Spectral Learning for Mixture of Markov Models

Y. Cem Sübakan[1], Barış Kurt[2], A. Taylan Cemgil[2], Bülent Sankur[3]

[1] University of Illinois at Urbana-Champaign, Computer Science
[2] Boğaziçi University, Computer Engineering
[3] Boğaziçi University, Electrical & Electronics Engineering

Abstract

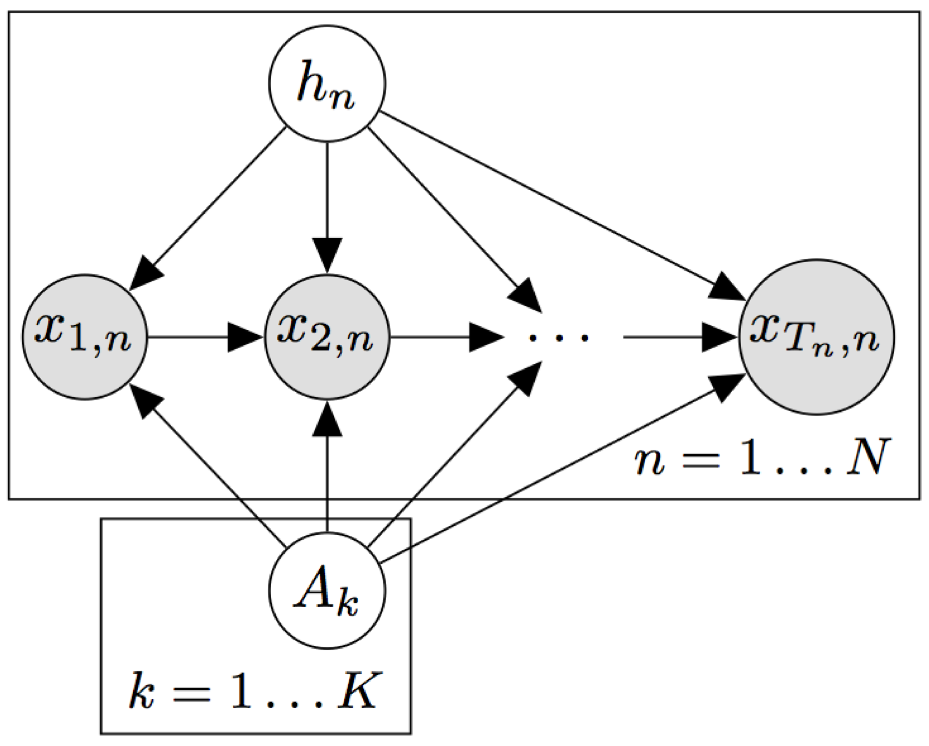
Problem statement: *Sequence clustering by spectral learning for Mixture of Markov models.*

Contribution: *Regular spectral learning algorithms for latent variable models require fifth order moment. We reduce the sample complexity by learning mixture of Dirichlet distributions.*

Conclusion: *We experimentally show the superiority of our approach compared to EM and regular spectral learning.*

1. Mixture of Markov Models

The graphical model is as follows:



- Observation at sequence n, time t: $x_{t,n}$.
- Cluster indicator of sequence n: h_n.
- Transition matrix of cluster k: A_k.

The likelihood of observing a sequence, $\mathbf{x}_n = (x_{1,n}, x_{2,n}, \dots, x_{T_n,n})$ of length T_n is defined as:

$$p(x_n|A_{1:K}) = \sum_{k=1}^{K} p(h_n = k) \prod_{t=1}^{T_n} p(x_{t,n}|x_{t-1,n}, h_n = k)$$
$$= \sum_{k=1}^{K} \pi_k \prod_{t=1}^{T_n} \prod_{l_1=1}^{L} \prod_{l_2=1}^{L} A_{k,l_1,l_2}^{[x_{t,n}=l_1][x_{t-1,n}=l_2]}$$

- The ultimate learning goal is to estimate the cluster assignments $h_{1:N}$ given sequences $\mathbf{x}_{1:N}$.

2. Spectral Learning of Mixture of Markov Models

Learning Strategy: Learn transition matrices $A_{1:K}$ given sequences $\mathbf{x}_{1:N}$. Then, learn the assignments $h_{1:N}$.

Claim: The model parameters $A_{1:K}$ can be uniquely identified using fifth and fourth order moments:

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

Drawback: This spectral learning approach requires moments up-to order five.

