Spectral Learning of Hidden Markov Models with Group Persistence

Abstract

In this paper, we develop a general Method of Moments (MoM) based parameter estimation framework for Switching Hidden Markov Model (SHMM) variants. The main obstacle for deriving a straightforward MoM algorithm for these models is the inherent permutation ambiguity in the parameter estimation, which causes the parameters of individual HMM groups to get mixed. We show that, as long as a global transition matrix has a group persistence property, it is possible to isolate the group parameters using a spectral de-permutation approach. We also provide a noise bound on the eigenvalues of the recovered transition matrix. We do experiments on synthetic data which shows the accuracy and the computational advantage of the proposed approach. We also perform a segmentation experiment on saxophone note sequences.

1. Introduction

Method of moments (MoM) based learning algorithms have been popular in machine learning community due to their computational advantages and uniqueness guarantees for parameter learning in latent variable models (LVMs). In the original line of work (Anandkumar et al., 2012b;a; Hsu & Kakade, 2013), the parameter learning is only applicable to few basic models such as mixture models and HMMs, and it is unclear as to how to derive a MoM algorithm for more complicated models.

Recently in (Subakan et al., 2014), a MoM algorithm for Mixture of Hidden Markov Models (Smyth, 1997) (MHMM) is proposed. The main idea is to reduce the MHMM learning to an HMM learning with a larger state space, since it can be easily seen that an MHMM is an HMM with a larger state space which has a block diagonal transition matrix. Therefore, it is possible to use a MoM based HMM learning procedure to estimate the model parameters up to a permutation ambiguity, which is a fundamental nuisance for MHMM learning since it causes the parameters of the individual clusters to get mixed. However, due to the extreme block-diagonal nature of the transition matrix it is possible to de-permute to recover a block diagonal structure, given that the noise from the parameter estimation step is not too strong (Subakan et al., 2014).

In this paper, we extend the de-permutation idea introduced in (Subakan et al., 2014) to a more general class of HMMs which does not necessarily have a complete block-diagonal transition matrix. Group persistence is defined as the tendency of a Markov chain to stay within a subset of states, rather than switching to another subset. We show that as long as a global transition matrix possesses a group persistence property (or anti group persistence), it is possible to de-permute it.

We use the switching HMM (SHMM) (Murphy, 2002) as the most general case, since it is possible to adjust the level of group persistence with a group transition matrix. Also, by setting certain parameters of the SHMM, it is possible to obtain useful models used in practice such as an HMM with mixture observations (Rabiner, 1989) or MHMM.

In this paper, we also propose a methodology to exactly recover the specific structure of the transition matrix. Once the permutation mapping is estimated, since we recover the emission and transition matrices separately as proposed in (Chaganty & Liang, 2014), we are able to formulate a convex optimization problem with constraints that enforce a specific transition matrix structure.

The organization of the paper is as follows: In Section 2, we give the definitions of the models used in the paper. In Section 3 we describe the specifics of the learning procedure. In Section 4 we provide experiments on synthetic and real data.

2. The Setup

2.1. On Notation

Throughout the paper, we use the MATLAB notation $A(:, i)$, which takes the i'th column of a matrix A. We do not use boldface to distinguish between matrices and scalars, since it should be clear from the context. $1_M \in \mathbb{R}^M$ denotes an all-ones vector of length M. We use subscripts to denote matrix blocks: $A_{i,j}$ denotes the (i, j)'th block taken from the A matrix. $A_{i,j}(k, l)$ denotes the

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(k,l)'th entry of block (i,j). We use the $\otimes$ symbol to denote outer product such that, $(a \otimes b)_{ij} = a_i b_j$ (unless otherwise noted).

2.2. Hidden Markov Model

The generative model of an Hidden Markov Model (HMM) is defined as follows:

$$\begin{aligned} r_1 &\sim \mathcal{D}(\nu), \\ r_t | r_{t-1} &\sim \mathcal{D}(A(:, r_{t-1})),\ \forall t \in \{2, \dots, T\} \\ x_t | r_t &\sim p(x_t | r_t),\ \forall t \in \{1, \dots, T\}, \end{aligned}$$

where $\mathcal{D}(\nu)$ denotes a discrete distribution with weights ν. The HMM is parametrized by $\theta = (O, A, \nu)$, which are respectively the emission matrix, transition matrix, and the initial state distribution. An observation sequence is denoted with $x_{1:T} = \{x_1, x_2, \dots, x_T\}$, and each observation $x_t \in \mathbb{R}^L$ is generated from the emission density $p(x_t|r_t)$, which is parametrized by the emission matrix $O \in \mathbb{R}^{L \times M}$. The columns of the emission matrix corresponds to the parameters of the emission distribution. E.g., in the discrete case, the columns correspond to the emission probability of the symbols, or in the spherical fixed variance Gaussian case the columns correspond the means of the Gaussians in each dimension.

