Online EM Algorithm for hidden markov models

Computational Statistics

Adil Rhoulam
Ayoub Ghriss

MVA, 2017

Contents

Why HMM ?

- Considered a key concept of statistical time series analysis
- Sufficiently simple to give rise to efficient inference procedures

Hidden Markov Model

An HMM is specified by the following components:

- χ=1,2,3...K : the latent variables space
- $A = (a_{i,j})$ a transition probability matrix, each $a_{i,j}$ representing the probability of moving from state i to state j
- a sequence (Y_i) of T observations,
- $p(y_i/x_j)$ a sequence of observation likelihoods
- π_0, a special initial distribution of over the states

Likelihood Computation

For a particular sequence of hidden states $X_1, X_2, ...X_T$, the likelihood of observing the sequence is :

$$P(Y|X) = \Pi_i P(y_i|x_i)$$

We also define two quantities:

- alpha-message : $\alpha_t(x_t) = p(x_t, y_0, \dots, y_t)$
- beta-message : $\beta_t(x_t) = p(y_{t+1} \dots y_T | x_t)$

Forward-Backward Recursion

With with $\beta_T = 1$ and $\alpha_0 = \pi_0$, the two messages follow the recursion:

- $\alpha_{t+1}(x_{t+1}) = p(y_{t+1}|x_{t+1}) \sum_{x_t} p(x_{t+1}|x_t)\alpha(x_t)$
- $\beta_t(x_t) = \sum_{x_{t+1}} p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})\beta_{t+1}(x_{t+1})$

Using alpha and beta messages, we easily obtain the following probabilities:

- $\gamma_t(x_t) = p(x_t|y_0, \dots, y_T) \propto \alpha_t(x_t)\beta_t(x_t)$
- For all $t < T$, $\phi_t(x_t, x_{t+1}) = p(x_t, x_{t+1}|y_0, \dots, y_T) \propto \alpha_t(x_t)\, \beta_{t+1}(x_{t+1})\, p(x_{t+1}|x_t)$

Under the probability q of having observed $(y_1, \ldots .y_t)$, we have:

$$\mathbb{E}_q[l_c(\theta)] = \sum_{i=1}^{K} \gamma_0(i) \log((\pi_0)_i)$$

$$+ \sum_{t=0}^{T-1} \sum_{i,j=1}^{K} \phi_t(i,j) \log(A_{i,j}) + \sum_{t=0}^{T} \sum_{i=1}^{K} \gamma_t(i) \log p(\overline{y_t}, i)$$

EM for HMM

We define the matrix ξ_t as follows :

$$\xi_t(i,j) \propto= \alpha_t(i) A_{i,j} p(y_{t+1}|j) \beta_{t+1}(j)$$

If we consider EM algorithm to optimise the transition matrix, the the transition matrix update $\hat{A}$ at the M-step would be :

$$\hat{A}_{i,j} = \frac{\sum_t \xi_t(i,j)}{\sum_j \sum_t \xi(i,j)}$$

EM for Gaussian Mixture HMM

For Gaussian Mixture HMM, we assume that:

$$p(y_t|x_t = i) \sim N(\mu_i, \Sigma_i)$$

Applying the EM for Gaussian HMM, we update the transition matrix A as previously showed and the means (μ_i) and matrix (Σ_i) as follows :

$$\hat{\mu}_i = \frac{\sum_t \gamma_t(i) y_t}{\sum_t \gamma_t(i)}$$

$$\hat{\Sigma}_i = \frac{\sum_t \gamma_t(i)(y_t - \hat{\mu}_i)(y_t - \hat{\mu}_i)^T}{\sum_t \gamma_t(i)}$$

Online EM - Assumptions

The usual EM for HMM imposes computations over all the observed data $(y_1, ..; y_T)$ which can be time consuming when T is large. We represent an online EM for HMM over exponential families:

$$p_\theta(x_t, y_t | x_{t1}) = h(x_t, y_t) exp(< \psi(\theta), s(x_{t1}, x_t, y_t) > -A(\theta))$$

We also assume we have an Explicit M-step:

$$\nabla_\theta \psi(\theta) S - \nabla_\theta A(\theta) = 0$$

as a unique solution (in the set of sufficient statistics) denoted by $\bar{(}\theta(S))$.