3. Spectral Learning of Mixture of Dirichlet Distributions

Learning Strategy: Alternatively, one can learn a mixture of posterior distributions of transition matrices, which is mixture of Dirichlet distributions, to estimate $h_{1:N}$.

Using conjugate Dirichlet prior, the posterior is also Dirichlet:

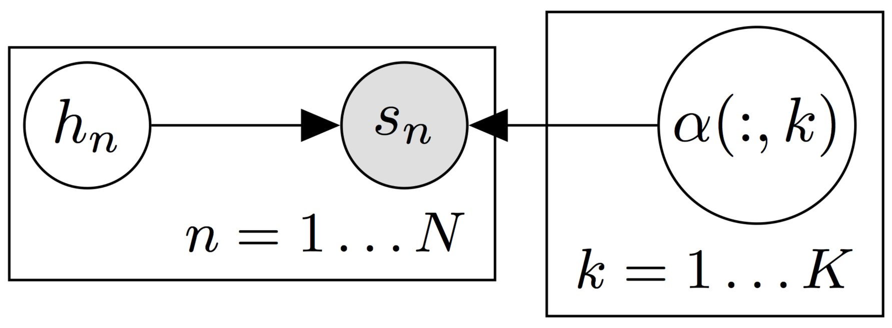
$$p(A_{h_n}|\mathbf{x}_n, h_n) \propto \quad p(\mathbf{x}_n|A_{h_n}, h_n)p(A_{h_n})$$
$$\propto \left(\prod_{t=1}^{T_n} \prod_{l_1=1}^{L} \prod_{l_2=1}^{L} A_{h_n,l_1,l_2}^{[x_{t,n}=l_1][x_{t-1,n}=l_2]} \right) \prod_{l_1=1}^{L} \prod_{l_2=1}^{L} A_{h_n,l_1,l_2}^{\beta-1}$$
$$\propto \prod_{l_1=1}^{L} \prod_{l_2=1}^{L} A_{h_n,l_1,l_2}^{c^n_{l_1,l_2}+\beta-1}$$
$$= \text{Dirichlet}(c^n_{1,1} + \beta - 1, c^n_{1,2} + \beta - 1, \dots, c^n_{L,L} + \beta - 1)$$

where, $c^n_{l_1,l_2}$ stores the state transition counts of $\mathbf{x}_n$.

Setting $\beta = 1$ (having a uniform prior), the posterior distribution becomes; $p(A_{h_n}|\mathbf{x}_n, h_n) =$ Dirichlet$(c^n_{1,1}, c^n_{1,2}, \dots, c^n_{L,L})$.

We treat the normalized state transition count matrices $s^n_{l_1,l_2} = c^n_{l_1,l_2} / \sum_{l_1,l_2} c^n_{l_1,l_2}$ as a sample from the posterior of the transition matrix.

So, the graphical model becomes:



- Observations, normalized transition counts: s_n.
- Dirichlet parameters of cluster k: $\alpha(:, k) = \alpha_k$.

Claim: The posterior parameters of each cluster and observable moments are related as the following:

$$m := \mathbb{E}[s]; \quad t^l \in \mathbb{R}^{L^2}, \quad t^l_i := \mathbb{E}[s_i^2] - \frac{1}{\alpha_0 + 1}\mathbb{E}[s_i] \quad if \quad i = l,$$
$$t^l_i := \mathbb{E}[s_l s_i] \quad if \quad i \neq l$$
$$M_2 := \mathbb{E}[s \otimes s] - \frac{1}{\alpha_0 + 1}\text{diag}(m)$$
$$M_3 := \mathbb{E}[s \otimes s \otimes s] - \frac{1}{\alpha_0 + 2}\left(\sum_{l=1}^{L^2} (e_l \otimes e_l \otimes t^l) + (e_l \otimes t^l \otimes e_l) + (t^l \otimes e_l \otimes e_l) \right) - \frac{2}{(\alpha_0 + 1)(\alpha_0 + 2)} \left(\sum_{l=1}^{L^2} m_l (e_l \otimes e_l \otimes e_l) \right)$$

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.

Having the parameters in the symmetric tensor form, we can apply the existing spectral learning procedures to estimate $\alpha_{1:K}$.

4. Experimental Results

- We generated 100 data sets. Each set is composed of 60 sequences, with $K = 3$.
- The prior cluster probabilities $p(h_n)$ and transition matrices $A_{1:3}$ are generated randomly.