Latent states $r_{1:T}$ form a Markov chain with the transition matrix $A \in \mathbb{R}^{M \times M}$, and the initial state distribution $\nu \in \mathbb{R}^M$. Finally note that the HMM full joint distribution is defined as follows:

$$p(x_{1:T}, r_{1:T}) = p(r_1) \prod_{t=2}^{T} p(x_t | r_t) p(r_t | r_{t-1}). \tag{1}$$

2.3. The general case: Switching HMM

We use the Switching State Space HMM (SHMM) (Murphy, 2002) as the general case from which we derive the other models as special cases. It is parametrized by parameters $\theta_{1:K} = (O_{1:K}, A_{1:K,1:K}, \nu_{1:K}, B, \pi)$. The generative model is specified as follows,

$$\begin{aligned} &h_1 \sim \mathcal{D}(\pi),\ r_1 | h_1 \sim \mathcal{D}(\nu_{h_1}), \\ &h_t | h_{t-1} \sim \mathcal{D}(B(:, h_{t-1}))\ \forall t \in \{2, \dots T\}, \\ &r_t | r_{t-1}, h_t, h_{t-1} \sim [h_t = h_{t-1}] \mathcal{D}(A_{h_t, h_t}(:, r_{t-1})) + \dots \\ &\qquad \dots [h_t \neq h_{t-1}] \mathcal{D}(A_{h_{t-1}, h_t}(:, r_{t-1}))\ \forall t \in \{2, \dots T\}, \\ &x_t | h_t, r_t \sim p(x_t | h_t, r_t), \end{aligned}$$

where, $\mathcal{D}(\pi)$ denotes a discrete distribution with weights specified in the π vector, $h_t \in \{1, \dots, K\}$ is the latent *regime* indicator, and $r_t \in \{1, \dots, M\}$ is the latent state indicator. Model parameters are as follows: $B \in \mathbb{R}^{K \times K}$ is the *regime transition matrix*, and $A_{h_t, h_{t-1}} \in \mathbb{R}^{M \times M}$ is the state transition matrix which corresponds to the regime indicators h_t, h_{t-1}. If $h_t = h_{t-1}$, then the *intra-regime* transition matrix A_{h_t, h_t} is used and if $h_t \neq h_{t-1}$, then the *inter-regime* transition matrix A_{h_{t-1}, h_t} is used. Finally, $\pi \in \mathbb{R}^K$ is the *initial regime distribution* and $\nu_{h_1} \in \mathbb{R}^M$ is the initial state distribution which corresponds to the initial regime h_1.

2.3.1. Special Cases

By setting certain parameters of the SHMM we can obtain several models which are useful in practice. The graphical models of these models are given in Table 1.

HMM with a Mixture Observation Model:
For $A_{i,j} = \nu_i 1_M^\top$, $\forall\ (i,j)$, SHMM reduces to an HMM with a mixture observation model (HMM-M) (Murphy, 2002; Rabiner, 1989).

Mixture of HMMs:
For $B = I$, and $A_{i,j} = 0$ for $i \neq j$, SHMM reduces to an MHMM (Smyth, 1997).

Mixture of Mixtures:
For $B = I$, $A_{i,j} = \mathbf{0}$, $\forall\ i \neq j$, and $A_{i,i} = \nu_i 1_M^\top$, SHMM reduces to a Mixture of Mixtures (MM). Note that in this model a sequence is generated using a single mixture model.

Uniform Transition Switching HMM: (USHMM)
For $A_{i,j} = \frac{1}{M} 1_M 1_M^\top, \forall\ i \neq j$, we get a practical switching HMM with uniform transitions when a regime change occurs. We use this SHMM in the experiment in Section 4.1.

Note: MHMM and MM are *sequence clustering* models, and therefore we observe more than one sequence. SHMM and HMM-M are *sequence segmentation* models and they are typically used to segment a single sequence.

2.4. Global HMM Interpretation

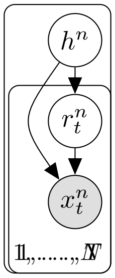
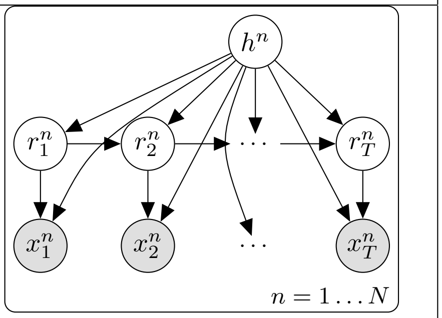
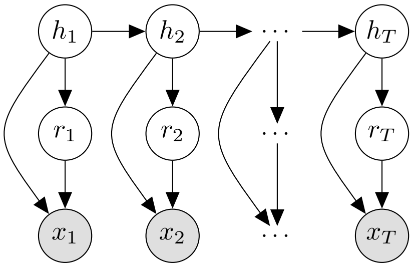
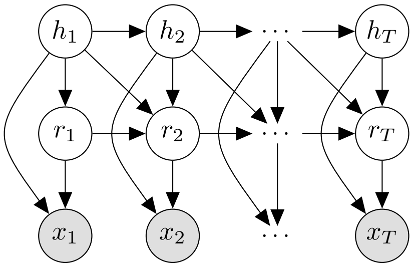
Lemma 1: Let $\theta_{1:K} = (O_{1:K}, A_{1:K,1:K}, \nu_{1:K}, B, \pi)$ be the parameters describing a general HMM with local components. The general HMM is equivalent to a standard HMM with global parameters $\theta = (O, A, \nu)$, where

$$\begin{aligned} O &= \begin{bmatrix} O_1 & O_2 & \dots & O_K \end{bmatrix}, \\ \nu &= \begin{bmatrix} \pi_1 \nu_1^\top & \pi_2 \nu_2^\top & \dots & \pi_K \nu_K^\top \end{bmatrix}^\top, \\ A &= \begin{bmatrix} B_{1,1} A_{1,1} & B_{1,2} A_{1,2} & \dots & B_{1,K} A_{1,K} \\ B_{2,1} A_{2,1} & B_{2,2} A_{2,2} & \dots & B_{2,K} A_{2,K} \\ & & \ddots & \\ B_{K,1} A_{K,1} & B_{K,2} A_{K,2} & \dots & B_{K,K} A_{K,K} \end{bmatrix}, \end{aligned} \tag{2}$$

where $O \in \mathbb{R}^{L \times (MK)}$, $A \in \mathbb{R}^{MK \times MK}$, and $\nu \in \mathbb{R}^{MK \times 1}$.

