`osl-dynamics`: HMM Cost Function

C. Gohil

OHBA, Department of Psychiatry, Warneford Hospital, Oxford, OX3 7JX

(March 6, 2023)

Abstract

We describe the calculation of the cost function used to update the observation model parameters (state means and covariances) in the `osl-dynamics` implementation of a Hidden Markov Model (HMM).

1 Variational Free Energy

In variational Bayesian inference we infer a posterior distribution for model parameters, $q(.)$, by minimising the *variational free energy*, $\mathcal{F}$, given some data we have observed, x_t. For the HMM, our model parameters are:

- The hidden state at each time point, s_t.
- The state transition probability at each time point, π_t, which is dependent on s_{t-1}.
- The initial state probability, π_1.
- The observation model parameters, θ_{obs}.

Therefore, we infer our model parameters by minimising the following variational free energy[1] [1]

$$\mathcal{F} = \iiiint q(s_{1:T})q(\pi_t)q(\pi_1)q(\theta_{\text{obs}}) \log \left[\frac{q(s_{1:T})q(\pi_t)q(\pi_1)q(\theta_{\text{obs}})}{p(x_{1:T}, s_{1:T}, \pi_t, \pi_1, \theta_{\text{obs}})} \right] ds_{1:T} d\pi_t d\pi_1 d\theta_{\text{obs}}, \tag{1}$$

where $s_{1:T}$ and $x_{1:T}$ denote $s_1, ..., s_T$ and $x_1, ..., x_T$ respectively. However, in the `osl-dynamics` implementation of an HMM, we will not be Bayesian on θ_{obs}, instead of learning $q(\theta_{\text{obs}})$ we will learn point estimates for θ_{obs}. We will learn the posterior distributions $q(s_{1:T})$, $q(\pi_t)$, $q(\pi_1)$ and point estimates for θ_{obs} by minimising the following variational free energy,

$$\mathcal{F} = \iiint q(s_{1:T})q(\pi_t)q(\pi_1) \log \left[\frac{q(s_{1:T})q(\pi_t)q(\pi_1)}{p(x_{1:T}, s_{1:T}, \pi_t, \pi_1)} \right] ds_{1:T} d\pi_t d\pi_1. \tag{2}$$

We will show that Eq. (2) implicitly depends on the point estimates for θ_{obs} below.

2 Generative Model

The term $p(x_{1:T}, s_{1:T}, \pi_t, \pi_1)$ is determined by our generative model. For the HMM, if we were being fully Bayesian this would be [1]

$$p(x_{1:T}, s_{1:T}, \pi_t, \pi_1, \theta_{\text{obs}}) = p(x_1|s_1, \theta_{\text{obs}})p(s_1|\pi_1)p(\pi_1)p(\theta_{\text{obs}}) \prod_{t=2}^{T} p(x_t|s_t, \theta_{\text{obs}})p(s_t|s_{t-1}, \pi_t)p(\pi_t). \tag{3}$$

[1] We have used the mean field approximation.

However, because we are learning point estimates for θ_{obs} we do not have the prior $p(\theta_{\text{obs}})$. We will use the following generative model,

$$p(x_{1:T}, s_{1:T}, \pi_t, \pi_1) = p(x_1|s_1, \theta_{\text{obs}})p(s_1|\pi_1)p(\pi_1)\prod_{t=2}^{T} p(x_t|s_t, \theta_{\text{obs}})p(s_t|s_{t-1}, \pi_t)p(\pi_t), \quad (4)$$

where θ_{obs} are point estimates. We assume a multivariate normal distribution for the observed data,

$$p(x_t|s_t = k, \theta_{\text{obs}}) = \mathcal{N}(m_k, C_k), \quad (5)$$

where m_k and C_k are the mean and covariance for state k respectively. Our observation model parameters θ_{obs} are the set of state means and covariances, $\theta_{\text{obs}} = \{m_k, C_k\}$.

