Spectral Learning for Mixture of Markov Models Supplemental Material

Y.Cem Sübakan, Barış Kurt, A. Taylan Cemgil, Bülent Sankur

October 9,2013

1 Spectral Learning for Mixture of Markov Models

In lieu of using local search algorithms such as EM, we can estimate the transition matrix A_k of each cluster with spectral learning, solely based on some observable moments, and in a local optima-free fashion. According to the method of moments based learning algorithms described in [1], the trick is to express the moments of the distribution as a matrix multiplication (or possibly tensor as in our case), so that an eigen-decomposition can be applied. The latter reveals information about the model parameters, and they can be obtained using a function of some observable moments. To demonstrate the idea and set the notation let us first apply the procedure on a multi-view mixture model [1].

1.1 Spectral Learning for multi-view mixture model

We will demonstrate how to derive the learning algorithm of [1] for a multi-view mixture model. Let us denote the cluster indicator variable with $h \in \{1, \dots K\}$. The prior distribution on clustering assignments is defined as $\pi = p(h) \in \mathbb{R}^L$. The observations are denoted by $x_1, x_2, x_3 \in \{1 \dots, L\}$. The observation model is defined as $O = p(x_1|h) = p(x_2|h) = p(x_3|h) \in \mathbb{R}^{L \times K}$. We use the diag(.) operator to construct a diagonal matrix out of a vector. We also use the MATLAB array notation: $A(:, k)$, $A(k, :)$, which respectively picks k'th column and row of A matrix. Having a discrete observation model, the second and third order moments can be written down as follows:

$$p(x_1, x_2) = p(x_1|h)\text{diag}(\pi)p(x_2|h)^T \tag{1}$$

$$p(x_1, x_2, x_3 = i) = p(x_1|h)\text{diag}(\pi)\text{diag}(p(x_3 = i|h))p(x_2|h)^T \tag{2}$$

Then, by inserting the term $I = p(x_1'|h)^{-1}p(x_1'|h)$, we see that;

$$\begin{aligned} p(x_1, x_2, x_3 = i) =& p(x_1|h)\text{diag}(\pi)\text{diag}(p(x_3 = i|h))Ip(x_2|h)^T \\ =& p(x_1|h)\text{diag}(p(x_3 = i|h))I\text{diag}(\pi)p(x_2|h)^T \\ =& p(x_1|h)\text{diag}(p(x_3 = i|h))p(x_1'|h)^{-1} \\ & \times \underbrace{p(x_1'|h)\text{diag}(\pi)p(x_2|h)^T}_{p(x_1', x_2)} \end{aligned}$$

Then, we observe that,

$$\begin{aligned} B_i :=& p(x_1, x_2, x_3 = i)p(x_1', x_2)^{-1} \\ =& O\text{diag}(O(i,:))O^{-1} \end{aligned} \tag{3}$$

So, we conclude that the observation matrix can be obtained by eigendecomposition of some auxiliary B_i matrix, which can be computed solely by using observable moments. That is, eigenvectors of B_i are columns of the O matrix, and also the eigenvalues form the i'th row of O. Note that we assumed O to be invertible for the sake of simplicity. If not, one can make it invertible via singular value decomposition as in [1, 2]. The general idea is to express the observable moments so that we end up with an eigen-decomposition form, from which we can identify the parameters. Having set the notation and introduced the general idea behind the method of moments approach of [1], let us apply the same procedure on mixture of Markov models.

1.2 Derivation of spectral learning algorithm for mixture of Markov models

Similarly to the mixture model, the goal is to obtain an eigen-decomposition such as:

$$B_{i,j,k} := A(:,i,:)\text{diag}(A(j,k,:))A^{-1}(:,i,:) \tag{4}$$

So that, we can "read" the parameters from the eigenvalues or eigenvectors of $B_{i,j,k}$ matrix. Let us start with the third order observable moment of mixture of Markov models.

Claim: It is not possible to obtain the form in Equation (4) from the third and second order moments.

