244854&contentType=Journals+%26+Magazines&searchField%3DSearch\\_All%26queryText%3Dboudaren](http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6244854&contentType=Journals+%26+Magazines&searchField%3DSearch_All%26queryText%3Dboudaren)), M. Y. Boudaren, E. Monfrini, and W. Pieczynski, Unsupervised segmentation of random discrete data hidden with switching noise distributions, IEEE Signal Processing Letters, Vol. 19, No. 10, pp. 619-622, October 2012.
14. Lawrence R. Rabiner (February 1989). "A tutorial on Hidden Markov Models and selected applications in speech recognition" (<http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorial%20on%20hmm%20and%20applications.pdf>) (PDF). *Proceedings of the IEEE*. **77** (2): 257–286. doi:10.1109/5.18626 (<https://doi.org/10.1109%2F5.18626>). [1] (<http://www.cs.cornell.edu/courses/cs481/2004fa/rabiner.pdf>)
15. Newberg, L. (2009). "Error statistics of hidden Markov model and hidden Boltzmann model results" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2722652>). *BMC Bioinformatics*. **10**: 212. doi:10.1186/1471-2105-10-212 (<https://doi.org/10.1186%2F1471-2105-10-212>). PMC 2722652 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2722652>)  PMID 19589158 (<https://www.ncbi.nlm.nih.gov/pubmed/19589158>). 
16. Sipos, I. Róbert. *Parallel stratified MCMC sampling of AR-HMMs for stochastic time series prediction*. In: Proceedings, 4th Stochastic Modeling Techniques and Data Analysis International Conference with Demographics Workshop (SMTDA2016), pp. 295-306. Valletta, 2016. PDF ([http://1drv.ms/b/s!ApL\\_0Av0YGDGlgIwEOv1aYAGbmQeL](http://1drv.ms/b/s!ApL_0Av0YGDGlgIwEOv1aYAGbmQeL))
17. Chatzis, Sotirios P.; Kosmopoulos, Dimitrios I. (2011). "A variational Bayesian methodology for hidden Markov models utilizing Student's-t mixtures" (<http://www.sciencedirect.com/science/article/pii/S0031320310004383>). *Pattern Recognition*. **44**: 295–306. doi:10.1016/j.patcog.2010.09.001 (<https://doi.org/10.1016%2Fj.patcog.2010.09.001>).
18. R. Paroli. et al., *Poisson hidden Markov models for time series of overdispersed insurance counts* ([http://actuarie.s.org/ASTIN/Colloquia/Porto\\_Cervo/Paroli\\_Redaeli\\_Spezia.pdf](http://actuarie.s.org/ASTIN/Colloquia/Porto_Cervo/Paroli_Redaeli_Spezia.pdf))
19. Sipos, I. Róbert; Ceffer, Attila; Levendovszky, János (2016). "Parallel Optimization of Sparse Portfolios with AR-HMMs". *Computational Economics*. **49**: 563–578. doi:10.1007/s10614-016-9579-y (<https://doi.org/10.1007%2Fs10614-016-9579-y>).
20. Petropoulos, Anastasios; Chatzis, Sotirios P.; Xanthopoulos, Stylianos (2016). "A novel corporate credit rating system based on Student's-t hidden Markov models" (<http://www.sciencedirect.com/science/article/pii/S0957417416000257>). *Expert Systems with Applications*. **53**: 87–105. doi:10.1016/j.eswa.2016.01.015 (<https://doi.org/10.1016%2Fj.eswa.2016.01.015>).