Proof:

Table 1. Directed Graphs of HMMs in Section 2.3.1

Mixture of Mixtures	Mixture of HMMs
	 h^n r_1^n r_2^n $\cdots$ r_T^n x_1^n x_2^n $\cdots$ x_T^n $n = 1 \dots N$
Hidden Markov Model with Mixture Observations	**Switching HMM**
 h_1 h_2 $\cdots$ h_T r_1 r_2 $\cdots$ r_T x_1 x_2 $\cdots$ x_T	 h_1 h_2 $\cdots$ h_T r_1 r_2 $\cdots$ r_T x_1 x_2 $\cdots$ x_T

The full joint distribution for SHMM is defined as follows:

$$\begin{aligned}&p(x_{1:T}, r_{1:T}, h_{1:T})\\ =&p(r_1|h_1)p(h_1)\prod_{t=1}^{T} p(x_t|r_t,h_t)p(r_t|r_{t-1},h_{t-1})p(h_t|h_{t-1}),\\ =&p(r_1,h_1)\prod_{t=2}^{T} p(x_t|r_t,h_t)p(r_t,h_t|r_{t-1},h_{t-1}).\end{aligned}$$

At this point, we see that this expression is same as the HMM joint distribution expression in Equation (1), if we define a new variable $rh_t : (r_t \otimes h_t)$, which is defined on the product space of r_t and h_t. Therefore, SHMM is equivalent to an HMM with MK states. □

It is interesting to note that with the global HMM formulation the special cases given in Section 2.3.1 reduce to HMMs with a particular transition matrix structure. The transition matrices of these special cases are specified in Table 2.

3. Learning

3.1. Estimation of the Emission Matrix

Given that few assumptions hold, it is possible to learn the global emission matrix O (up-to a permutation) ambiguity, by learning a mixture model (Kontorovich et al., 2013).

Assumption 1: (Accessibility for SHMM and HMM-M) The global transition matrix A has a stationary distribution $\xi \succ 0$ (all elements of the stationary distribution are greater than zero) and the sequence length T is large enough to observe from all columns of the global emission matrix $O = [O_1, O_2, \dots, O_k]$.

Assumption 2: (Accessibility for MHMM and MM) All the mixing weights are greater then zero $\pi \succ 0$, and the transition matrix A_k has a stationary distribution $\xi_k \succ 0$. Furthermore, the number of sequences N is large enough and the individual sequences are long enough to observe from all columns of the global emission matrix O.

Assumption 3: (Non-Degeneracy) $L \geq K$ and the global emission matrix O has full column rank.

Theorem 1: Given that the assumption 1 and assumption 3 hold for SHMM and HMM-M (2 and 3 for MHMM and MM), a MoM algorithm (e.g. Tensor Power method) can recover the global transition matrix O, and the stationary distribution ξ up-to a permutation $\mathcal{P}$ of the columns, and elements, respectively.

Proof: For SHMM and HMM-M, since the global transition matrix A defines a fully connected Markov chain, after a burn-in period the Markov chain converges to some stationary distribution ξ, and the model can be seen as a mixture distribution with mixing weights ξ. Therefore, one can estimate O and ξ with the Tensor power method, using the second and third order moments $\mathbb{E}[x_2 x_1^\top]$, $\mathbb{E}[x_3 \otimes x_2 \otimes x_1]$ by treating the data as if it was i.i.d.. For MHMM and MM, the model can be seen as a simple mixture model with mix-

Table 2. Transition matrix structures for several HMM variants

Mixture of Mixtures	**Mixture of HMMs**
$\begin{bmatrix} \nu_1 \mathbf{1}_M^\top & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \nu_2 \mathbf{1}_M^\top & \dots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & \nu_K \mathbf{1}_M^\top \end{bmatrix}$	$\begin{bmatrix} A_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & A_2 & \dots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & A_K \end{bmatrix}$
Hidden Markov Model with Mixture Observations	**Practical SHMM**
$\begin{bmatrix} B_{1,1}\ \nu_1 \mathbf{1}_M^\top & B_{1,2}\ \nu_1 \mathbf{1}_M^\top & \dots & B_{1,K}\ \nu_1 \mathbf{1}_M^\top \\ B_{2,1}\ \nu_2 \mathbf{1}_M^\top & B_{2,2}\ \nu_2 \mathbf{1}_M^\top & \dots & B_{2,K}\ \nu_2 \mathbf{1}_M^\top \\ & & \ddots & \\ B_{K,1}\ \nu_K \mathbf{1}_M^\top & B_{K,2}\ \nu_K \mathbf{1}_M^\top & \dots & B_{K,K}\ \nu_K \mathbf{1}_M^\top \end{bmatrix}$	$\begin{bmatrix} B_{1,1}\ A_{1,1} & B_{1,2}\ \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top & \dots & B_{1,K}\ \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top \\ B_{2,1}\ \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top & B_{2,2}\ A_{2,2} & \dots & B_{2,K}\ \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top \\ & & \ddots & \\ B_{K,1}\ \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top & B_{K,2}\ \frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top & \dots & B_{K,K}\ A_{K,K} \end{bmatrix}$

ing weights $\xi = [\pi_1 \xi_1, \pi_2 \xi_2, \dots, \pi_K \xi_K]$. □

Note that, if the observation model is spherical Gaussian then, one can use the single view moments $\mathbb{E}[x_t x_t^\top]$ and $\mathbb{E}[x_t \otimes x_t \otimes x_t]$ (Hsu & Kakade, 2013). In fact, this is true for all observation models provided that the moment equations can be worked out.

