Chapter 1 - Mixtures & Markov Chains

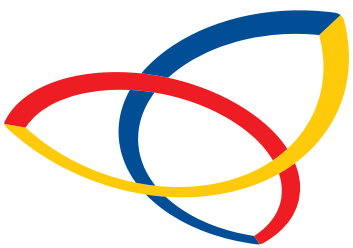
02433 - Hidden Markov Models

Tryggvi Jónsson, Martin Wæver Pedersen, Henrik Madsen

Course week 1

JKMO, compiled February 9, 2015

LUNDS UNIVERSITET
Lunds Tekniska Högskola



Interreg IV A
ÖRESUND – KATTEGAT – SKAGERRAK

DTU

Hidden Markov Models (HMMs)

Informal Definition: Models in which the distribution generating observations depends on an unobserved Markov process.

Common applications:

- Speech recognition (Rabiner, 1989).
- Bioinformatics (Krogh and Brown, 1994).
- Environmental processes (Lu and Berliner, 1999).
- Econometrics (Rydén et al., 1998).
- Image processing and computer vision (Li et al., 2002).
- Animal behaviour (Patterson et al., 2009; Zucchini et al., 2008).
- Wind power forecasting (Pinson and Madsen, accepted).

Available HMM books

General textbooks:

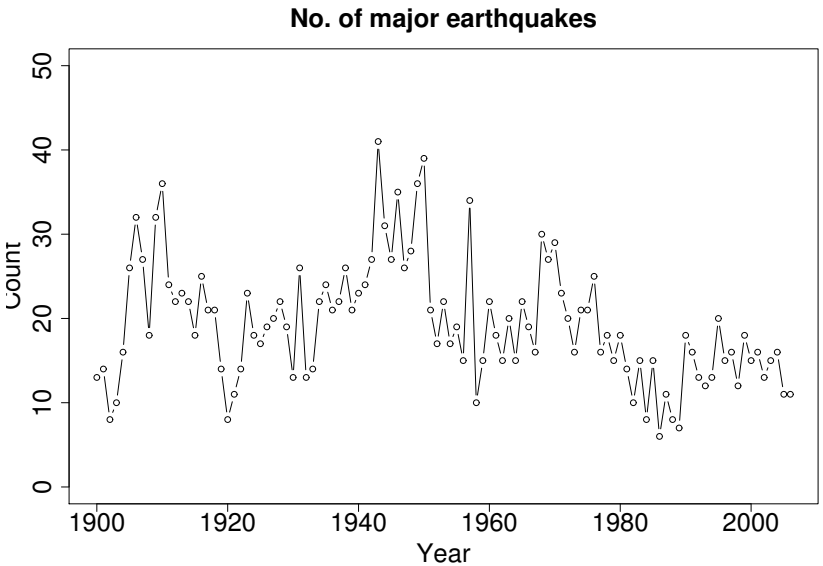
- Zucchini, W. and I. MacDonald. 2009. Hidden Markov Models for Time Series. Chapman & Hall/CRC, London.
- MacDonald, I. and W. Zucchini. 1997. Hidden Markov and other models for discrete-valued time series. CRC Press.
- Cappe, O., E. Moulines, and T. Ryden. 2005. Inference in hidden Markov models. Springer Verlag.

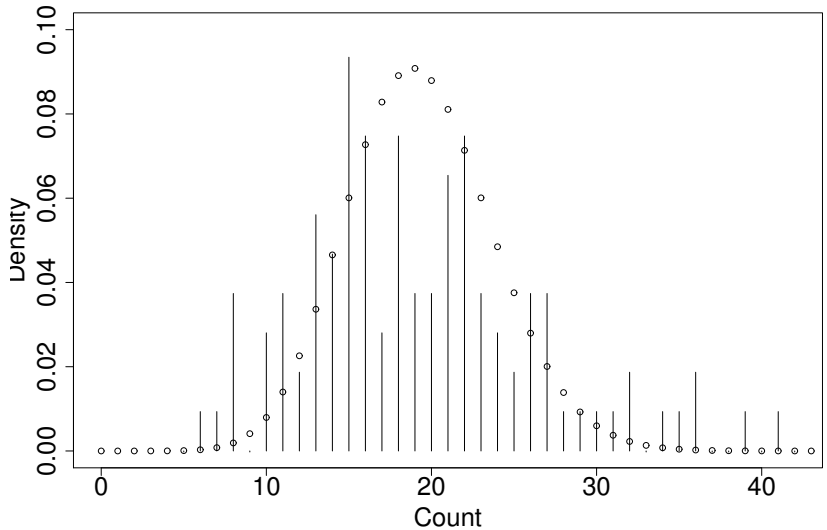
Specialised textbooks:

- Elliott, R., L. Aggoun, and J. Moore. 1995. Hidden Markov models: estimation and control. Springer.
- Li, J. and R. Gray. 2000. Image segmentation and compression using hidden Markov models. Kluwer Academic Publishers.
- Koski, T. 2001. Hidden Markov models for bioinformatics. Springer Netherlands.
- Durbin, R. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge Univ Pr.
- Bunke, H. and T. Caelli, 2001. Hidden Markov Models Applications in Computer Vision, Series in Machine Perception and Artificial Intelligence, vol. 45.
- Bhar, R. and S. Hamori. 2004. Hidden Markov models: applications to financial economics. Kluwer Academic Pub.

Example: US major earthquake count

Possible assumption: Earthquake counts (X) within a year are Poisson distributed.





From data we observe:

- Overdispersion: $\mathrm{E}(X) = 19.36 \neq \mathrm{Var}(X) = 51.57$, i.e. the property of the Poisson distribution that the mean equals the variance is not fulfilled.
- Serial dependence (nonzero autocorrelation): $\rho(X_t, X_{t-1}) = 0.57$.

HMMs can account for these features (and more).

Independent Mixture Models

An independent mixture model consists of $m < \infty$ component distributions with probability functions p_i for $i \in \{1, \ldots, m\}$ and a "mixing distribution". The mixing is performed by a discrete random variable C:

$$C = \begin{cases} 1 & \text{with probability } \delta_1 \\ \vdots & \vdots \\ i & \text{with probability } \delta_i \\ \vdots & \vdots \\ m & \text{with probability } \delta_m = 1 - \sum_{i=1}^{m-1} \delta_i \end{cases},$$

Thus $\Pr(C = i) = \delta_i$ must obey $0 < \delta_i < 1$ and that $\sum_{i=1}^{m} \delta_i = 1$.

Moments of Mixture Models

For a discrete random variable X described by a mixture model consisting of m components it holds that:

$$p(X) = \sum_{i=1}^{m} \delta_i p_i(X) \implies \Pr(X = x) = \sum_{i=1}^{m} \Pr(X = x|C = i)\Pr(C = i).$$

Hence, letting Y_i denote the random variable with probability function p_i

$$\mathrm{E}(X) = \sum_{i=1}^{m} \Pr(C = i)\mathrm{E}(X|C = i) = \sum_{i=1}^{m} \delta_i \mathrm{E}(Y_i)$$

and

$$\mathrm{Var}(X) = \sum_{i=1}^{m} \delta_i [\mathrm{Var}(Y_i) + \sum_{j=i+1}^{m} \delta_j (\mathrm{E}(Y_i) - \mathrm{E}(Y_j))^2].$$

Parameter estimation

ML estimation of mixture distribution is done my maximizing the combined likelihood of the components:

$$L(\theta_1, \ldots, \theta_m, \delta_1, \ldots, \delta_m | x_1, \ldots, x_n) = \prod_{j=1}^{n} \sum_{i=1}^{m} \delta_i p_i(x_j, \theta_i),$$

where $\theta_1, \ldots, \theta_m$ are the parameter vectors of the component distributions, and $x_1, \ldots, x_n$ are the n observations from the system. Difficult to find the maximum analytically. Instead maximize numerically, e.g. using the `flexmix` package in R.

Parameter estimation of a Poisson mixture

Consider a mixture of m Poisson components. Independent parameters are $\lambda_1, \ldots, \lambda_m$ and $\delta_1, \ldots, \delta_{m-1}$ leading to the likelihood function:

$$
\begin{aligned}
&L(\lambda_1, \ldots, \lambda_m, \delta_1, \ldots, \delta_{m-1} | x_1, \ldots, x_n) \\
&= \prod_{j=1}^{n} \left[\sum_{i=1}^{m-1} \left(\delta_i \frac{\lambda_i^{x_j} e^{-\lambda_i}}{x_j!} \right) + (1 - \sum_{i=1}^{m-1} \delta_i) \frac{\lambda_m^{x_j} e^{-\lambda_m}}{x_j!} \right].
\end{aligned}
$$

Since the parameters have to fulfill $\sum_i \delta_i = 1$, $\delta_i > 0$ and $\lambda_i > 0$ for all i we reparameterize (or transform) by

$$
\begin{aligned}
\text{Log-transformation} \quad & \eta_i = \log \lambda_i, i = 1, \ldots, m \\
\text{Logit-transformation} \quad & \tau_i = \log \left(\frac{\delta_i}{1 - \sum_{j=2}^{m} \delta_j} \right) i = 2, \ldots, m
\end{aligned}
$$

Then the likelihood can be maximized using the transformed parameters.