21. NICOLAI, CHRISTOPHER (2013). "SOLVING ION CHANNEL KINETICS WITH THE QuB SOFTWARE". *Biophysical Reviews and Letters*. **8** (3n04): 191–211. doi:10.1142/S1793048013300053 (<https://doi.org/10.1142%2FS1793048013300053>).
22. Stigler, J.; Ziegler, F.; Gieseke, A.; Gebhardt, J. C. M.; Rief, M. (2011). "The Complex Folding Network of Single Calmodulin Molecules". *Science*. **334** (6055): 512–516. doi:10.1126/science.1207598 (<https://doi.org/10.1126%2Fscience.1207598>). PMID 22034433 (<https://www.ncbi.nlm.nih.gov/pubmed/22034433>).
23. Wong, W.; Stamp, M. (2006). "Hunting for metamorphic engines". *Journal in Computer Virology*. **2** (3): 211–229. doi:10.1007/s11416-006-0028-7 (<https://doi.org/10.1007%2Fs11416-006-0028-7>).
24. Wong, K. -C.; Chan, T. -M.; Peng, C.; Li, Y.; Zhang, Z. (2013). "DNA motif elucidation using belief propagation" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3763557>). *Nucleic Acids Research*. **41** (16): e153. doi:10.1093/nar/gkt574 (<https://doi.org/10.1093%2Fnar%2Fgkt574>). PMC 3763557 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3763557>). PMID 23814189 (<https://www.ncbi.nlm.nih.gov/pubmed/23814189>).
25. Baker, J. (1975). "The DRAGON system—An overview". *IEEE Transactions on Acoustics, Speech, and Signal Processing*. **23**: 24–29. doi:10.1109/TASSP.1975.1162650 (<https://doi.org/10.1109%2FTASSP.1975.1162650>).
26. Jelinek, F.; Bahl, L.; Mercer, R. (1975). "Design of a linguistic statistical decoder for the recognition of continuous speech". *IEEE Transactions on Information Theory*. **21** (3): 250. doi:10.1109/TIT.1975.1055384 (<https://doi.org/10.1109%2FTIT.1975.1055384>).
27. Xuedong Huang; M. Jack; Y. Ariki (1990). *Hidden Markov Models for Speech Recognition*. Edinburgh University Press. ISBN 0-7486-0162-7.
28. Xuedong Huang; Alex Acero; Hsiao-Wuen Hon (2001). *Spoken Language Processing*. Prentice Hall. ISBN 0-13-022616-5.
29. M. Bishop and E. Thompson (1986). "Maximum Likelihood Alignment of DNA Sequences". *Journal of Molecular Biology*. **190** (2): 159–165. doi:10.1016/0022-2836(86)90289-5 (<https://doi.org/10.1016%2F0022-2836%2886%2990289-5>). PMID 3641921 (<https://www.ncbi.nlm.nih.gov/pubmed/3641921>). (subscription required) 
30. Durbin, Richard M.; Eddy, Sean R.; Krogh, Anders; Mitchison, Graeme (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (<http://www.cambridge.org/gb/knowledge/isbn/item1158701>) (1st ed.), Cambridge, New York: Cambridge University Press, doi:10.2277/0521629713 (<https://doi.org/10.2277%2F0521629713>), ISBN 0-521-62971-3, OCLC 593254083 (<https://www.worldcat.org/oclc/593254083>)
31. Sotirios P. Chatzis, "Hidden Markov Models with Nonelliptically Contoured State Densities," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 12, pp. 2297-2304, Dec. 2010. [2] (<http://ieeexplore.ieee.org/abstract/document/5551154/>)
32. Ghahramani, Zoubin; Jordan, Michael I. (1997). "Factorial Hidden Markov Models". *Machine Learning*. **29** (2/3): 245–273. doi:10.1023/A:1007425814087 (<https://doi.org/10.1023%2FA%3A1007425814087>).
33. Pieczynski, Wojciech (2002). "Chaînes de Markov Triplet" (<http://www.sciencedirect.com/science/article/pii/S1631073X02024627>). *Comptes Rendus Mathématique*. **335**: 275–278. doi:10.1016/S1631-073X(02)02462-7 (<https://doi.org/10.1016%2FS1631-073X%2802%2902462-7>).
34. Sotirios P. Chatzis, Dimitrios Kosmopoulos, "Visual Workflow Recognition Using a Variational Bayesian Treatment of Multistream Fused Hidden Markov Models," IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 7, pp. 1076-1086, July 2012. [3] (<http://ieeexplore.ieee.org/document/6164251/>)
35. Chatzis, Sotirios P.; Demiris, Yiannis (2012). "A Reservoir-Driven Non-Stationary Hidden Markov Model" (<http://www.sciencedirect.com/science/article/pii/S0031320312001987>). *Pattern Recognition*. **45** (11): 3985–3996. doi:10.1016/j.patcog.2012.04.018 (<https://doi.org/10.1016%2Fj.patcog.2012.04.018>).
36. M. Lukosevicius, H. Jaeger (2009) Reservoir computing approaches to recurrent neural network training, *Computer Science Review* **3**: 127–149.

