# Spectral Learning of Hidden Markov Models with Group Persistence, Appendix

## 1. Proofs:

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

**Proof:** Let us consider the product $A^\top(b \otimes 1_M)$, where $\otimes$ denotes the Kronecker product, and $b \in \mathbb{R}^M$ is an eigenvector of $B^\top$. Let $\lambda_b$ denote the corresponding eigenvalue.

$$A^\top(b \otimes 1_M) = \begin{bmatrix} \vdots \\ \sum_k B_{k,1} b_k A_{k,1}^\top 1_M \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \lambda_b b_k 1_M \\ \vdots \end{bmatrix} = \lambda_b(b \otimes 1_M).$$

Note that each block $A_{i,j}$ is stochastic matrix with column sums equal to 1, so every $A_{i,j}^\top$ has an eigenvector equal to with eigenvalue one. And, since $A$ and $A^\top$ (and $B$, $B^\top$) have the same eigenvalues, we conclude that $\lambda_b$ is an eigenvector of $A$. This argument applies to all eigenvectors $b$ of $B^\top$. $\square$

**Theorem 2:**

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

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

**Proof:** We know from (Bhatia et al., 1997) that for two diagonalizable matrices $A = V \Lambda V^{-1}$ and $\widehat{A} = \widehat{V}\widehat{\Lambda}\widehat{V}^{-1}$,

$$\|\mathcal{P}(\Lambda) - \widehat{\Lambda}\|_F \leq \sqrt{\kappa(A)\kappa(\widehat{A})}\, \|A - \widehat{A}\|_F. \tag{1}$$

So, proving the bound amounts to upper bounding the deviation in the estimation $\|A - \widehat{A}\|_F$. We use the estimator given in Equation 5(in the paper, pseudo inverse estimator) for the sake of analysis.

$$\begin{aligned}
&\|A - \widehat{A}\|_F \qquad (2)\\
&= \|O^\dagger S_{1,2} \xi^{-1} (O^\top)^\dagger - \widehat{O}^\dagger \widehat{S_{1,2}} \widehat{\xi}^{-1} (\widehat{O}^\top)^\dagger\|_F \\
&\leq \|O^\dagger S_{1,2} \xi^{-1} (O^\top)^\dagger - O^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger\|_F \\
&\quad + \|O^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger - \widehat{O}^\dagger \widehat{S_{1,2}} \widehat{\xi}^{-1} (\widehat{O}^\top)^\dagger\|_F \\
&\leq \|O^\dagger S_{1,2} \xi^{-1} (O^\top)^\dagger - O^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger\|_F \\
&\quad + \|O^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger - \widehat{O}^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger\|_F \\
&\quad + \|\widehat{O}^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger - \widehat{O}^\dagger \widehat{S_{1,2}} \widehat{\xi}^{-1} (\widehat{O}^\top)^\dagger\|_F, \\
&\leq \|O^\dagger S_{1,2} \xi^{-1} (O^\top)^\dagger - O^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger\|_F \\
&\quad + \|O^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger - \widehat{O}^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger\|_F \\
&\quad + \|\widehat{O}^\dagger \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger - \widehat{O}^\dagger \widehat{S_{1,2}} \widehat{\xi}^{-1} (O^\top)^\dagger\|_F, \\
&\quad + \widehat{O}^\dagger \widehat{S_{1,2}} \widehat{\xi}^{-1} (O^\top)^\dagger - \widehat{O}^\dagger \widehat{S_{1,2}} \widehat{\xi}^{-1} (\widehat{O}^\top)^\dagger\|_F, \\
&= \|O^\dagger (S_{1,2} - \widehat{S_{1,2}}) \xi^{-1} (O^\top)^\dagger\|_F \\
&\quad + \|(O^\dagger - \widehat{O}^\dagger) \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger\|_F \\
&\quad + \|\widehat{O}^\dagger \widehat{S_{1,2}} (\xi^{-1} - \widehat{\xi}^{-1}) (O^\top)^\dagger\|_F \\
&\quad + \|\widehat{O}^\dagger \widehat{S_{1,2}} \widehat{\xi}^{-1} ((O^\top)^\dagger - (\widehat{O}^\top)^\dagger)\|_F,
\end{aligned}$$

where the inequalities are due to triangular inequality. The next step is to bound the individual terms:

$$\begin{aligned}
&\|O^\dagger (S_{1,2} - \widehat{S_{1,2}}) \xi^{-1} (O^\top)^\dagger\| \qquad (3)\\
&\leq \|O^\dagger\| \|S_{1,2} - \widehat{S_{1,2}}\| \|\xi^{-1}\| \|(O^\top)^\dagger\| \\
&\leq \|O^\dagger\|^2 \|\xi^{-1}\| \frac{1 + \log(1/\delta)}{\sqrt{T}}, \text{with probability } 1 - \delta,
\end{aligned}$$