3.2. Estimation of the Transition Matrix

Lemma 1: For models in Section 2.3.1, second order temporal statistics of the global HMM factorizes as follows,

$$\mathbb{E}[x_{t+1} x_t^\top] = O A \text{diag}(\xi) O^\top \tag{3}$$

where, O and A are respectively the global emission matrix and the global transition matrix and $\xi \in \mathbb{R}^{KM}$ is the stationary distribution of the global HMM for SHMM and HMM-M, and $[\pi_1 \xi_1, \pi_2 \xi_2, \dots \pi_K \xi_K]$ vector for MHMM and MM.

Proof: It is easy to see that the moment expression holds for SHMM and HMM-M by considering that they have a global HMM interpretation (The proof for the second order moment of an HMM is in (Anandkumar et al., 2012b)). For MHMM and MM we see that the moment expression factorizes as,

$$\begin{aligned} S_{2,1} := \mathbb{E}[x_{t+1} x_t^\top] =& \sum_{k=1}^{K} \pi_k O_k A_k \text{diag}(\xi_k) O_k, \\ =& O A \text{diag}([\pi_1 \xi_1, \dots, \pi_K \xi_K]) O, \\ =& O A \text{diag}(\xi) O. \end{aligned} \tag{4}$$

This concludes the proof. □

Having observed that the moment expression factorizes as indicated above, given an estimate $\widehat{O}$ for the emission matrix, an estimate $\widehat{\xi}$ for the stationary distribution, and an empirical estimate $\widehat{S_{2,1}}$ for the second order moment, an estimator $\widehat{A}$ can be derived by using the pseudo inverse of $\widehat{O}$ and the reciprocal of $\widehat{\xi}$:

$$\widehat{A} = \widehat{O}^\dagger \widehat{S_{2,1}} \text{diag}(1/\widehat{\xi}) (\widehat{O}^\top)^\dagger. \tag{5}$$

The drawback of this estimator is that it can result in negative entries. To alleviate this issue, we formulate a constrained optimization problem:

Lemma 2: Given that $\widehat{O} = O P^\top$, $\widehat{\xi} = \xi P^\top$ and, $\widehat{S_{2,1}} = S_{2,1}$, the solution A^* of the convex optimization problem is equal to the permuted true transition matrix $\mathcal{P}(A)$.

$$\begin{aligned} \min_{A} \quad & \| \widehat{S_{2,1}} - \widehat{O} A \text{diag}(\widehat{\xi}) \widehat{O}^\top \|_F \\ s.t. \quad & 1_{MK}^\top A = 1_{MK}^\top \\ & A \geq 0, \end{aligned} \tag{6}$$

is deterministically related to the true transition matrix A up-to the permutation $\mathcal{P}$ of the state labels.

Proof: In the case where we have the true moment $S_{2,1}$, and $\widehat{O}$ has correct columns up-to a permutation ambiguity $\mathcal{P}$, the objective function achieves zero at the solution of the problem $A^* = P A P^\top$. □

Also note the optimization problem given in Lemma 2, is convex since the constraints form a convex set and the objective function is convex (Boyd & Vandenberghe, 2004). The constraints ensure that A is a conditional probability table. Note that, it would have been possible to impose structural constraints on A to recover the exact model structures given in Table 2, if we had known the permutation $\mathcal{P}$. In Section 3.4, we propose running another convex optimization procedure to recover an exact structure.

3.3. The Depermutation

Let θ denote the true global parameters. After the parameter estimation steps given in Section 3.1 and Section 3.2, the original parameters are contaminated with an estimation noise $\epsilon : \theta \rightarrow \theta_\epsilon$, and they are permuted according to the permutation mapping $\mathcal{P} : \theta_\epsilon \rightarrow \mathcal{P}(\theta_\epsilon)$, such that the output $\mathcal{P}(\theta_\epsilon) = (\widehat{O}, \widehat{A}, \widehat{\nu})$ is,

$$\widehat{O} = O_\epsilon P^\top, \quad \widehat{A} = P A_\epsilon P^\top, \quad \widehat{\nu} = \nu_\epsilon P^\top.$$

where P is the permutation matrix which corresponds to the permutation mapping $\mathcal{P}$. The permutation $\mathcal{P}$ is a fundamental nuisance in learning HMMs with several layers such as MHMM and SHMM since it causes individual HMM parameters $\theta_{1:K}$ to get mixed. The goal in the depermutation stage is to find a depermutation mapping $\widetilde{\mathcal{P}}$: $\mathcal{P}(\theta_\epsilon) \rightarrow \widetilde{\mathcal{P}}(\mathcal{P}(\theta_\epsilon))$, such that $\widetilde{\mathcal{P}}(\widehat{A})$ is a *sufficiently block-diagonal* matrix, which would allow us to isolate individual HMM components. We use the de-permutation algorithm proposed in (Subakan et al., 2014), and our goal in this section is provide a case study for SHMM special cases and develop an insight regarding choice of the particular HMM model, and difficulty of the depermutation problem.

Definition 1: For an SHMM, the matrix $\widetilde{\mathcal{P}}(\widehat{A})$ is sufficiently block diagonal if,

$$\|\widetilde{\mathcal{P}}(\widehat{A}_{i.j}) - B_{\mathcal{P}_2(i,j)}\mathcal{P}_1(A_{\mathcal{P}_2(i,j)})\|_F \leq \gamma, \ \forall (i,j),$$

where γ is a small enough number, $\mathcal{P}_1$ and $\mathcal{P}_2$ are the depermutation mappings which respectively correspond to the permutation within the blocks $\widehat{A}_{i,j}$ and between blocks $(i,j), \forall (i,j)$.

To get a sense of the problem let us start with the cases that correspond to the models in Section 2.3.1. For simplicity we assume the estimation noise is negligible.