Parameter estimation of a Poisson mixture

Once the likelihood is maximized the original parameters can be recovered by a back transformation:

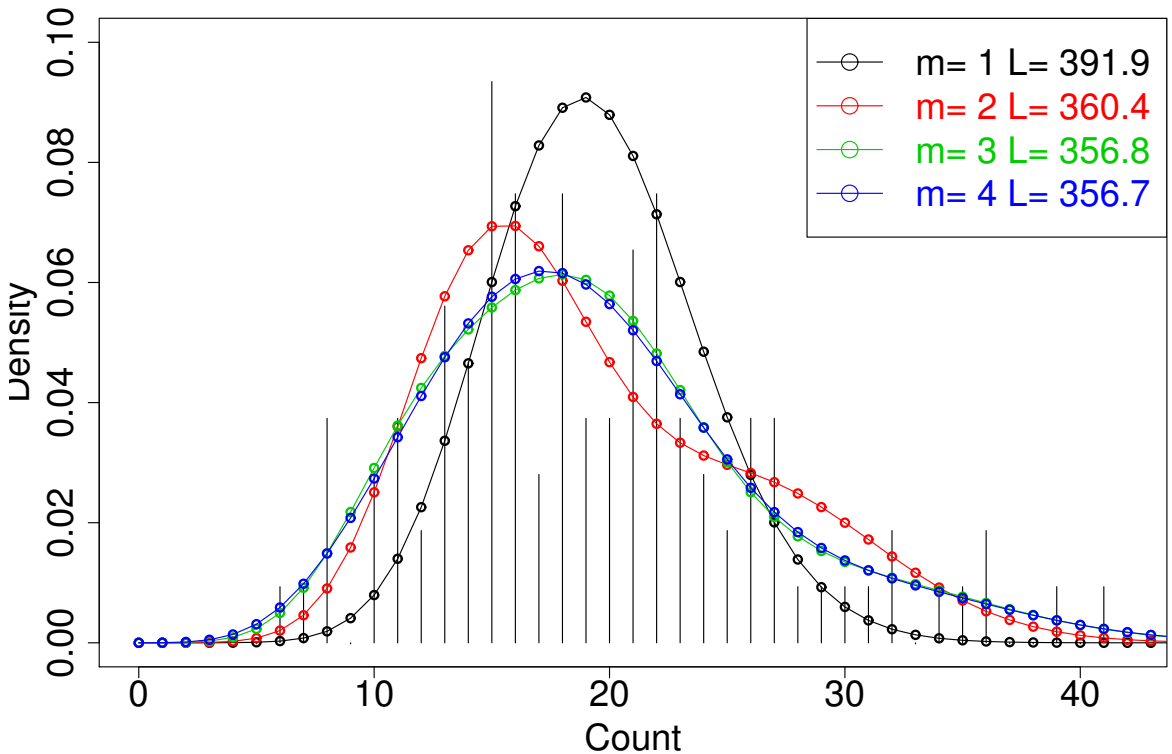
$$\text{exp-transformation} \quad \lambda_i = e^{\eta_i}, i = 1, \ldots, m$$

$$\text{inverse logit} \quad \delta_i = \frac{e^{\tau_i}}{1 + \sum_{j=2}^{m} e^{\tau_j}}, i = 2, \ldots, m$$

and finally $\delta_1 = 1 - \sum_{j=2}^{m} \delta_j$.

Example: US major earthquake count

Recall the earthquake data from slide 4. Four Poisson mixture models are fitted to the data for $m = 1, \ldots, 4$. The resulting mixture distributions (with maximum likelihood values) are shown below. It is evident from the graphs (and from the likelihoods) that the difference between the $m = 3$ and $m = 4$ models is minimal.



Estimation of mixtures of continuous distributions

Issue with unbounded likelihood

In estimation of mixtures of continuous distributions the likelihood is calculated from probability density functions instead of probability distributions. Thus, the likelihood can become unbounded in the vicinity of certain parameter combinations. For example in mixtures of normal distributions, if a variance parameter in the maximization process shrinks toward zero the density value at the mean goes to infinity thus ruining the estimation. This can be prevented by discretisizing the density (or equivalently the likelihood) and integrating over the small intervals $[a_j, b_j]$:

$$L = \prod_{j=1}^{n} \sum_{i=1}^{m} \delta_i \int_{a_j}^{b_j} p_i(x, \theta_i) dx$$

where a_j and b_j are the upper and lower bounds respectively of the jth interval out of n intervals in total.

