Hidden Markov models
Previous work
Main results
Elements of proof

Consistency of the maximum likelihood estimator for general hidden Markov models

Jimmy Olsson
Centre for Mathematical Sciences
Lund University

Nordstat 2012
Umeå, Sweden
This is a joint work with

- **Randal Douc**, Télécom SudParis,
- **Eric Moulines**, Télécom ParisTech,
- **Ramon van Handel**, Princeton University.

Plan of the talk

1. Hidden Markov models
2. Previous work
3. Main results
4. Elements of proof
1. Hidden Markov models

2. Previous work

3. Main results

4. Elements of proof
Hidden Markov models (HMMs)

An HMM comprises

- an unobservable Markov chain $X \triangleq (X_n)_{n \geq 0}$ on $(\mathcal{X}, \mathcal{X})$ with transition kernel $Q_\theta$ and initial distribution $\nu$, i.e. $X_0 \sim \nu$ and

$$X_{n+1} | X_n \sim Q_\theta(X_n, \cdot).$$

- observations $(Y_n)_{n \geq 0}$ in $(\mathcal{Y}, \mathcal{Y})$ being conditionally independent given $X$ such that

$$Y_n | X \overset{d}{=} Y_n | X_n \sim G_\theta(X_n, \cdot).$$

We assume that $G_\theta$ has a transition density $g_\theta$, i.e.

$$G_\theta(x, A) = \int_A g_\theta(x, y) \, \mu(dy), \quad A \in \mathcal{Y}.$$

Here $\theta$ is a parameter vector belonging to some compact metric space $\Theta$. 
HMMs (cont.)

Graphical representation:

\[ Y_{n-1} \quad Y_n \quad Y_{n+1} \]

\[ X_{n-1} \quad X_n \quad X_{n+1} \]

(Observations)

(Markov chain)

\[ Y_n | X_n \sim G_\theta(X_n, \cdot) \]
\[ X_{n+1} | X_n \sim Q_\theta(X_n, \cdot) \]
\[ X_0 \sim \nu \]
Throughout this talk we fix a distinguished element $\theta^* \in \Theta$, interpreted as the true parameter value. Having done this, we

- presume that $Q_{\theta^*}$ possesses a unique invariant distribution $\pi_{\theta^*}$,
- denote by $\mathbb{P}_{\theta^*}$ the law of the stationary HMM under $\theta^*$,
- and assume that we are given observations $Y_0^n = (Y_0, \ldots, Y_n)$ sampled from the distribution $\mathbb{P}_{\theta^*}$.

In this setting, we estimate $\theta^*$ using the maximum likelihood estimator (MLE), i.e. the argument $\hat{\theta}_n$ maximizing

$$\Theta \ni \theta \mapsto p_{\theta}^\nu(Y_0^n),$$

where $p_{\theta}^\nu(y_0^n)$ denotes the density of the measure $\mathbb{P}_{\theta}^\nu(Y_0^n \in \cdot)$.
We want to show that the MLE is *strongly consistent* in the sense that $\hat{\theta}_n \to \theta^*$, $\mathbb{P}_{\theta^*}$-a.s., as $n \to \infty$.

For HMMs, this problem has been addressed in a long series of papers, where the most significant quantum leaps are:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>Contributors</th>
</tr>
</thead>
<tbody>
<tr>
<td>finite</td>
<td>finite</td>
<td>Baum &amp; Petrie (1966)</td>
</tr>
<tr>
<td>finite</td>
<td>general</td>
<td>Leroux (1992)</td>
</tr>
<tr>
<td>compact</td>
<td>general</td>
<td>Douc &amp; Mathias (2001)</td>
</tr>
</tbody>
</table>
Hidden Markov models

Previous work

Main results

Elements of proof

We are here ➔ •

1. Hidden Markov models

2. Previous work

3. Main results

4. Elements of proof

J. Olsson

Consistency of the MLE for general HMMs
Common thread of previous contributions:

(1) Show that for all $\theta \in \Theta$ there is a constant $H(\theta, \theta^*)$, the asymptotic contrast, such that, $\bar{\mathbb{P}}_{\theta^*}$-a.s.,

$$
\lim_{n \to \infty} n^{-1} \log p^\nu_\theta(Y_0^n) = \lim_{n \to \infty} n^{-1} \bar{E}_{\theta^*} \left( \log p^\nu_\theta(Y_0^n) \right) = H(\theta, \theta^*).
$$

(2) Establish identifiability, i.e. that the relative entropy

$$
\mathcal{E}(\theta, \theta^*) \triangleq H(\theta^*, \theta^*) - H(\theta, \theta^*)
$$

is minimized only at $\theta = \theta^*$.