where we omitted the subscript $F$ to save space, used the property $\|AB\|_F \leq \|A\|_F \|B\|_F$ of the Frobenius norm, and the second inequality is from (Hsu et al., 2009). The three remaining terms are also handled using the same property of the Frobenius norm,

$$\begin{aligned}
&\|(O^\dagger - \widehat{O}^\dagger) \widehat{S_{1,2}} \xi^{-1} (O^\top)^\dagger\| \qquad (4)\\
&\leq \|(O^\dagger - \widehat{O}^\dagger)\| \|\widehat{S_{1,2}}\| \|\xi^{-1}\| \|(O^\top)^\dagger\|,
\end{aligned}$$

$$\|\widehat{O}^{\dagger}\widehat{S_{1,2}}(\xi^{-1}-\widehat{\xi}^{-1})(O^{\top})^{\dagger}\|_F \leq \|\widehat{O}^{\dagger}\|\|\widehat{S_{1,2}}\|\|(\xi^{-1}-\widehat{\xi}^{-1})\|\|(O^{\top})^{\dagger}\| \tag{5}$$

$$\|\widehat{O}^{\dagger}\widehat{S_{1,2}}\widehat{\xi}^{-1}((O^{\top})^{\dagger}-(\widehat{O}^{\top})^{\dagger})\| \leq \|\widehat{O}^{\dagger}\|\|\widehat{S_{1,2}}\|\|\widehat{\xi}^{-1}\|\|O^{\dagger}-\widehat{O}^{\dagger}\|. \tag{6}$$

By arranging terms we see that $c_1 = \|O^{\dagger}\|^2\|\xi^{-1}\|$, $c_2 = \|\widehat{S_{1,2}}\|\|\xi^{-1}\|\|(O^{\top})^{\dagger}\| + \|\widehat{O}^{\dagger}\|\|\widehat{S_{1,2}}\|\|\widehat{\xi}^{-1}\|$, $c_3 = \|\widehat{O}^{\dagger}\|\|\widehat{S_{1,2}}\|\|(O^{\top})^{\dagger}\|$, and $c_4 = \sqrt{\kappa(A)\kappa(\widehat{A})}$. □

## 2. Additional Experiment

We do the depermutation experiment in Section 4.1 also for the HMM-M model. We could not include this result in the manuscript due to space limitations. It is given in Figure 1 of the appendix.

## References

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

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.

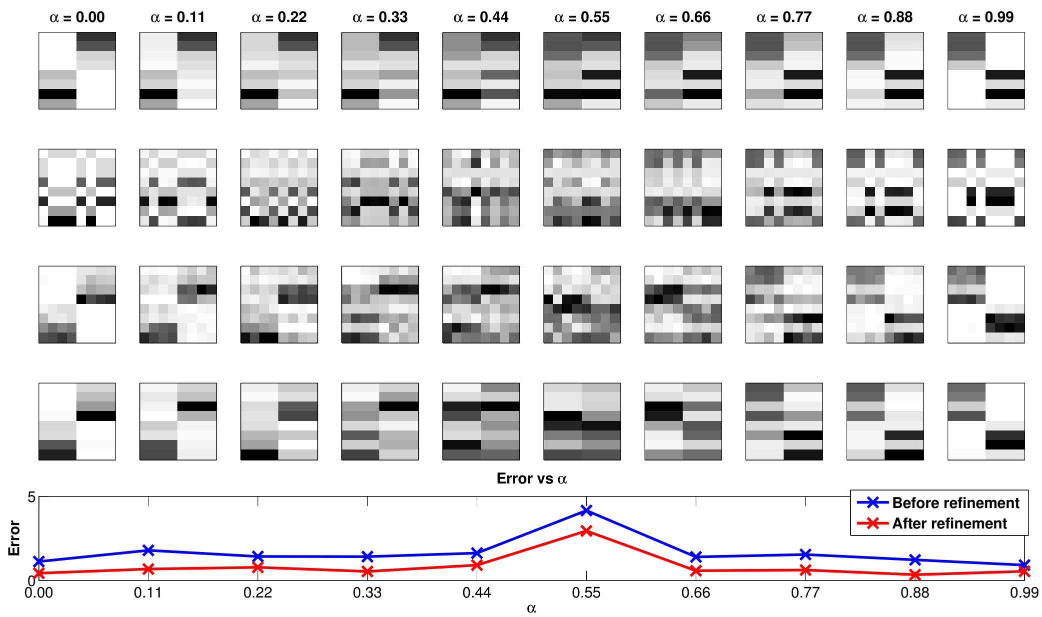


*Figure 1.* Depermuting estimated transition matrices on synthetic data, (First row) True Transition Matrices, (Second Row) Low Rank Reconstruction $\widehat{A}_r$, (Third Row) $\widetilde{\mathcal{P}}(A)$ before refinement, (Fourth Row) $\widetilde{\mathcal{P}}(A)$ after refinement, (Fifth Row) The 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.