école normale supérieure paris—saclay

Online EM Algorithm for Hidden Markov Model

Analysis of an article of Olivier Cappé

Célia DOCLOT and Yoach LACOMBE

MVA Students

Computational Statistics Course

ENS Paris-Saclay University

April 2, 2022

Table Of Contents

Overview and Definitions

Goal : estimate in an online manner the parameters of an HMM.

Définition - Online: we only scan each observation once.

Définition - Hidden Markov Model (HMM) : Time-series consisting of state sequences (X_t) and observations Y_t, generated by unknown parameters θ^* through the joint density P_{ν,θ^*} where ν is an initial pdf giving the initial state X_0. Here, we suppose that the X_t takes values in a finite set.

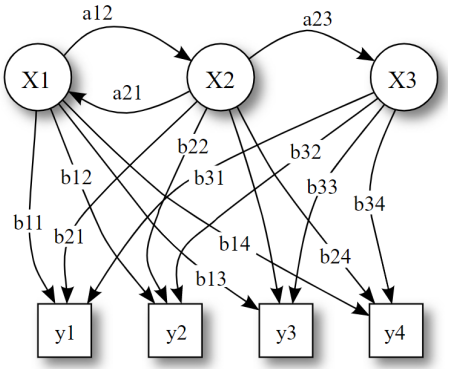

Figure: credits: Tdunning - Wikipedia

Main idea and assumptions

Main Idea: The author introduces the idea of an online EM algorithm twisted in such a manner that each step only depends on one sample.

Main Hypothesis:

- **Exponential family:**

$$p_\theta(x_t, y_t|x_{t-1}) = h(x_t, y_t)exp(\psi(\theta)^T s(x_{t-1}, x_t, y_t) - A(\theta))$$

  Where s is the complete-data sufficient statistics which takes values in $\mathcal{S}$.

- **Explicit M-Step:** For all S in $\mathcal{S}$, $\nabla_\theta \psi(\theta)S - \nabla_\theta A(\theta) = 0$ has an unique solution denoted $\bar{\theta}(S)$.

Building up the algorithm

- **E-step :** $S_{k+1} = \frac{1}{n}\mathbb{E}_{\nu,\theta_k}\left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) | Y_{0:n}\right]$
- **M-step :** $\theta_{k+1} = \bar{\theta}(S_{k+1})$

Note : The initial term $\log p_{\nu,\theta}(x_0, y_0)/n$ is vanishing.

Problem : Computing S_{k+1}

Solution : recursive form of smoothing $S_{n+1} = \sum_{x\in\mathcal{X}} \phi_{n,\nu,\theta}(x)\rho_{n,\nu,\theta}(x)$

With :

$$\phi_{n,\nu,\theta}(x) = \mathbb{P}_{\nu,\theta}(X_n = x | Y_{0:n})$$

$$\rho_{n,\nu,\theta}(x) = \frac{1}{n}\mathbb{E}_{\nu,\theta}\left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) | Y_{0:n}, X_n = x\right]$$

The Algorithm

Initialisation, $\hat{\theta} = \theta_0$:

- For x in $\mathcal{X}$
  - $\hat{\phi}_0(x) = \frac{\nu(x) g_{\hat{\theta}}(x, Y_0)}{\sum_{x' \in \mathcal{X}} \nu(x') g_{\hat{\theta}}(x', Y_0)}$
  - $\hat{\rho}_0(x) = 0$

Recursion on n :

- For x in $\mathcal{X}$
  - $\hat{\phi}_{n+1}(x) = \sum_{x' \in \mathcal{X}} \hat{\phi}_n(x) q_{\hat{\theta}}(x', x) g_{\hat{\theta}}(x, Y_{n+1})$ (normalized)
  - $\hat{\rho}_{n+1}(x) = \sum_{x' \in \mathcal{X}} \gamma_{n+1} s(x', x, Y_{n+1}) + (1 - \gamma_{n+1}) \hat{\rho}_n(x') \hat{r}_{n+1,\hat{\theta}}(x'|x)$
    with : $\hat{r}_{n+1,\hat{\theta}}(x'|x) = \hat{\phi}_n(x') q_{\hat{\theta}}(x', x)$ (normalized). $\hat{r}_{n+1,\hat{\theta}}(x'|x)$ can be interpreted as $\mathbb{P}_{\nu,\hat{\theta}}(X_n = x' | X_{n+1} = x, Y_{0:n})$ (backward retrospective probability).
- $\hat{\theta} \leftarrow \bar{\theta}\left(\sum_{x \in \mathcal{X}} \hat{\phi}_{n+1}(x) \hat{\rho}_{n+1}(x)\right)$

Algorithm - Experiment

The experiment : The article studied in depth a simple problem: a 2-state Markov Chain with an additive Gaussian noise.