(3) Prove that $\hat{\theta}_n$ converges (a.s.) to the maximizer of $H(\theta, \theta^*)$. 
Previous work

In the literature, establishing the existence of $H(\theta, \theta^*)$ has (for $\theta \neq \theta^*$) required very restrictive assumptions that are rarely satisfied in practice (particularly in the non-compact case), such as the following.

**Assumptions (Global Doeblin)**

*There are constants $0 < \epsilon_- < \epsilon_+$ and a probability measure $\eta$ on $(\mathcal{X}, \mathcal{X})$ such that for all $x \in \mathcal{X}$ and $A \in \mathcal{X}$,*

$$\epsilon_- \eta(A) \leq Q(x, A) \leq \epsilon_+ \eta(A).$$

Thus, a general consistency result has hitherto remained lacking.
1 Hidden Markov models

2 Previous work

3 Main results

4 Elements of proof

We are here →
Exponential separability

Definition (Exponential separability)

For each \( n \), Let \( Q_n \) and \( P_n \) be probability measures on some measurable space \((Z_n, \mathcal{Z}_n)\). Then \( (Q_n) \) is exponentially separated from \( (P_n) \), denoted \( (Q_n) \ STATES (P_n) \), if there exists a sequence \( (A_n) \) of sets \( A_n \in \mathcal{Z}_n \) such that

\[
\liminf_{n \to \infty} P_n(A_n) > 0, \quad \limsup_{n \to \infty} n^{-1} \log Q_n(A_n) < 0.
\]

Using standard properties of the Kullback-Leibler (KL) divergence one may prove the following.

Theorem

If \( (Q_n) \ STATES (P_n) \), then \( \liminf_{n \to \infty} n^{-1} KL(P_n\|Q_n) > 0. \)
Main result

Assumptions

1. The Markov kernel $Q_{\theta^*}$ is positive Harris recurrent.
2. For some integer $\ell \geq 1$, each $Q^\ell_{\theta}$ has a bounded density $q_\theta$ with respect to some $\sigma$-finite measure $\lambda$.
3. For every $\theta \neq \theta^*$,

$$\left( P_\theta^n(Y_0^n \in \cdot) \right) \not\sim \left( P_{\theta^*}(Y_0^n \in \cdot) \right).$$

4. ...

Theorem

Under the assumptions above, the MLE is strongly consistent.
Verifying separability

The assumption (3) of exponential separability is nontrivial. However, assume that for all $\theta \neq \theta^*$,

(i) $Q_\theta$ is ergodic and

(ii) $\bar{P}_{\theta^*}^Y \neq \bar{P}_\theta^Y$.

By (ii) there are $s < \infty$ and $h : Y^{s+1} \to \mathbb{R}$ such that

$$\bar{E}_\theta (h(Y_0^s)) = 0 \quad \text{and} \quad \bar{E}_{\theta^*} (h(Y_0^s)) = 1;$$

Then for sets

$$A_n \triangleq \left\{ y_0^n \in Y^{n+1} : \frac{1}{n-s} \sum_{i=s}^{n-s} h(y_i^{i+s}) > \frac{1}{2} \right\}$$

it holds, by (i), that $\bar{P}_{\theta^*}(A_n) \to 1$ and $P_\theta^\nu(A_n) \to 0$ as $n \to \infty$. 
A useful deviation inequality

Exponential separability follows if we can show that $\mathbb{P}_{\theta}^{\nu}(A_n) \to 0$ occurs at an exponential rate. We thus provide the following result.

**Proposition (Azuma-Hoeffding inequality for Markov chains)**

Assume that $Q$ is $V$-uniformly ergodic with invariant distribution $\pi$. Then there is a constant $c > 0$ such that for all $t > 0$ and all bounded functions $h : X \to \mathbb{R}$,

$$\mathbb{P}^{\eta} \left( \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i) - \pi(h) \right| \geq t \right) \leq c \nu(V) \exp \left( -\frac{n}{c} \left( \frac{t^2}{\|h\|_\infty^2} \vee \frac{t}{\|h\|_\infty} \right) \right).$$
Verifying separability

Write the quantity of interest as a telescoping of form

\[
\sum_{i=1}^{n} h(Y_i^{i+s}) = \sum_{j=0}^{s} \left( \sum_{i=1}^{n} \xi_{i,j} \right) + \sum_{i=1}^{n} \mathbb{E}_\theta(h(Y_i^{i+s})|X_{0}^{i-1}, Y_{0}^{i-1}),
\]

where \((\xi_{i,j})\) is a martingale increment sequence for each \(j\), separability is verified by combining the inequality above with the standard Azuma-Hoeffding inequality for martingale increment sequences.