Case 1: (MM) As seen in Table 2, for Mixture of Mixtures (MM) model columns of the global transition which correspond to a group are identical to each other. As also can be seen in the middle column of Figure 1, a block diagonal structure can easily be obtain by simply clustering the columns of $\mathcal{P}(A)$, which is shown in the right column of Figure 1.

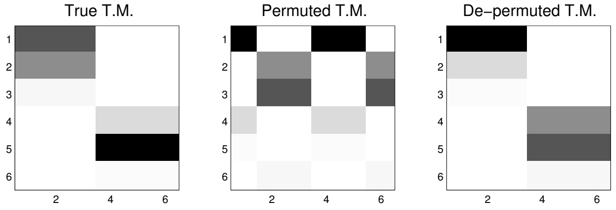


Figure 1. (left) True Global Transition Matrix A, (middle) Permuted Global Transition Matrix $\mathcal{P}(A)$, (right) De-permuted Global Transition Matrix $\widetilde{\mathcal{P}}(\mathcal{P}(A))$, for an example Mixture of Mixtures (MM) model.

Case 2: (MHMM) As can be seen from Table 2 the transition structure is not as simple as MM. However, as can be seen easily, assuming that each block $A_{k,k}$ has one eigenvalue which is one, the transition matrix $\mathcal{P}(A)$ has K eigenvalues which are one. Since the eigenvalue decomposition is invariant with respect to $\mathcal{P}$, $\mathcal{P}(A)^\infty$ reveals the same K cluster structure (Subakan et al., 2014), as the permuted MM transition matrix of the previous example. Therefore, it is possible to find a mapping $\widetilde{P}$ to make $\mathcal{P}(A)$, by clustering the columns of $\mathcal{P}(A)^\infty$.

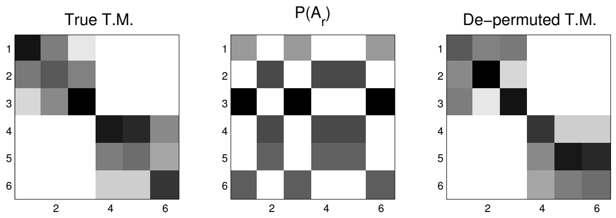


Figure 2. (left) True Global Transition Matrix A, (middle) Permuted Global Transition Matrix to the power $\mathcal{P}(A)$, (right) De-permuted Global Transition Matrix $\widetilde{\mathcal{P}}(\mathcal{P}(A))$, for an example MHMM.

Case 3: (HMM-M) Despite the fact that HMM blocks are connected, This case is easy because the individual blocks are already converged and it is easy to make $\mathcal{P}(A)$ sufficiently block diagonal by clustering the columns. This is shown in Figure 3.

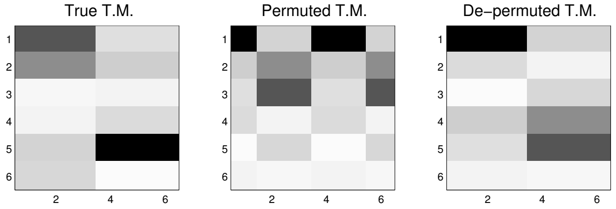


Figure 3. (left) True Global Transition Matrix A, (middle) Permuted Global Transition Matrix $\mathcal{P}(A)$ of the Permuted Global Transition Matrix $\mathcal{P}(A)$, (right) De-permuted Global Transition Matrix $\widetilde{\mathcal{P}}(\mathcal{P}(A))$, for an example HMM-M.

Case 4: (SHMM) SHMM corresponds to the most difficult cases because of the presence of off-diagonal blocks in the transition matrix and the fact that individual blocks are not converged as HMM-M. In the case, where all blocks are non-zero, the Markov chains becomes fully connected and $\mathcal{P}(A)^\infty$ converges to a stationary distribution, as shown in Figure 4. However, if the eigenvalue ordering is preserved, it is possible to recover the K cluster structure using a low rank reconstruction.

Lemma 3: *K eigenvalues of an SHMM global transition matrix are same as the eigenvalues of its corresponding regime transition matrix B.*

Proof: See appendix.

Observation: If $|\lambda_K(A)| = |\lambda_K(B)|$, then there exists a

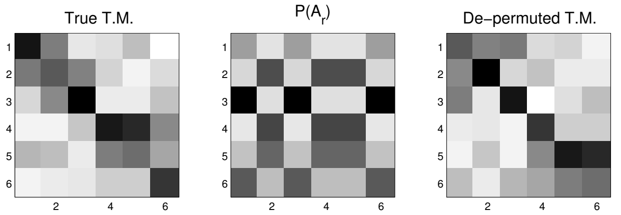


Figure 4. (left) True Global Transition Matrix A, (middle) Low rank reconstruction $\mathcal{P}(A_r)$ of the Permuted Global Transition Matrix $\mathcal{P}(A)$, (right) De-permuted Global Transition Matrix $\widetilde{\mathcal{P}}(\mathcal{P}(A))$, for an example SHMM.

matrix $A_r = V\bar{\Lambda}V^{-1}$ such that,

$$\sum_{i=1}^{KM} \|A_r(:,i) - C(:,h_i)\|_2 = 0 \tag{7}$$

where, V is the eigenvector matrix, $\bar{\Lambda}$ is a diagonal matrix with only the first K largest eigenvalues being non-zero, and C is the cluster centroid matrix, with h_i being the cluster indicator of the i'th column.

So, we conclude that if B has group persistence (or anti-persistence) with large $|\lambda_K(B)|$ it is much more likely that eigenvalues of B will be the top K eigenvalues and therefore, a permutation mapping $\widetilde{\mathcal{P}}$ is recoverable by clustering the columns of the rank-K reconstruction A_r to make $\mathcal{P}(A)$ sufficiently block diagonal, with $\gamma = 0$ in Definition 1. Now, we will look at the case where the estimation noise ϵ is not negligible.