Markov Chains

Definition: A sequence of discrete random variables $\{C_t : t \in \mathbb{N}\}$ is said to be a (discrete time) Markov chain (MC) if for all $t \in \mathbb{N}$ it satisfies the Markov property: $\Pr(C_{t+1}|C_t, \ldots, C_1) = \Pr(C_{t+1}|C_t)$, i.e. that the future of the chain is independent of the past conditional on the present.

Important quantities and aspects related to MCs:

- Transition probabilities: $\gamma_{ij}(t) = \Pr(C_{s+t} = j|C_s = i)$
- Homogeneity: $\gamma_{ij}(t) = \Pr(C_{s+t} = j|C_s = i)$ is independent of s
- Transition probability matrix: $\mathbf{\Gamma}(t) = \begin{pmatrix} \gamma_{11}(t) & \cdots & \gamma_{1m}(t) \\ \vdots & \ddots & \vdots \\ \gamma_{m1}(t) & \cdots & \gamma_{mm}(t) \end{pmatrix}$
- Chapman-Kolmogorov equations: $\mathbf{\Gamma}(t+u) = \mathbf{\Gamma}(t)\mathbf{\Gamma}(u)$
- Short-hand for the one-step transition probability matrix: $\mathbf{\Gamma} = \mathbf{\Gamma}(1)$.

Markov Chains

More definitions related to MCs:

- The distribution of C_t at index t (where t typically is time) is contained in the row vector: $\mathbf{u}(t) = (\Pr(C_t = 1), \ldots, \Pr(C_t = m))$.
- The evolution in time of $\mathbf{u}(t)$ is described by $\mathbf{\Gamma}(t)$ in that $\mathbf{u}(t+s) = \mathbf{u}(t)\mathbf{\Gamma}(s)$.
- The stationary distribution of a Markov chain is $\boldsymbol{\delta}$ if $\boldsymbol{\delta}\mathbf{\Gamma}(s) = \boldsymbol{\delta}$ for all $s \leq 0$, and $\boldsymbol{\delta}\mathbf{1}' = 1$. The stationary distribution can be found either by solving $\boldsymbol{\delta}(\mathbf{I}_m - \mathbf{\Gamma} + \mathbf{U}) = \mathbf{1}$, where $\mathbf{U}$ is a $m \times m$ matrix of ones, or by substituting one of the equations in $\boldsymbol{\delta}\mathbf{\Gamma} = \boldsymbol{\delta}$ with $\sum_i \delta_i = 1$.
- Reversibility: A random process is said to be reversible if its finite-dimensional distributions are invariant under reversal of time. A stationary irreducible Markov chain satisfying the "detailed balance" equations, $\delta_i \gamma_{ij} = \delta_j \gamma_{ji}$ is reversible. $\forall i, j$

Autocorrelation function of an MC

Define the vector $\mathbf{v} = (1, 2, \ldots, m)$ and the matrix $\mathbf{V} = \text{diag}(1, 2, \ldots, m)$. Then for all integers $k > 0$

$$\text{Cov}(C_t, C_{t+k}) = \boldsymbol{\delta}\mathbf{V}\boldsymbol{\Gamma}^k\mathbf{v}' - (\boldsymbol{\delta}\mathbf{v}')^2.$$

Now, $\boldsymbol{\Gamma} = \mathbf{U}\boldsymbol{\Omega}\mathbf{U}^{-1}$, where $\boldsymbol{\Omega} = \text{diag}(1, \omega_2, \omega_3, \ldots, \omega_m)$ and the columns of $\mathbf{U}$ are corresponding right eigenvectors of $\boldsymbol{\Gamma}$. Then

$$\text{Cov}(C_t, C_{t+k}) = \underbrace{\boldsymbol{\delta}\mathbf{V}\mathbf{U}}_{\mathbf{a}}\boldsymbol{\Omega}^k\underbrace{\mathbf{U}^{-1}\mathbf{v}'}_{\mathbf{b}} - (\boldsymbol{\delta}\mathbf{v}')^2 = \mathbf{a}\boldsymbol{\Omega}^k\mathbf{b}' - a_1 b_1 = \sum_{i=2}^{m} a_i b_i \omega_i^k.$$

This implies that $\text{Var}(C_t) = \sum_{i=2}^{m} a_i b_i$, and therefore that the autocorrelation function is:

$$\rho(k) = \text{Corr}(C_t, C_{t+k}) = \frac{\sum_{i=2}^{m} a_i b_i \omega_i^k}{\sum_{i=2}^{m} a_i b_i}.$$

Note that the $\rho(k)$ is calculated only based on $\boldsymbol{\Gamma}$ and known quantities.