Proof: The goal is to obtain a form similar (if not exactly the same form) to Equation (4). We want to have potential over the hidden variables at the middle diagonal term, and the i'th column of the transition matrix for each cluster as eigenvectors. With this aim, we clamp x_2 and x_1 to particular values in third order moment.

$$p(x_3, x_2 = i, x_1 = j) = p(x_3|x_2 = i, h)\text{diag}(p(x_2 = i|x_1 = j, h))\text{diag}(\pi)p(x_1 = j|h)^T \tag{5}$$

Next, we multiply in the $I = p(x_2|x_1 = j, h)^{-1}p(x_2|x_1 = j, h)$ term as did before in the multi-view mixture model. We choose a term which involves x_1 (and consequently x_2') in order to be able to merge the introduced term and $p(x_1 = j|h)$.

$$\begin{aligned} p(x_3, x_2 = i, x_1 = j) =& p(x_3|x_2 = i, h)\text{diag}(p(x_2 = i|x_1 = j, h))I \\ & \times \text{diag}(\pi)p(x_1 = j|h)^T \\ =& p(x_3|x_2 = i, h)\text{diag}(p(x_2 = i|x_1 = j, h))p(x_2'|x_1 = j, h)^{-1} \\ & \times \underbrace{p(x_2'|x_1 = j, h)\text{diag}(\pi)p(x_1 = j|h)^T}_{p(x_2', x_1 = j)} \end{aligned}$$

However, since the term $p(x_2', x_1 = j)$ is a vector and can not be inverted, we can not recover the parameters from $p(x_3, x_2 = i, x_1 = j)$. $\square$
So, we move up to the fourth order moment.

Claim: It is possible to obtain the form in Equation (4) from the fourth and third order moments, but there is no way to resolve the permutation ambiguity over the clusters.

Proof: Let us write down the fourth order moment, with x_3 and x_2 clamped to particular values:

$$\begin{aligned} p(x_4, x_3 = i, &x_2 = j, x_1) \\ &= p(x_4|x_3 = i, h)\text{diag}(p(x_3 = i|x_2 = j, h))p(x_2 = j, x_1, h)^T \end{aligned} \tag{6}$$

Then, we insert $I = p(x_3'|x_2 = j, h)^{-1}p(x_3'|x_2 = j, h)$:

$$\begin{aligned} p(x_4, x_3 = i, &x_2 = j, x_1) \\ &= p(x_4|x_3 = i, h)\text{diag}(p(x_3 = i|x_2 = j, h))p(x_3'|x_2 = j, h)^{-1} \\ &\quad \times \underbrace{p(x_3'|x_2 = j, h)p(x_2 = j, x_1, h)^T}_{p(x_3', x_2 = i, x_1)} \end{aligned} \tag{7}$$

So we conclude that,

$$\begin{aligned} B_{i,j} =& p(x_4, x_3 = i, x_2 = j, x_1)p(x_3', x_2 = i, x_1)^{-1} \\ =& A(:, i, :)\text{diag}(A(i, j, :))A(:, j, :)^{-1} \end{aligned} \tag{8}$$

At this stage it appears that we can obtain an eigen-decomposition form by using the fourth and third order moments. In this case, the eigenvectors are $p(x_4|x_3 = i, h)$ and $p(x_3^{'}|x_2 = j, h)$ and eigenvalues are $p(x_3 = i|x_2 = j, h)$. However, we realize that following this approach there exists no way to resolve the permutation ambiguity since the ordering of the eigenvalues are arbitrary. Therefore, it is not possible for us to recover the transition matrix of a cluster since columns of an estimated transition matrix can be permuted with those of another cluster's transition matrix. $\square$

The condition to resolve this nuisance is to have the indices of the eigenvalue matrix independent of the eigenvector indices (as in Equation (4)), which can ensure the correctness in the columns of the transition matrices. For this reason we go up to the fifth order moment.

Claim: We can identify the A from the fifth and fourth order moments.

Proof: Let us write down the fifth order moment, with x_4, x_2 and x_1 clamped to particular values.

$$
\begin{aligned}
p(x_5, x_4 = i, x_3, & x_2 = j, x_1 = k) \\
&= p(x_5|x_4 = i, h)\text{diag}(p(x_2 = j|x_1 = k, h)) \\
&\quad \times \text{diag}(p(x_1 = k, h))p(x_4 = i, x_3|x_2 = j, h)^T
\end{aligned} \tag{9}
$$

Then, we insert $I = p(x_5'|x_4 = i, h)^{-1}p(x_5'|x_4 = i, h)$:

$$
\begin{aligned}
p(x_5, x_4 = i, x_3, & x_2 = j, x_1 = k) \\
&= p(x_5|x_4 = i, h)\text{diag}(p(x_2 = j|x_1 = k, h))p(x_5'|x_4 = i, h)^{-1} \\
&\quad \times \underbrace{p(x_5'|x_4 = i, h)\text{diag}(p(x_1 = k, h))p(x_4 = i, x_3|x_2 = j, h)^T}_{p(x_5', x_4 = i, x_3, x_2 = j)}
\end{aligned} \tag{10}
$$

where, we have assumed that stationarity of the Markov chain so that $p(x_1) = p(x_2)$. So we conclude that,