3.3.1. Sensitivity of eigenvalues of $\widehat{A}$ to noise

Theorem 2:

$$\|\mathcal{P}(\Lambda) - \widehat{\Lambda}\|_F \leq c_3\Bigg(c_1\frac{1+\log(1/\delta)}{\sqrt{T}} + c_2\|O^\dagger - \widehat{O}^\dagger\|_F + c_3\|\xi^{-1} - \widehat{\xi}^{-1}\|_F\Bigg), \text{with probability } 1-\delta.$$

where, $c_1 = \|O^\dagger\|^2\|\xi^{-1}\|$, $c_2 = \|\widehat{S_{2,1}}\|\|\xi^{-1}\|\|(O^\top)^\dagger\| + \|\widehat{O}^\dagger\|\|\widehat{S_{2,1}}\|\|\widehat{\xi}^{-1}\|$, $c_3 = \|\widehat{O}^\dagger\|\|\widehat{S_{2,1}}\|\|(O^\top)^\dagger\|$, and $c_4 = \sqrt{\kappa(A)\kappa(\widehat{A})}$ and, T is the number data items used for $\widehat{S_{2,1}}$. We denote $\text{diag}(1/\xi)$ and $\text{diag}(1/\widehat{\xi})$ respectively by, ξ^{-1} and $\widehat{\xi}^{-1}$ to save space.

Proof: See appendix.

We see from this Theorem that if $\widehat{O} = O$, and $T \to \infty$, the deviation between the eigenvalues of A and $\widehat{A}$ goes to zero.

3.4. Refinement of the parameters

Once the de-permutation mapping $\widetilde{\mathcal{P}}$ is obtained, it's possible to refine the parameter estimates to exactly match the corresponding model structure by imposing additional convex constraints depending on the transition structure of the model (given in Table 2) in the convex optimization problem introduced in Section 3.2:

$$\begin{aligned} \min_{A} \;\; & \|\widehat{S_{2,1}} - \widetilde{\mathcal{P}}(\widehat{O})A\text{diag}(\widetilde{\mathcal{P}}(\widehat{\xi}))\widetilde{\mathcal{P}}(\widehat{O}^\top)\|_F \\ s.t. \;\; & 1_{MK}^\top A = 1_{MK}^\top \\ & A \geq 0, \\ & f(A) = b \end{aligned} \tag{8}$$

Here are some choices for the constraint $f(A) = b$, depending on the model structure:

MHMM: $A.{*}\Theta = 0$, where $.{*}$ is the element-wise product, Θ is a binary mask, zero on the block diagonals and one on the off block diagonals. Note that doing this masking in Section 3.2 was not possible because we did not know $\mathcal{P}$, and hence it was not possible to construct a valid Θ.

HMM-M: $A_{i,j}D = \mathbf{0}$, $\forall (i,j)$, where columns of $D \in \mathbb{R}^{M\times\binom{M}{2}}$ have two non-zero elements, one of which is 1 and the other one is -1. This ensures that $A_{i,j}(:,k) = A_{i,j}(:,l), \forall(k,l), k \neq l$.

MM: $A.*\Theta = 0$, and $A_{i,j}D = \mathbf{0}$, $\forall (i,j)$ can be used to impose the model structure. θ and D are defined same as above.

SHMM: $1_M^\top A_{i,j}D = \mathbf{0}$, $\forall(i,j)$, where D is defined same as above. In this case the constraint ensures that all columns within the block $A_{i,j}$ sum up to the same number, which is in accordance with the transition structure of the SHMM.

USHMM: $A.*\Theta = (\widehat{B} \otimes 1_M 1_M^\top).*\Theta$, where Θ is defined same as above, $\widehat{B}$ is an estimate from $\widetilde{\mathcal{P}}(A)$, where, $\widehat{B}_{i,j} = \frac{1}{K}\sum_{l,k}(\widetilde{\mathcal{P}}(A))_{i,j}(l,k)$. This ensures uniformity in off-block diagonals.

After this refinement step we have a transition matrix which is in an exact form, and it is straightforward to extract the local components $\theta_{1:K}$.

4. Experiments

4.1. Depermutation Experiment on Synthetic Data

In this experiment, we investigate the effect of group persistence on the parameter estimation for SHMM. We use uniform transition SHMM model (USHMM) defined in Section 2.3.1. We generated a sequence of length 1000. The emission model is unit variance spherical Gaussian with columns of the emission matrix drawn from a zero mean spherical Gaussian with variance 5. The regime transition matrix is of the form $B = \begin{bmatrix} \alpha & \frac{1-\alpha}{K} \\ \frac{1-\alpha}{K} & \alpha \end{bmatrix}$. The columns of

