Stable Spectral Learning Based on Schur Decomposition

Nicolò Colombo
Luxembourg Centre for Systems Biomedicine
University of Luxembourg
nicolo.colombo@uni.lu

Nikos Vlassis
Adobe Research
San Jose, CA
vlassis@adobe.com

Abstract

Spectral methods are a powerful tool for inferring the parameters of certain classes of probability distributions by means of standard eigenvalue-eigenvector decompositions. Spectral algorithms can be orders of magnitude faster than log-likelihood based and related iterative methods, and, thanks to the uniqueness of the spectral decomposition, they enjoy global optimality guarantees. In practice, however, the applicability of spectral methods is limited due to their sensitivity to model misspecification, which can cause instability issues in the case of non-exact models. We present a new spectral approach that is based on the Schur triangularization of an observable matrix, and we carry out the corresponding theoretical analysis. Our main result is a bound on the estimation error that is shown to depend linearly on the condition number of the ground-truth conditional probability matrix and inversely on the eigengap of an observable matrix. Numerical experiments show that the proposed method is more stable, and performs better in general, than the classical spectral approach using direct matrix diagonalization.

1 INTRODUCTION

The problem of learning mixtures of probability distributions from sampled data is central in the statistical literature (Titterington, 1985; Lindsay, 1995). In pioneering work, Chang (1996) showed that it is possible to learn a mixture of product distributions via the spectral decomposition of 'observable' matrices, that is, matrices that can be estimated directly from the data using suitable combinations of the empirical joint probability distributions (Chang, 1996). Extensions and improvements of this idea have been developed more recently in a series of works, where the spectral technique is applied to a larger class of probability distributions, including Gaussian mixtures, Hidden Markov models, stochastic languages, and others (Mossel and Roch, 2006; Hsu et al., 2012; Anandkumar et al., 2012a,c; Balle et al., 2014; Kuleshov et al., 2015). Some the most widely studied algorithms include Chang's spectral technique (Chang, 1996; Mossel and Roch, 2006), a tensor decomposition approach (Anandkumar et al., 2012a), and an indirect learning method for inferring the parameters of Hidden Markov Models (Hsu et al., 2012).

Spectral algorithms are typically much faster than iterative solvers such as the EM algorithm (Dempster et al., 1977), and thanks to the uniqueness of the spectral decomposition, they enjoy strong optimality guarantees. However, spectral algorithms are more sensitive to model misspecification than algorithms that maximize log-likelihood. Studies in the field of linear system subspace identification have shown that the solutions obtained via matrix decomposition methods can be suboptimal (Favoreel et al., 2000; Buesing et al., 2012). On the other hand, good results have been obtained by using the output of a spectral algorithm to initialize the EM algorithm (Zhang et al., 2014).

The practical implementation of the spectral idea is a non-trivial task because the stability of spectral decomposition strongly depends on the spacing between the eigenvalues of the empirical matrices (Anandkumar et al., 2012a; Hsu and Kakade, 2012). Mossel and Roch (2006) obtain certain eigenvalue separation guarantees for Chang's spectral technique via the contraction of higher (order three) moments through Gaussian random vectors. Anandkumar et al. (2012a) describe a tensor decomposition method that generalizes deflation methods for matrix diagonalization to the case of symmetric tensors of order three. Another algorithmic variant involves replacing the random contracting vector of Chang's spectral technique with an 'anchor observation', which guarantees the presence of at least one well separated eigenvalue (Arora et al., 2012; Song and Chen, 2014) (See also Kuleshov et al. (2015) for a similar idea). Finally, Zou et al. (2013) have presented a technique for learning mixtures of product distributions in the presence of a background model.

In this article we propose an alternative and more stable approach to Chang's method that is based on Schur decomposition (Konstantinov et al., 1994). We show that an approximate triangularization of all observables matrices appearing in Chang's spectral method can be obtained by means of the orthogonal matrices appearing in the Schur decomposition of their linear combination, an idea that has been suggested earlier (Corless et al., 1997; Anandkumar et al., 2012a). Our main result is a theoretical bound on the estimation error that is based on a perturbation analysis of Schur decomposition (Konstantinov et al., 1994). In analogy to related results in the literature, the bound is shown to depend directly on the model mispecification error and inversely on an eigenvalue separation gap. However, the major advantage of the Schur approach is that the bound depends very mildly on the condition number of the ground-truth conditional probability matrix (see discussion after Theorem 1). We compare numerically the proposed Schur decomposition approach with the standard spectral technique (Chang, 1996; Mossel and Roch, 2006), and we show that the proposed method is more stable and does a better job in recovering the parameters of misspecified mixtures of product distributions.

2 SPECTRAL LEARNING VIA SCHUR DECOMPOSITION

Here we discuss the standard spectral technique (Chang, 1996; Mossel and Roch, 2006), and the proposed Schur decomposition, in the context of learning mixtures of product distributions. The complete algorithm is shown in Algorithm 1. Its main difference to previous algorithms is step 13 (Schur decomposition).

The spectral approach in a nutshell. Consider ℓ distinct variables taking values in a discrete set with finite number of elements $\{1, \ldots, d\}$, and a sample S consisting of a number of independent joint observations. The empirical distribution corresponding to these observations is computed by counting the frequencies of all possible joint events in the sample (step 3), and it is modeled (approximated) by a mixture of product distributions with a given number p of mixture components. For every $p < d$, spectral methods allow one to recover the parameters of this approximation by means of the simultaneous diagonalisation of a set of 'observable' nearly diagonalizable matrices $\{\hat{M}_1, \ldots, \hat{M}_p\}$, computed from the sample S (step 10).