Estimating transition probabilities

A realization of a Markov chain with three state could read:

2332111112 3132332122 3232332222 3132332212 3232132232
3132332223 3232331232 3232331222 3232132123 3132332121

A matrix $\mathbf{F}$ of transition counts is

$$\mathbf{F} = \begin{pmatrix} 4 & 7 & 6 \\ 8 & 10 & 24 \\ 6 & 24 & 10 \end{pmatrix}$$

where f_{ij} denotes the number of transitions from i to j.

Estimating transition probabilities

An estimate of $\mathbf{\Gamma}$ intuitively found to be

$$\mathbf{\Gamma} = \begin{pmatrix} 4/17 & 7/17 & 6/17 \\ 8/42 & 10/42 & 24/42 \\ 6/40 & 24/40 & 10/40 \end{pmatrix}$$

by letting

$$\gamma_{ii} = \frac{f_{ii}}{\sum_{j=1}^{m} f_{ij}} \text{ and } \gamma_{ij} = \frac{f_{ij}\gamma_{ii}}{f_{ii}} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}}$$

It can be shown that this is equivalent to the maximum likelihood estimate of the transition probability matrix (see p. 21 in Zucchini09 for the proof).

Probability rules important for this course (1/2)

Joint probability

$$\Pr(A, B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A).$$

Bayes' rule

$$\Pr(A|B) = \Pr(B|A)\frac{\Pr(A)}{\Pr(B)}.$$

For disjoint events $B_1, \ldots, B_m$ then

$$\Pr(A) = \sum_{i=1}^{m} \Pr(A, B_i) = \sum_{i=1}^{m} \Pr(A|B_i)\Pr(B_i). \quad \text{(marginalization)}$$

If A and C are conditional independent given B then

$$\Pr(A|B) = \Pr(A|B, C).$$

Probability rules important for this course (2/2)

Define $A \in \{1, \ldots, i, \ldots, m\}$ and $B \in \{1, \ldots, j, \ldots, n\}$.
Expectations

$$\mathrm{E}(A) = \sum_{i=1}^{m} i \Pr(A = i)$$

$$\mathrm{E}(A^k) = \sum_{i=1}^{m} i^k \Pr(A = i),$$

$$\mathrm{E}(AB) = \sum_{i=1}^{m} \sum_{j=1}^{n} ij \Pr(A = i, B = j) = \sum_{i=1}^{m} \sum_{j=1}^{n} ij \Pr(A = i | B = j) \Pr(B = j).$$

Variance

$$\mathrm{Var}(A) = \mathrm{E}(A^2) - [\mathrm{E}(A)^2].$$

Covariance

$$\mathrm{Cov}(A, B) = \mathrm{E}(AB) - \mathrm{E}(A)\mathrm{E}(B)$$

Exercises

3,6,8,10,(12,15)

References

Krogh, A. and I. Brown. 1994. Hidden Markov models in computational biology. J. Mol. Bioi **235**:1501–1531.

Li, J., A. Najmi, and R. Gray. 2002. Image classification by a two-dimensional hidden Markov model. Signal Processing, IEEE Transactions on **48**:517–533.

Lu, Z. and L. Berliner. 1999. Markov switching time series models with application to a daily runoff series. Water Resources Research **35**:523–534.

Patterson, T., M. Basson, M. Bravington, and J. Gunn. 2009. Classifying movement behaviour in relation to environmental conditions using hidden Markov models. Journal of Animal Ecology **78**:1113–1123.

Pinson, P. and H. Madsen. accepted. Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models. Journal of Forecasting .

Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications inspeech recognition. Proceedings of the IEEE **77**:257–286.

Rydén, T., T. Terasvirta, and S. Åsbrink. 1998. Stylized facts of daily return series and the hidden Markov model. Journal of Applied Econometrics **13**:217–244.

Zucchini, W., D. Raubenheimer, and I. MacDonald. 2008. Modeling Time Series of Animal Behavior by Means of a Latent-State Model with Feedback. Biometrics **64**:807–815.