First Experiment:

- Comparison of clustering accuracies of the spectral learning and EM algorithms (4 algorithms).
- Results for varying sequence lengths are shown in Figure 3.

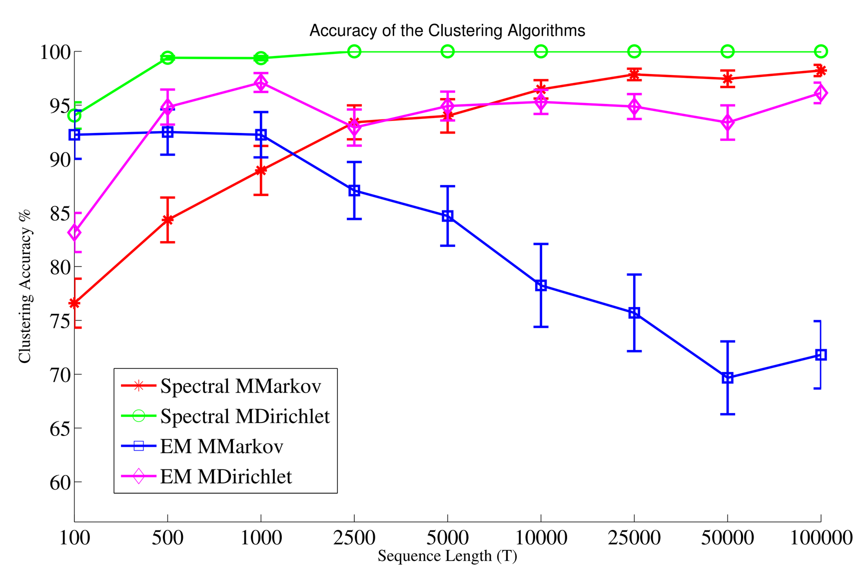


Figure 3: *Comparison of clustering accuracies on synthetic data for differing sequence lengths*

- Mixture of Dirichlet distributions yield the highest clustering accuracies for all sequence lengths.
- Given sufficient data, spectral algorithms give higher clustering accuracies compared to their EM counterparts.

Second Experiment:

- We next investigate the effect of changing L (cardinality of observations). Results for differing L are given in Figure 4.