$$
\begin{aligned}
B_{i,j,k} &= p(x_5, x_4 = i, x_3, x_2 = j, x_1 = k)p(x_5', x_4 = i, x_3, x_2 = j)^{-1} \\
&= A(:, i, :)\text{diag}(A(j, k, :))A^{-1}(:, i, :)
\end{aligned} \tag{11}
$$

which is exactly the form we are looking for. Now, the good news is, for differing x_4 we have the same eigenvalues if we use the same x_2 and x_1. Therefore, we can use the eigenvalue-eigenvector correspondence to ensure the consistency in the columns of the transition matrix estimates (In practice, eigenvalues for every $B(i, j, k)$ may not be the same, but one can order the eigenvectors according to the eigenvalue ordering.) $\square$

However, the downside of it is that we have to use the fifth order moment. To have accurate estimates of moments this order, we need large number of samples. In this respect, it is simply not practical to use this method, as in Equation (11) with the fifth order moment. In the next section, we propose an alternative scheme for learning mixture of Markov models, to reduce the sample complexity.

2 Spectral Learning for a Mixture of Dirichlet Distributions

Claim: The parameters of each cluster can be estimated using the following relationship between the model parameters and observable moments:

$$m := \mathbb{E}[s]; \quad t^l \in \mathbb{R}^{L^2}, \quad t^l_i := \mathbb{E}[s_i^2] - \frac{1}{\alpha_0+1}\mathbb{E}[s_i] \quad if \quad i = l, \quad t^l_i := \mathbb{E}[s_l s_i] \quad if \quad i \neq l$$

$$M_2 := \mathbb{E}[s \otimes s] - \frac{1}{\alpha_0+1}\text{diag}(m)$$

$$M_3 := \mathbb{E}[s \otimes s \otimes s] - \frac{1}{\alpha_0+2}\left(\sum_{l=1}^{L^2}(e_l \otimes e_l \otimes t^l) + (e_l \otimes t^l \otimes e_l) + (t^l \otimes e_l \otimes e_l)\right) - \frac{2}{(\alpha_0+1)(\alpha_0+2)}\left(\sum_{l=1}^{L^2} m_l(e_l \otimes e_l \otimes e_l)\right)$$

then,

$$M_2 = \frac{1}{\alpha_0(\alpha_0+1)}\sum_{k=1}^{K}\pi_k\,(\alpha_k \otimes \alpha_k)$$

$$M_3 = \frac{1}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\sum_{k=1}^{K}\pi_k\,(\alpha_k \otimes \alpha_k \otimes \alpha_k)$$

where, $\alpha_0 = \sum_{k=1}^{K}\alpha_k$, e_l is the canonical basis for $\mathbb{R}^{L^2}$ and $\otimes$ is the outer product operator.

Proof: Let us start with M_2:

$$\begin{aligned}
\mathbb{E}[s \otimes s] =& \mathbb{E}[\mathbb{E}[s \otimes s|h]] = \sum_{k=1}^{K}\pi_k\mathbb{E}[s \otimes s|h=k] \\
=& \sum_{k=1}^{K}\pi_k\left(\frac{\alpha_k \otimes \alpha_k + \text{diag}(\alpha_k)}{\alpha_0(\alpha_0+1)}\right) = \frac{1}{\alpha_0(\alpha_0+1)}\sum_{k=1}^{K}\pi_k\,(\alpha_k \otimes \alpha_k) + \sum_{k=1}^{K}\pi_k\left(\frac{\text{diag}(\alpha_k)}{\alpha_0(\alpha_0+1)}\right) \\
=& \frac{1}{\alpha_0(\alpha_0+1)}\sum_{k=1}^{K}\pi_k\,(\alpha_k \otimes \alpha_k) + \left(\frac{1}{\alpha_0+1}\text{diag}\left(\sum_{k=1}^{K}\pi_k\alpha_k/\alpha_0\right)\right) \\
=& \frac{1}{\alpha_0(\alpha_0+1)}\sum_{k=1}^{K}\pi_k\,(\alpha_k \otimes \alpha_k) + \left(\frac{1}{\alpha_0+1}\text{diag}\left(\mathbb{E}[s]\right)\right) \\
=& \frac{1}{\alpha_0(\alpha_0+1)}\sum_{k=1}^{K}\pi_k\,(\alpha_k \otimes \alpha_k) + \left(\frac{1}{\alpha_0+1}\text{diag}\left(m\right)\right)
\end{aligned}$$