Online EM - Assumptions

For a more specific form, we write the transition probabilities as :

$$q_\theta(x', x) = h^q(x', x) exp(< \psi^q(\theta), s(x', x) > -A^q(\theta))$$

and the observations likelihoods:

$$g_\theta(x, y) = h^g(x, y) exp(< \psi^g(\theta), s(x, y) > -A^g(\theta))$$

In which case :

$\psi_\theta = [\psi^q(\theta), \psi^g(\theta)]$ and $s(x', x, y) = [s(x', x), s(x, y)]$

The Usual k+1 iteration:
E-Step would be to evaluate:

$$S_{k+1} = \frac{1}{T} E_{\nu,\theta_k}[\sum_t S(X_{t1}, X_t, Y_t) | Y0 : n]$$

M-Step:

$$\theta_{k+1} = \overline{\theta(S_{k+1})}$$

The idea of the online EM is to only compute at iteration n : $(S(X_{t1}, X_t, Y_n))$ with other quantities that do not depend on the newly observed data Y_n

Online EM: The Algorithm

Initialise the parameters of the HMM model, n_{min} and choose a step-size sequence $(\gamma_n)_n$, which satisfy:

$$\begin{cases} \sum_n \gamma_n = \infty \\ \sum_n \gamma_n^2 < \infty \end{cases}$$

Initialisation of the online EM:

$$\begin{cases} \phi_0(x) = \frac{\nu(x) g_{\theta_0}(x, Y_0)}{\sum_{x'} \nu(x') g_{\theta_0}(x', Y_0)} \\ \rho_0(x) = 0 \end{cases}$$

The loop: n=0

n $\rightarrow n+1$:

$$\forall x \in \chi : \phi_{n+1}(x) = \frac{\sum_{x'} \phi_n(x) q_{\theta_n}(x',x) g_{\theta_n}(x, Y_{n+1})}{\sum_{x',x"=1} \phi_n(x') q_{\theta_n}(x',x") g_{\theta_n}(x", Y_{n+1})}$$

$$\rho_{n+1}(x) = \sum_{x'} \left(\gamma_{n+1} s(x', x, Y_{n+1} + (1-\gamma_{n+1})\rho_n(x')\right) r_{n+1,\nu,\theta}(x'|x)$$

If n>n$_{min}$:

$$\theta_{n+1} = \overline{\theta}(\textstyle\sum_x \rho_{n+1}(x)\phi_{n+1}(x))$$

Where:

r$_{n+1,\nu,\theta}(x'|x) = \frac{\phi_{n,\nu,\theta}(x') q_\theta(x',x)}{\sum_{x''} \phi_{n,\nu,\theta}(x'') q_\theta(x'',x)} = P_{\nu,\theta}(X_n = x' \Big| X_{n+1} = x_{n+1}, Y0 : n)$

Still missing but some results were proven using Douc. et Al. assumptions (χ finite, the parameter space is compact, the transition matrix is lower bounded by an ϵ>0 ...).
When we don't do the step M, ρ_n converges to a deterministic quantity that doesn't depends on x.

In the gaussian HMMs we suppose that, if $(X_1, ..., X_n)$ are latent variables and $(Y_1, ..., Y_n)$ are observations, then we have:

$$\begin{cases} \theta_t = \{\nu, q_t(X_{t-1} = i, X_t = j), \mu_k, \Sigma_k\} \\ p_{\theta\,t}(X_{t+1}/X_t) = q_t(X_t, X_{t+1}) \\ p_{\theta\,t}(Y_t/X_t = k) = g_{\theta_t}(k, Y_t) = N(Y_t/\mu_k{,}\Sigma_k) \end{cases}$$

Then we get:

$$\mathrm{p}_\theta(X_t, Y_t/X_{t-1}) = \sum_{i,j} \delta(X_{t-1} = i, X_t = j) \log(q(i,j)) + \sum_i \delta(X_t = i)\left[\frac{-1}{2} Y_t^T \Sigma^{-1} Y_t + \mu^T \Sigma^{-1} Y_t - \frac{-1}{2}\mu^T \mu - log(\Sigma^{-1})\right] + ...$$

We pick the sufficient statistics:

$$\begin{cases} s^q(X_{t-1}=i, X_t=j) = \delta(X_{t-1}=i, X_t=j, Y_t) \\ s_1^g(X_t=i, Y_t) = \delta(X_t=i) \\ s_2^g(X_t=i, Y_t) = \delta(X_t=i)Y_t \\ s_3^g(X_t=i, Y_t) = \delta(X_t=i)Y_t^T Y_t \end{cases}$$