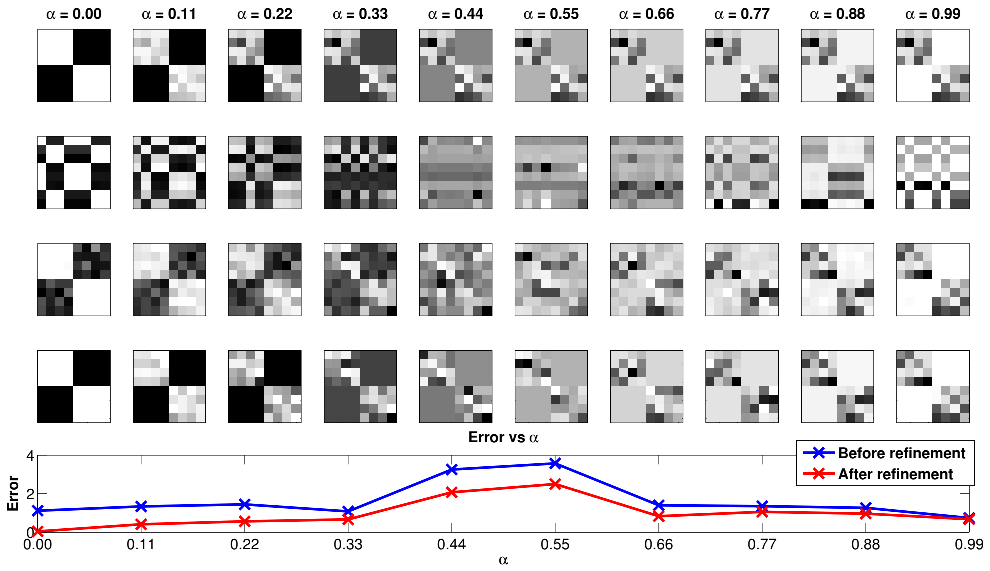


Figure 5. Depermuting estimated transition matrices on synthetic data, (First row) True Transition Matrices, (Second Row) Low Rank Reconstruction $\widehat{A}_r$, (Third Row) $\widetilde{\mathcal{P}}(A)$ before refinement, (Fourth Row) $\widetilde{\mathcal{P}}(A)$ after refinement, (Fifth Row) The error $\sum_{i,j} \|\widetilde{\mathcal{P}}(\widehat{A}_{i.j}) - B_{\mathcal{P}_2(i,j)}\mathcal{P}_1(A_{\mathcal{P}_2(i,j)})\|_F$, before and after refinement. (Note: In this colormap darker tones show larger numbers.)

the blocks $A_{i,j}$ are drawn from a Dirichlet$(1, \dots, 1)$ distribution. We use the tensor power method (Anandkumar et al., 2012a) to recover the columns of O. As can be seen in Figure 5, best results are obtained for large values of α, which correspond to the group persistent case. We compute the error $\sum_{i,j} \|\widetilde{\mathcal{P}}(\widehat{A}_{i.j}) - B_{\mathcal{P}_2(i,j)}\mathcal{P}_1(A_{\mathcal{P}_2(i,j)})\|_F$, by resolving the permutation between and within blocks. It is interesting to note that the permutation is still recoverable for small values of α, which correspond to the anti-group persistent case. In this case the B matrix has negative eigenvalues, but their magnitude is large. Furthermore, we see that the refinement step helps lowering the estimation error for all α values. We also do the same experiment for HMM-M, we include it in the appendix due to space limitations.

4.2. Segmentation Experiment on Synthetic Data

We compared three methods to learn the parameters of an SHMM for segmenting a data sequence. The first is [MoM + spectral], which is the algorithm proposed in the paper, the second is [k-means + spectral], which uses a k-means algorithm to estimate the emission matrix up-to a permutation, followed by a label counting step to estimate the global transition matrix up-to permutation. The spectral algorithm in (Subakan et al., 2014) is used to depermute the parameters. The third method uses randomly generated parameters. In addition, we looked at how useful these parameters are for initializing EM. For randomly initialized EM, we do ten random restarts and take the solution with the highest log likelihood.

We generated $L = 20$ dimensional 100 data sequences of varying lengths for $K = 3$ regimes and $M = 3$ states per regime. We used a unit variance spherical Gaussian emission model with columns of the emission matrix drawn from a zero mean spherical Gaussian with variance 5. The transition structure was set same as the experiment in the previous section with $\alpha = 0.95$. Figure 6 shows the average segmentation accuracy results for the HMM-M and SHMM models, for 100 sequences.

Given a learned parameter set, we used the standard Viterbi decoding to segment the observation sequence by regime (group) label. As expected, a random initialization performs much worse on average for all sequence lengths compared with either the [MoM + spectral] or [k-means + spectral] method. Furthermore, [MoM + spectral] improves over [k-means + spectral] for both models. In the case of the SHMM, the result of EM initialized with either method typically leads to an improvement in segmentation accuracy. In the case of the HMM-M, however, EM can make the solution worse. This is probably due to the fact that EM for learning the HMM-M treats the parameters associated with each state (i.e. $O_k, \nu_k, k = 1, \dots, K$) more independently in the M step than EM for learning the SHMM. Because of this, it has trouble identifying the correct permutation of the states even with a decent initialization. As expected, EM initialized with random parameters fares quite poorly even when taking the solution with the highest log likelihood out of 10 attempts.