where we have used the facts that all clusters share the same precision parameter α_0, and $m = \mathbb{E}[s] = \mathbb{E}[\mathbb{E}[s|h]] = \mathbb{E}[\alpha_k/\alpha_0] = \sum_{k=1}^{K} \pi_k \alpha_k/\alpha_0$. So, we conclude that $\mathbb{E}[s \otimes s] - \left(\frac{1}{\alpha_0+1}\text{diag}\,(m)\right) = M_2$. For M_3, we also take the same approach:

$$\mathbb{E}[s \otimes s \otimes s] = \mathbb{E}[\mathbb{E}[s \otimes s \otimes s|h]] = \sum_{k=1}^{K} \pi_k \mathbb{E}[s \otimes s \otimes s|h = k]$$

$$= \frac{1}{\alpha_0(\alpha_0+1)(\alpha_0+2)} \sum_{k_1}^{K} \pi_k \Bigg(\alpha_k \otimes \alpha_k \otimes \alpha_k + \sum_{l=1}^{L^2} \alpha_{l,k}(\alpha_k \otimes e_l \otimes e_l)$$

$$+ \sum_{l=1}^{L^2} \alpha_{l,k}(e_l \otimes \alpha_k \otimes e_l) + \sum_{l=1}^{L^2} \alpha_{l,k}(e_l \otimes e_l \otimes \alpha_k) + 2\sum_{l=1}^{L^2} \alpha_{l,k}(e_l \otimes e_l \otimes e_l) \Bigg)$$

$$= \frac{1}{\alpha_0(\alpha_0+1)(\alpha_0+2)} \sum_{k_1}^{K} \pi_k \Bigg(\alpha_k \otimes \alpha_k \otimes \alpha_k \Bigg) + \frac{1}{\alpha_0+2} \sum_{k_1}^{K} \pi_k \Bigg(\sum_{l=1}^{L^2} \frac{\alpha_{l,k}}{\alpha_0(\alpha_0+1)}(\alpha_k \otimes e_l \otimes e_l)$$

$$+ \sum_{l=1}^{L^2} \frac{\alpha_{l,k}}{\alpha_0(\alpha_0+1)}(e_l \otimes \alpha_k \otimes e_l) + \sum_{l=1}^{L^2} \frac{\alpha_{l,k}}{\alpha_0(\alpha_0+1)}(e_l \otimes e_l \otimes \alpha_k) + 2\sum_{l=1}^{L^2} \frac{\alpha_{l,k}}{\alpha_0(\alpha_0+1)}(e_l \otimes e_l \otimes e_l) \Bigg)$$

Then, we observe that the term $\sum_{k=1}^{K} \pi_k \alpha_{l,k}(\alpha_k \otimes e_l \otimes e_l) = (t^l \otimes e_l \otimes e_l)$, where we note that $\sum_{k=1}^{K} \pi_k \alpha_{l,k}^2/\alpha_0(\alpha_0+1) = \mathbb{E}[s_l^2] - \frac{1}{\alpha_0+1}\mathbb{E}[s_l]$, and $\sum_{k=1}^{K} \pi_k \alpha_{l,k}\alpha_{i,k}/\alpha_0(\alpha_0+1) = \mathbb{E}[s_l s_i]$, for $i \neq l$. So, the whole thing becomes;

$$\mathbb{E}[s \otimes s \otimes s] =$$

$$\frac{1}{\alpha_0(\alpha_0+1)(\alpha_0+2)} \sum_{k_1}^{K} \pi_k \Bigg(\alpha_k \otimes \alpha_k \otimes \alpha_k \Bigg) + \frac{1}{\alpha_0+2} \Bigg(\sum_{l=1}^{L^2} (t^l \otimes e_l \otimes e_l) + (e_l \otimes t^l \otimes e_l)$$

$$+ (e_l \otimes e_l \otimes t^l) \Bigg) + \frac{2}{(\alpha_0+1)(\alpha_0+2)} \sum_{l=1}^{L^2} m_l (e_l \otimes e_l \otimes e_l)$$

So, we conclude that we obtain M_3 by subtracting all the terms but the first term from $\mathbb{E}[s \otimes s \otimes s]$. Note that the terms to be subtracted can be computed in terms of observable moments. This concludes the proof. □

References

[1] Anandkumar, A., D. Hsu and S. Kakade, "A Method of Moments for Mixture Models and Hidden Markov Models", *COLT*, 2012.

[2] Hsu, D., S. M. Kakade and T. Zhang, "A Spectral Algorithm for Learning Hidden Markov Models A Spectral Algorithm for Learning Hidden Markov Models", *Journal of Computer and System Sciences*, , No. 1460-1480, 2009.