3 Cost Function for Learning $\theta_{\text{obs}} = \{m_k, C_k\}$

We update our point estimate for θ_{obs} by minimising Eq. (2). We separate Eq. (2) into the following terms[2]

$$\begin{aligned}
\mathcal{F} = &-\iiint q(s_{1:T})q(\pi_t)q(\pi_1)\log\left[p(x_{1:T}, s_{1:T}, \pi_t, \pi_1)\right] ds_{1:T}d\pi_t d\pi_1 \\
&+ \iiint q(s_{1:T})q(\pi_t)q(\pi_1)\log\left[q(s_{1:T})q(\pi_t)q(\pi_1)\right] ds_{1:T}d\pi_t d\pi_1 \\
\mathcal{F} = &-\iiint q(s_{1:T})q(\pi_t)q(\pi_1)\log\left[p(x_{1:T}, s_{1:T}, \pi_t, \pi_1)\right] ds_{1:T}d\pi_t d\pi_1 \\
&+ \int q(s_{1:T})\log\left[q(s_{1:T})\right] ds_{1:T} + \int q(\pi_t)\log\left[q(\pi_t)\right] d\pi_t + \int q(\pi_1)\log\left[q(\pi_1)\right] d\pi_1
\end{aligned} \quad (6)$$

Only the first term depends on θ_{obs} so the rest can be ignored. Substituting Eq. (4) into the first term, we have

$$\begin{aligned}
\mathcal{F} &\propto -\iiint q(s_{1:T})q(\pi_t)q(\pi_1)\log\left[p(x_{1:T}, s_{1:T}, \pi_t, \pi_1)\right] ds_{1:T}d\pi_t d\pi_1 \\
&\propto -\iiint q(s_{1:T})q(\pi_t)q(\pi_1) \\
&\qquad \log\left[p(x_1|s_1, \theta_{\text{obs}})p(s_1|\pi_1)p(\pi_1)\prod_{t=2}^{T} p(x_t|s_t, \theta_{\text{obs}})p(s_t|s_{t-1}, \pi_t)p(\pi_t)\right] \\
&\qquad ds_{1:T}d\pi_t d\pi_1.
\end{aligned} \quad (7)$$

Again, only retaining the factors that depend on θ_{obs}, we have

$$\begin{aligned}
\mathcal{F} &\propto -\int q(s_{1:T})\log\left[\prod_{t=1}^{T} p(x_t|s_t, \theta_{\text{obs}})\right] ds_{1:T} \\
&\propto -\sum_{t=1}^{T}\int q(s_{1:T})\log\left[p(x_t|s_t, \theta_{\text{obs}})\right] ds_{1:T} \\
&\propto -\sum_{t=1}^{T}\int \dots \int q(s_1)\dots q(s_T)\log\left[p(x_t|s_t, \theta_{\text{obs}})\right] ds_1\dots ds_T \\
&\propto -\sum_{t=1}^{T}\int q(s_t)\log\left[p(x_t|s_t, \theta_{\text{obs}})\right] ds_t = \mathcal{L}.
\end{aligned} \quad (8)$$

[2] We have used $\int q(\xi)d\xi = 1$ to evaluate some of the integrals.

Here, we have defined the negative log-likelihood loss, $\mathcal{L}$, which is minimised via stochastic gradient descent to learn the parameters θ_{obs}. As $q(s_t)$ is a discrete probability distribution for the state, we can evaluate the integral as

$$\begin{aligned}\mathcal{L} &= -\sum_{t=1}^{T}\sum_{k=1}^{K} q(s_t = k) \log\left[p(x_t|s_t = k, \theta_{\text{obs}})\right] \\ &= -\sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_{kt} \log\left[p(x_t|s_t = k, \theta_{\text{obs}})\right],\end{aligned} \tag{9}$$

where K is the number of states and $q(s_t = k) = \gamma_{kt}$ is the probability of state k at time t. Substituting Eq. (5) into this we have

$$\mathcal{L} = -\sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_{kt} \log\left[\mathcal{N}(m_k, C_k)\right], \tag{10}$$

which is the log-likelihood loss function implemented in `osl-dynamics` for inferring the point estimates for the observation model parameters $\theta_{\text{obs}} = \{m_k, C_k\}$.

References

[1] I. Rezek and S. Roberts, Ensemble hidden Markov models with extended observation densities for biosignal analysis. Probabilistic modeling in bioinformatics and medical informatics. Springer, London, 419-450 (2005).