## External links

 Media related to Hidden Markov Model at Wikimedia Commons

# Concepts

- Teif, V. B.; Rippe, K. (2010). "Statistical–mechanical lattice models for protein–DNA binding in chromatin" (<http://arxiv.org/pdf/1004.5514>). *J. Phys.: Condens. Matter.* **22**: 414105. doi:10.1088/0953-8984/22/41/414105 (<https://doi.org/10.1088%2F0953-8984%2F22%2F41%2F414105>).
- A Revealing Introduction to Hidden Markov Models (<http://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf>) by Mark Stamp, San Jose State University.
- Fitting HMM's with expectation-maximization – complete derivation (<https://web.archive.org/web/20120415032315/http://www.ee.washington.edu/research/guptalab/publications/EMbookChenGupta2010.pdf>)
- A step-by-step tutorial on HMMs ([http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html\\_dev/main.html](http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html)) (*University of Leeds*)
- Hidden Markov Models (<http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/HiddenMarkovModels.html>) (*an exposition using basic mathematics*)
- Hidden Markov Models (<http://jedlik.phy.bme.hu/~gerjanos/HMM/node2.html>) (*by Narada Warakagoda*)
- Hidden Markov Models: Fundamentals and Applications Part 1 (<http://www.eecis.udel.edu/~lliao/cis841s06/hmmtutorialpart1.pdf>), Part 2 (<http://www.eecis.udel.edu/~lliao/cis841s06/hmmtutorialpart2.pdf>) (*by V. Petrushin*)
- Lecture on a Spreadsheet by Jason Eisner, Video ([http://videlectures.net/hltss2010\\_eisner\\_plm/video/2/](http://videlectures.net/hltss2010_eisner_plm/video/2/)) and interactive spreadsheet (<http://www.cs.jhu.edu/~jason/papers/eisner.hmm.xls>)

# Software

- Hidden Markov Model (HMM) Toolbox for Matlab (<http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html>) (*by Kevin Murphy*)
- Hidden Markov Model Toolkit (HTK) (<http://htk.eng.cam.ac.uk/>) (*a portable toolkit for building and manipulating hidden Markov models*)
- Hidden Markov Model R-Package (<https://cran.r-project.org/web/packages/HMM/index.html>) to set up, apply and make inference with discrete time and discrete space Hidden Markov Models
- zipHMMlib (<http://birc.au.dk/software/zipHMM>) (*a library for general (discrete) hidden Markov models, exploiting repetitions in the input sequence to greatly speed up the forward algorithm. Implementation of the posterior decoding algorithm and the Viterbi algorithm are also provided.*)
- GHMM Library (<http://www.ghmm.org/>) (*home page of the GHMM Library project*)
- Jahmm Java Library (<https://github.com/KommuSoft/jahmm>) (*general-purpose Java library*)
- HMM and other statistical programs (<http://www.kanungo.com/software/software.html>) (*Implementation in C by Tapas Kanungo*)
- The hmm package (<http://hackage.haskell.org/cgi-bin/hackage-scripts/package/hmm>) A Haskell (<http://www.haskell.org/>) library for working with Hidden Markov Models.
- GT2K (<http://gt2k.cc.gatech.edu/>) Georgia Tech Gesture Toolkit (referred to as GT2K)
- Hidden Markov Models -online calculator for HMM – Viterbi path and probabilities. Examples with perl source code. ([http://www.lwebzem.com/cgi-bin/courses/hidden\\_markov\\_model\\_online.cgi](http://www.lwebzem.com/cgi-bin/courses/hidden_markov_model_online.cgi))
- A discrete Hidden Markov Model class, based on OpenCV. (<http://sourceforge.net/projects/cvhmm/>)
- depmixS4 (<https://cran.r-project.org/web/packages/depmixS4/index.html>) R-Package (Hidden Markov Models of GLMs and Other Distributions in S4 )
- MLPACK contains a C++ implementation of HMMs
- Hidden Markov Models Java Library (<https://adrianulbona.github.io/hmm>) contains basic HMMs abstractions in Java 8
- SFIHMM (<http://tuvalu.santafe.edu/~simon/styled-8/>) high-speed C code for the estimation of Hidden Markov Models, Viterbi Path Reconstruction, and the generation of simulated data from HMMs.

---

Retrieved from "[https://en.wikipedia.org/w/index.php?title=Hidden\\_Markov\\_model&oldid=813449939](https://en.wikipedia.org/w/index.php?title=Hidden_Markov_model&oldid=813449939)"

---

This page was last edited on 3 December 2017, at 19:30.

Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. By using this site, you agree to the [Terms of Use](#) and [Privacy Policy](#). Wikipedia® is a registered trademark of the [Wikimedia Foundation, Inc.](#), a non-profit organization.