The model: Two states 0 and 1, with respective mean $\mu(0)$ and $\mu(1)$ and a transition matrix q. We have access to this state with centered Gaussian Noise with variance v.

$$Y_t = X_t + V_t \text{ where } V_t \overset{i.i.d}{\sim} \mathcal{N}(0, v)$$

Parameters: We want to estimate (q, μ, v).

Difficulty: The variance is chosen such that the problem is difficult.

Figure: The Markov Chain

$$v = 0.5, \mu = [0, 1]^T$$

Experiment - Set-Up

- We reproduced the article's experiments on Python 3.
- The problem's design allows to compute in a relatively simple manner the solution, since the problem respects the usual assumptions and that $\hat{\theta}$ and are easily computable.
- We compared the **online EM algorithm** to the **batch EM algorithm with backward-forward computation**.
- Each algorithm runs 100 independent times and with the same initial values.

Figure: μ_{init} and q_{init}

$$v_{init} = 2$$

Experiment - results - limits of the batch EM

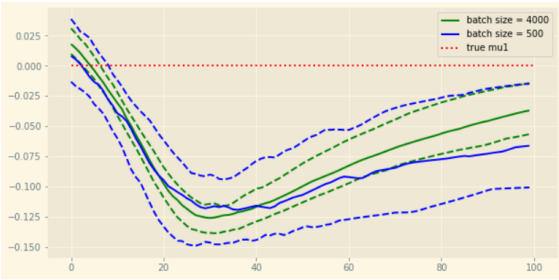

Figure: $\mu_{estimated}(1)$ as a function of the batch EM iterations.

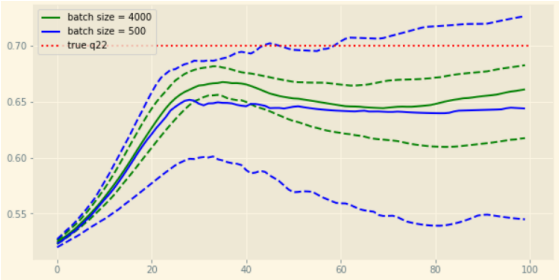

Figure: $_{estimated}(1, 1)$ as a function of the batch EM iterations.

Notice the bias that seems to appears (du to the initial values and despite theoretical consistency).

Experiment - results - $q(1,1)$

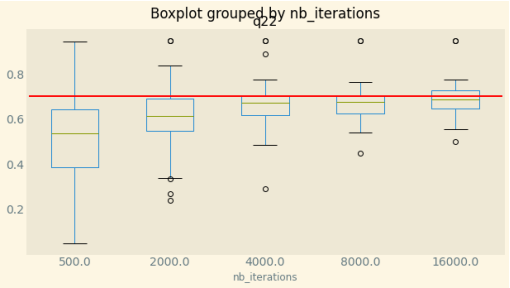

Figure: Boxplot of $q_{estimated}(1,1)$ after 100 runs of the online EM

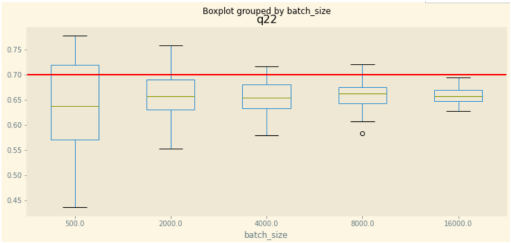

Figure: Boxplot of $q_{estimated}(1,1)$ after 100 runs of the batch EM with different batch size. Each EM has been iterated 50 times.

The batch EM's bias is confirmed. The online EM seems to have an overall better behaviour but still has some limits (see next slide).

Experiment - results - $\mu(1)$

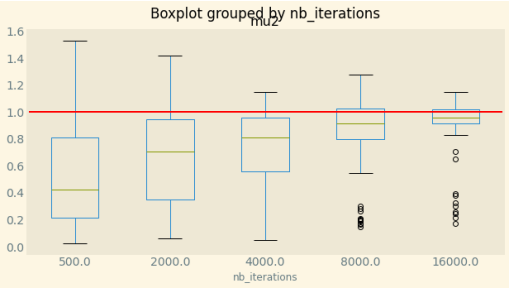

Figure: Boxplot of $\mu_{estimated}(1)$ after 100 runs of the online EM

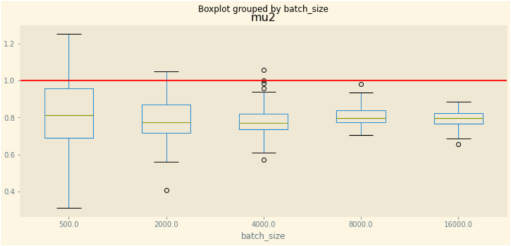

Figure: Boxplot of $\mu_{estimated}(1)$ after 100 runs of the batch EM with different batch size. Each EM has been iterated 50 times.