Thus,

\(V\)-uniform ergodicity \(\Rightarrow\) exponential separability!
Some examples

Our assumptions can be straightforwardly verified for, e.g., very general classes of

- HMMs with finite state space and
- nonlinear HMMs with possibly non-compact state space, e.g. the popular stochastic volatility model

\[
\begin{align*}
X_{k+1} &= \alpha X_k + \sigma \epsilon_{k+1}, \\
Y_k &= \beta \exp \left( \frac{X_k}{2} \right) \epsilon_k,
\end{align*}
\]

where \((\epsilon_k)\) and \((\epsilon_k)\) are sequences of i.i.d. Gaussian variables and \(\theta = (\alpha, \beta, \sigma)\) are unknown parameters with \(|\alpha| < 1\) (Taylor, 1982).
We are here

1. Hidden Markov models
2. Previous work
3. Main results
4. Elements of proof
Our approach: Main ideas

Instead of proving (1), i.e. for all $\theta \in \Theta$,

$$\lim_{n \to \infty} n^{-1} \log p_{\theta}^\nu(Y^n_0) = H(\theta, \theta^*), \quad \text{(a.s.)}$$

we note that it is enough to establish

1. the limit

$$\lim_{n \to \infty} n^{-1} \log p_{\theta^*}^\nu(Y^n_0) = H(\theta^*, \theta^*), \quad \text{(a.s.)}$$

which is easily obtained using the SBM theorem, and

2. the following bound: For all closed sets $C \subset \Theta$ with $\theta^* \notin C$,

$$\limsup_{n \to \infty} n^{-1} \sup_{\theta \in C} \log p_{\theta}^\nu(Y^n_0) < H(\theta^*, \theta^*) \quad \text{(a.s.)}$$
Our approach: A blocking technique

To establish the bound, consider blocks of observations of length $\ell$ and bound (roughly) the likelihood according to

$$
\log p^\nu_\theta(Y_0^n) \leq \frac{1}{\ell} \sum_{k=0}^{n-\ell} \log p^\lambda_\theta(Y_{k+\ell}^k) + o_{\text{a.s.}}(n)
$$

for some constant $c > 0$. This implies directly, via Birkhoff’s ergodic theorem,

$$
\limsup_{n \to \infty} n^{-1} \log p^\nu_\theta(Y_0^n) \leq \bar{E}_{\theta^*} \left( \ell^{-1} \log p^\lambda_\theta(Y_0^\ell) \right) \quad (\text{a.s.})
$$
Our approach: Identifiability

To complete the proof we use the assumed exponential separability and its connection to the KL divergence: For all $\theta \neq \theta^*$,

\[
\left( \mathbb{P}_\theta^n(Y_0^n \in \cdot) \right) \not\rightarrow \left( \mathbb{P}_{\theta^*}^n(Y_0^n \in \cdot) \right)
\]

Lemma

\[
\lim_{n \to \infty} \inf \mathbb{E}_{\theta^*} \left( n^{-1} \log \frac{\bar{p}_{\theta^*}(Y_0^n)}{p_{\theta}(Y_0^n)} \right) > 0
\]

\[
\Rightarrow \lim_{n \to \infty} \sup \mathbb{E}_{\theta^*} \left( n^{-1} \log p_{\theta}(Y_0^n) \right) < H(\theta^*, \theta^*)
\]

\[
\Rightarrow \exists \ell_{\theta} : \mathbb{E}_{\theta^*} \left( \ell_{\theta}^{-1} \log p_{\theta}(Y_0^{\ell_{\theta}}) \right) < H(\theta^*, \theta^*).
\]
Combining this with the previous gives

$$\limsup_{n \to \infty} n^{-1} \log p_{\theta}(Y_0^n) < H(\theta^*, \theta^*) \quad (\text{a.s.})$$

and, after some more work, using that the parameter space $\Theta$ is compact, for all closed sets $C \subset \Theta$ with $\theta^* \notin C$,

$$\limsup_{n \to \infty} n^{-1} \sup_{\theta \in C} \log p_{\theta}(Y_0^n) < H(\theta^*, \theta^*) \quad (\text{a.s.})$$

This completes the proof of consistency.
We have

- introduced HMMs and the problem of maximum likelihood-based inference in such models.
- proved that the MLE is strongly consistent under (what we believe) minimal assumptions.
- discussed an information-theoretic device, exponential separability, which is efficiently used in our proof to establish identifiability.
- Shown how exponential separability can be verified using a novel Azuma-Hoeffding-type inequality for \( V \)-uniformly ergodic Markov chains—a result of independent interest.