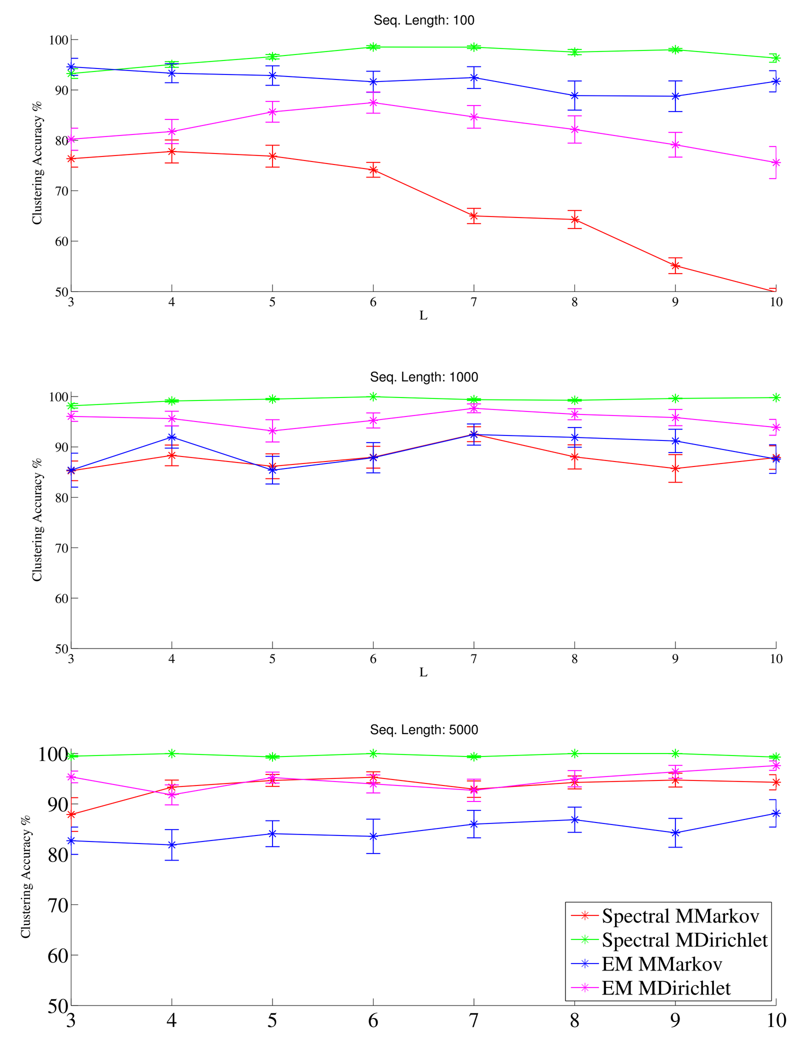


Figure 4: *Effect of changing L on clustering accuracy*

- Number of states L has the least significant effect on spectral mixture of Dirichlet algorithm.
- In experiments with short sequences, the spectral learning for mixture of Markov models is the most sensitive algorithm to increasing L.
- All algorithms become less sensitive to L as sequence length increases.

Third Experiment:

- Next, we investigate cluster similarity on clustering accuracy.
- For $K = 3$, transition matrices are generated as $A_k = (1 - \lambda)\tilde{A}_0 + \lambda\tilde{A}_k$, where $\tilde{A}_0, \tilde{A}_1, \tilde{A}_2, \tilde{A}_3 \sim$ Dirichlet(1,...,1). Results for differing λ are given in Figure 5.

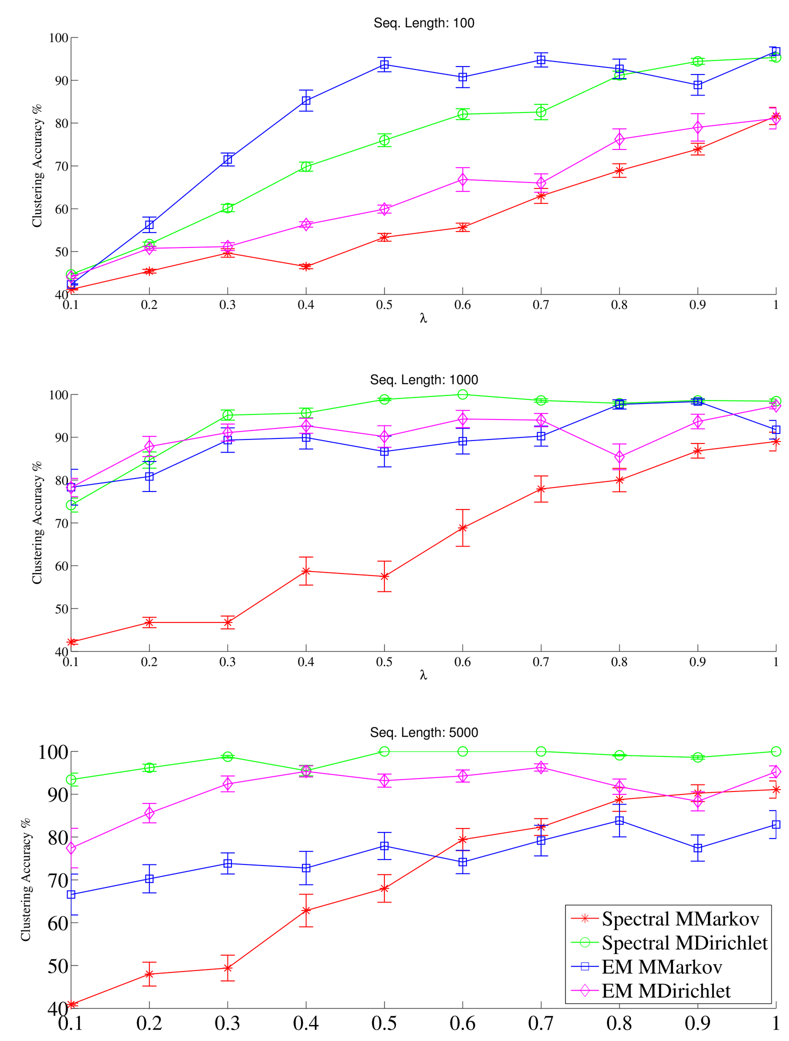


Figure 5: *Effect of cluster similarity on clustering accuracy*

- If there is enough data available, mixture of Dirichlet algorithm yields high accuracy, even when $\lambda = 0.1$.

5. Conclusions

- **Conclusion:** Experimental results suggests that proposed method outperforms EM and regular spectral learning approach in several regimes.
- **Future Work:** Application of the algorithm on real-world applications.