We observe a behaviour that was not indicated in the article. The online EM algorithm often falls into a trap where it only predicts one state.

Experiment - results - v

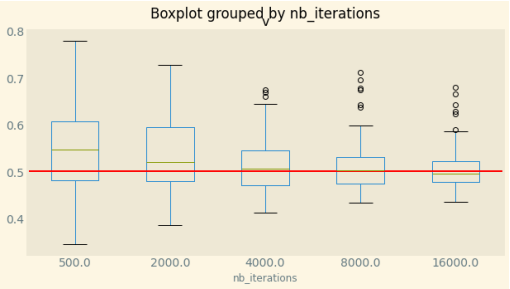

Figure: Boxplot of $v_{estimated}$ after 100 runs of the online EM

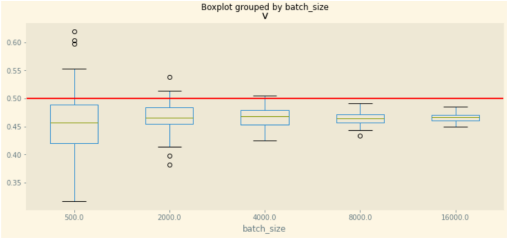

Figure: Boxplot of $v_{estimated}$ after 100 runs of the batch EM with different batch size. Each EM has been iterated 50 times.

The high predicted variances on some of the online EM iterations are du to this trap.

Experiment - results - Additional observations

- Online EM algorithm sometimes fails when the initial observation is not informative enough
- **Time Comparison:**

Algorithm	Batch length	Mean time (s)
Online EM	1	0.0001875
Online EM	16k	3
Batch EM (50 iterations)	500	1 to 2
Batch EM (50 iterations)	2k	4.5
Batch EM (50 iterations)	16k	30

What brings this article...

...Compared to other articles of the bibliography :

- *Online Learning with Hidden Markov Models*, Gianluigi Mongillo and Sophie Deneve

Reduced to the case $\mathcal{Y}$ finite, include in our article. Proposition for continuous observation space : subdividing into bins and recover the parameters of the pdf from the matrix.

.

- *Online EM Algorithm for Latent Data Models*, Olivier Cappé, Eric Moulines

One more important assumption : $\bar{s}(y;\theta) \equiv \mathbb{E}_\theta[S(X)|Y=y]$ is well defined for all $(y,\theta) \in \mathcal{Y} \times \Theta$ so the E-step becomes $\hat{s}_{n+1} = \hat{s}_n + \gamma_{n+1}(\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n)$ which is impossible in the example of our article.

What brings this article...

...Compared the algorithm of the course :

The three articles use **Online learning** : each iteration needs the previous one and a new observation point, not the whole data at every step.

Convergence Properties

We have for θ fixed, $\forall x \in \mathcal{X}, \hat{\rho}_n(x) \to \mathbb{E}_{\theta_*}[\mathbb{E}_\theta(s(X_{-1}, X_0, Y_0|Y_{-\infty:+\infty})], \mathbb{P}_{\theta^*}$ a.s. so

$$\bar{\theta}(S_{n+1}) \to \bar{\theta}(\mathbb{E}_{\theta_*}[\mathbb{E}_\theta(s(X_{-1}, X_0, Y_0|Y_{-\infty:+\infty})])$$

With new assumptions it is proved that the fixed points of $\theta_{k+1} = \bar{\theta}(\mathbb{E}_{\theta_*}[\mathbb{E}_\theta(s(X_{-1}, X_0, Y_0|Y_{-\infty:+\infty})])$ are the stationary points of $\mathbb{E}_{\theta_*}[\log l_\theta(Y_0|Y_{-\infty:-1})]$

...

$\to$ The convergence works in our example but isn't proved

$\to$ To have elements we have to fix θ

Conclusion

Strengths:

+ Really quick
+ Works in our example - doesn't seem to have bias
+ No hard constraints in assumptions

Weaknesses:

- No proof of convergence in general
- Problems when the states are close or with unbalanced transitions
- Crashes with the wrong initialization
- Most efficient when the number of states is low

Thank you

for listening.