Run times for several sequence lengths are shown in Fig-

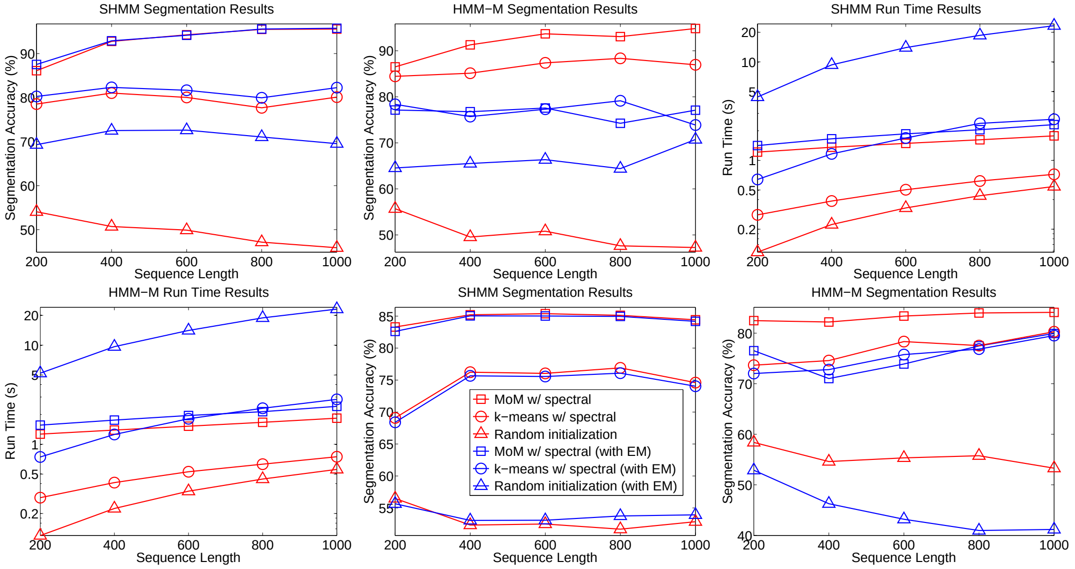


Figure 6. (Top Left) Accuracy for Synthetic Data, SHMM (Top Middle) Accuracy for Synthetic Data, HMM-M (Top Right) Run time Comparison for Synthetic Data, SHMM, (Bottom Left) Run time Comparison for Synthetic Data, SHMM-M, (Bottom Middle) Accuracy for Saxophone Data, SHMM, (Bottom Right) Accuracy for Saxophone Data, HMM-M. Note: The legend applies to all figures.

ure 6. Note that the vertical axis is on a log scale. The time taken for each method is an accumulation of parameter estimation, EM (if applicable), and segmentation. We can see that the [MoM + spectral] method with or without further EM refinement performs better than EM initialized with [k-means + spectral] for longer sequence lengths. This indicates that [MoM + spectral] provides a better initialization for EM than [k-means + spectral], leading to fast convergence. The line corresponding to EM initialized at random demonstrates the linear growth in the amount of data. This arises from the forward-backward passes along the data sequence in the E step. For much larger datasets, the [MoM + spectral] method provides an enormous computational advantage.

4.3. Saxophone note sequence segmentation

The saxophone sequence data consists of the concatenation of saxophone notes on 100 audio files. We generated raw audio data by (1) generating SHMM parameters, (2) sampling regime and state sequences with $K = 3$ and $M = 3$, and (3) building a time-domain signal according to the state sequence. Each switch state (regime) represented a set of note pitches associated with a major or minor key signature and each emission state represented a single note. A note was set to play for as long as the state indices didn't change. We associated each state with a duration of 1,024 samples. This resulted in about 8 time steps per second (at a sampling rate of 16 kHz).

We mapped this time-domain signal to the frequency domain with the Short-Time Fourier Transform (STFT) using window and hop sizes of 1,024 samples and subsequently computed the log-magnitude of the resulting spectra. This results in a sequence of sparse, 513-dimensional vectors. To reduce the dimensionality, we applied a projection to a random 20-dimensional basis. This dataset was then given as input to the learning algorithms. Given a learned parameter set, we used standard Viterbi decoding to segment the low-dimensional observation sequence by regime label. The projection step is beneficial for two reasons: (1) it compresses the (inherently sparse) STFT representation into fewer dimensions to reduce run time and (2) it makes the high-dimensional clusters corresponding to each state more spherical (Dasgupta, 2000). Furthermore, we can appeal to the Johnson-Lindenstrauss Lemma (Dasgupta & Gupta, 1999) to understand why the clusters remain separated after the projection. Figure 6 shows the segmentation accuracy results, for 100 sequences. We can see that they qualitatively mirror the results of the synthetic data trials.

5. Conclusions

In this paper, we have shown that it is possible to use a MoM learning procedure for learning special cases of SHMM. We have shown that the success of the learning algorithm depends on the group persistence property. We have also proposed a convex optimization procedure to be able to recover the model parameters in their true form, which is validated by the experiments. Experiments on sequence segmentation also confirm that the proposed MoM procedure is computationally efficient and accurate compared to the more traditional EM approach.

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S.M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *arXiv:1210.7559v2*, 2012a.

Anandkumar, A., Hsu, D., and Kakade, S.M. A method of moments for mixture models and hidden markov models. In *COLT*, 2012b.

Bhatia, Rajendra, Kittaneh, Fuad, and Li, Ren Cang. Some inequalities for commutators and an application to spectral variation. *Linear and Multilinear Algebra*, 1997.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

Chaganty, Arun Tejasvi and Liang, Percy. Estimating latent-variable graphical models using moments and likelihoods. In *International Conference of Machine Learning (ICML)*, 2014.

Dasgupta, Sanjoy. Experiments with random projection. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, UAI '00, pp. 143–151, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-709-9. URL http://dl.acm.org/citation.cfm?id=647234.719759.

Dasgupta, Sanjoy and Gupta, Anupam. An elementary proof of the johnson-lindenstrauss lemma. Technical report, 1999.

Hsu, Daniel and Kakade, Sham. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Fourth Innovations in Theoretical Computer Science*, 2013.

Hsu, Daniel, Kakade, Sham M., and Zhang, Tong. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, (1460-1480), 2009.

Kontorovich, Aryeh, Nadler, Boaz, and Weiss, Roi. On learning parametric-output hmms. In *International Conference of Machine Learning (ICML)*, 2013.

Murphy, Kevin. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, 2002.

Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition (1989). *Proceedings of the IEEE*, pp. 257–286, 1989.

Smyth, P. Clustering sequences with hidden markov models. In *Advances in neural information processing systems*, 1997.

Subakan, Y. Cem, Traa, Johannes D., and Smaragdis, Paris. Spectral learning of mixture of hidden markov models. In *Neural Information Processing Systems (NIPS)*, 2014.