Which leads to intermediate quantities:

$$\begin{cases} \rho_{n+1,\theta}^q(i,j,k) = \frac{1}{n}E_{v,\theta}\left[\sum_{t=0}^{n+1} s^q(X_{t-1}=i, X_t=j) \middle| Y_{0:n}, X_n=k\right] \\ \rho_{n+1,d,\theta}^g(i,k) = \frac{1}{n}E_{v,\theta}\left[\sum_{t=0}^{n+1} s_d^g(X_t=i, Y_t) \middle| Y_{0:n}, X_n=k\right]/d \in \{1,2,3\} \end{cases}$$

The Intialization step becomes:

$$\begin{cases} \rho^{q}_{0,\theta}(i,j,k) = 0 \\ \rho^{g}_{0,1,\theta}(i,k) = \delta(i-k) \\ \rho^{g}_{0,2,\theta}(i,k) = \delta(i-k) Y_{n+1} \\ \rho^{g}_{0,3,\theta}(i,k) = \delta(i-k) Y^{T}_{n+1} Y_{n+1} \end{cases}$$

The Filter update step becomes:

$$\phi_{n+1}(k) = \frac{\sum_{k'=1}^{K} \phi_n(k') q_n(k',k) g_{\theta_n}(k, Y_{n+1})}{\sum_{k',k"=1}^{K} \phi_n(k') q_n(k',k") g_{\theta_n}(k", Y_{n+1})}$$

The stochastic Approximation E-step becomes:

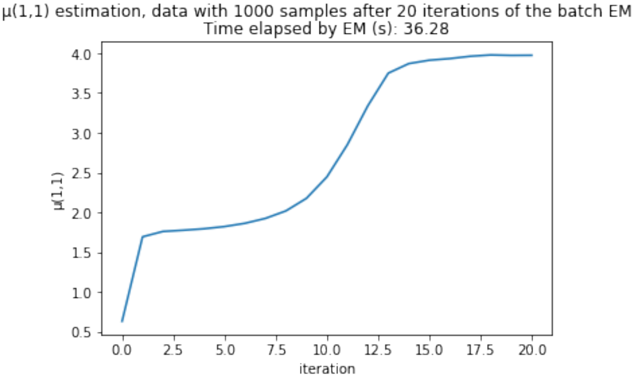
$$\rho^q_{n+1,\theta}(i,j,k) = \gamma_{n+1}\delta(j-k) r_{n+1}(i|j) + (1-\gamma_{n+1}) \sum_{k'=1}^{K} \rho^q_{n,\theta}(i,j,k') r_{n+1}(k'|k)$$
$$\rho^q_{n+1,1,\theta}(i,k) = \gamma_{n+1}\delta(i-k) + (1-\gamma_{n+1}) \sum_{k'=1}^{K} \rho^q_{n,1,\theta}(i,k') r_{n+1}(k'|k)$$
$$\rho^q_{n+1,2,\theta}(i,k) = \gamma_{n+1}\delta(i-k) Y_{n+1} + (1-\gamma_{n+1}) \sum_{k'=1}^{K} \rho^q_{n,1,\theta}(i,k') r_{n+1}(k'|k)$$
$$\rho^q_{n+1,3,\theta}(i,k) = \gamma_{n+1}\delta(i-k) Y_{n+1}^T Y_{n+1} + (1-\gamma_{n+1}) \sum_{k'=1}^{K} \rho^q_{n,1,\theta}(i,k') r_{n+1}(k'|k)$$

Where:

$$\mathrm{r}_{n+1}(i|j) = \frac{\phi_n(i) q_{\theta_n}(i,j)}{\sum_{k'=1}^{K} \phi_n(k') q_{\theta_n}(k',j)}$$

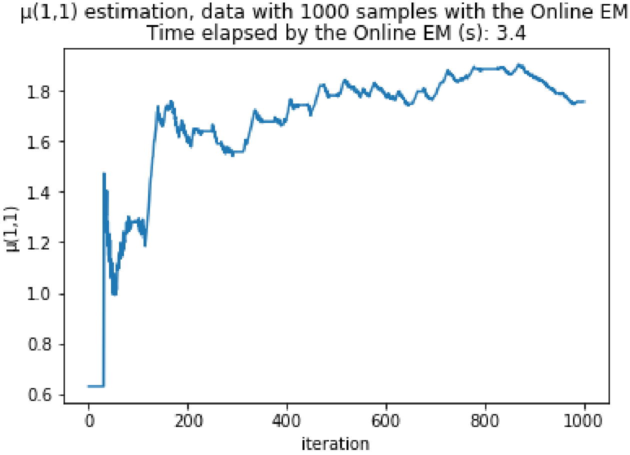
Implementation and results

We have implemented 2D gaussian HMM with K=4 (4 states) by using the usual EM and the Online EM. The initial means and covariances matrices are chosen randomly, The transition density is chosen $A_{i,i} = \frac{1}{2}$ and $A_{i,j} = \frac{1}{6}$ if i$\neq j$

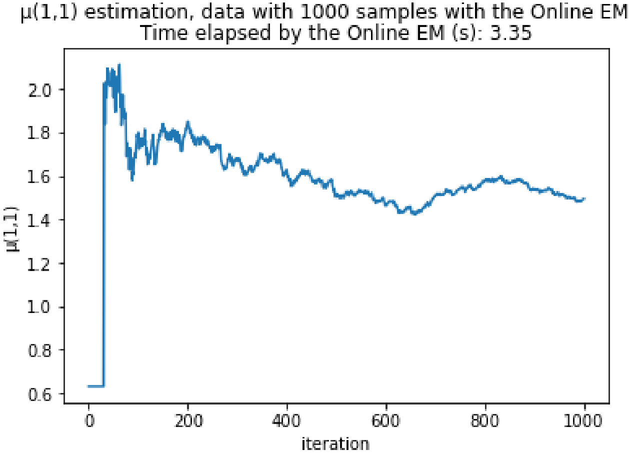


μ(1,1) estimation, data with 1000 samples after 20 iterations of the batch EM
Time elapsed by EM (s): 36.28
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
μ(1,1)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
iteration

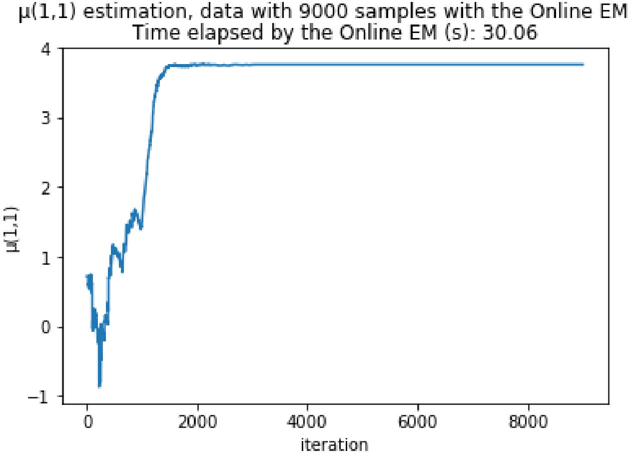
Implementation and results

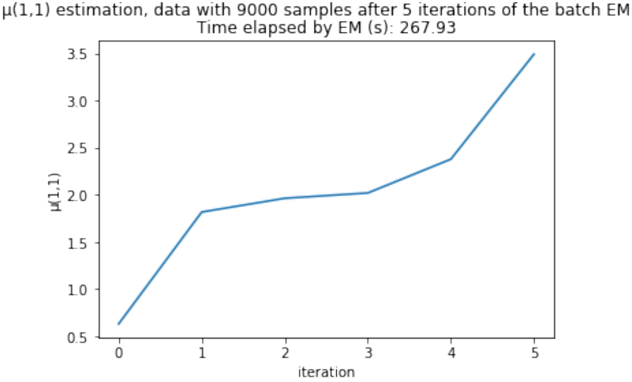


Implementation and results



Implementation and results




μ(1,1) estimation, data with 9000 samples after 5 iterations of the batch EM
Time elapsed by EM (s): 267.93
μ(1,1)
iteration
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
1
2
3
4
5

Conclusions

Pros : - Fast than the usual EM when comparing the convergence.

- More accurate than the batch EM.

- For larger samples, they are preferable since they converge rapidly.

Cons : - Using Online EM with small sample sizes gives very poor result in both terms (accuracy and variability).

- High variance of the estimations.

- complexity grows exponentially with the number of states.

- Depends highly on the choice of the step-size sequence.

- Theoretical analysis of the convergence are still missing.