If the sample is drawn exactly from a mixture of p components, and in the limit of an infinite amount of data, the mixture parameters, i.e., the conditional probability distributions and the mixing weights of the mixture, are contained exactly in the eigenvalues of the matrices $\hat{M}_i$ (Chang, 1996). If the sample is not drawn exactly from a mixture of p product distributions, and in the typical finite

Algorithm 1 Spectral algorithm via Schur decomposition

Input: data $s_n = [x_n, y_n, z_n] \in \mathscr{N}$, dimension d, number of mixture components p

Output: estimated conditional probability matrices $\hat{X}, \hat{Y}, \hat{Z}$ and mixing weights vector $\hat{w}$

1: $\hat{P} = 0$
2: **for** $s_n \in S$ **do**
3: $\quad \hat{P}_{x_n y_n z_n} = \hat{P}_{x_n y_n z_n} + 1$
4: **end for**
5: **for** $i = 1, \ldots d$ **do**
6: $\quad [\hat{P}_i^Y]_{jk} = \hat{P}_{jik},\ [\hat{P}_i^X]_{jk} = \hat{P}_{ijk},\ [\hat{P}_i^Z]_{jk} = \hat{P}_{jki}$
7: **end for**
8: compute $[\hat{P}_{xz}] = \sum_i [\hat{P}_i^Y]$, $[\hat{P}_{yz}] = \sum_i [\hat{P}_i^X]$, $[\hat{P}_{xy}] = \sum_i [\hat{P}_i^Z]$
9: **for** $i = 1, \ldots, d$ **do**
10: $\quad$ compute $\hat{M}_i = \hat{P}_i^Y \hat{P}_{xz}^{-1}$
11: **end for**
12: find $\theta \in \mathbf{R}^d$ such that $\hat{M} = \sum_i \theta_i \hat{M}_i$ has real non-degenerate eigenvalues.
13: find $\hat{U}$ such that $\hat{U}^T \hat{U} = 1$ and $\hat{U}^T \hat{M} \hat{U}$ is upper triangular (Schur decomposition)
14: **for** $i, j = 1, \ldots, d$ **do**
15: $\quad$ let $\hat{Y}_{i,j} = [\hat{U}^T \hat{M}_i \hat{U}]_{jj}$
16: $\quad$ set $\hat{Y}_{i,j} = 0$ if $[\hat{U}^T \hat{M}_i \hat{U}]_{jj} < 0$
17: **end for**
18: normalize to 1 the columns of $\hat{Y}$
19: **for** $i = 1, \ldots, d$ **do**
20: $\quad$ compute $\hat{M}_i^X = \hat{P}_i^X \hat{P}_{yz}^{-1}$
21: $\quad$ compute $\hat{M}_i^Z = \hat{P}_i^Z \hat{P}_{xy}^{-1}$
22: **end for**
23: **for** $i, j = 1, \ldots, d$ **do**
24: $\quad$ let $\hat{X}_{i,j} = [\hat{Y}^{-1} \hat{M}_i^X \hat{Y}]_{jj}$ and set $\hat{X}_{i,j} = 0$ if $[\hat{Y}^{-1} \hat{M}_i^X \hat{Y}]_{jj} < 0$
25: $\quad$ let $\hat{Z}_{i,j} = [\hat{X}^{-1} \hat{M}_i^Z \hat{X}]_{jj}$ and set $\hat{Z}_{i,j} = 0$ if $[\hat{X}^{-1} \hat{M}_i^Z \hat{X}]_{jj} < 0$
26: **end for**
27: normalize to 1 the columns of $\hat{X}$ and $\hat{Z}$
28: compute $\hat{w} = \hat{X}^{-1} \hat{P}_{xy} (\hat{Y}^T)^{-1}$ and normalize to 1

sample setting, the model is only an approximation to the empirical distribution, and as a result, the matrices $\hat{M}_i$ are no longer simultaneously diagonalizable and an approximate diagonalisation technique is required. The standard approach consists of choosing one particular observable matrix in the set, or a linear combination of all matrices, and use its eigenvectors to diagonalize each $\hat{M}_i$ (Mossel and Roch, 2006).

Here we propose a new approach that is based on the Schur decomposition of a linear combination of the observable matrices. In particular, we first mix the matrices $\hat{M}_i$ to compute a candidate matrix $\hat{M}$ (step 12), and then we apply Schur decomposition to $\hat{M}$ (step 13). The eigenvalues of each $\hat{M}_i$, and thereby the model parameters, are then extracted using the orthogonal matrix $\hat{U}$ of the Schur decomposition (steps 15-16). Effectively we exploit the fact that the real eigenvalues of a matrix A always appear on the diagonal of its Schur triangularization $T = U^T AU$, even though the entries of the strictly upper diagonal part of T may not be unique. Using the perturbation analysis of the Schur system of a matrix by Konstantinov et al. (1994), we obtain a theoretical bound on the error of such eigenvalue estimation as a function of the model misspecification error, the condition number of the ground-truth matrix X, and the separation of the eigenvalues of $\hat{M}$ (Theorem 1).

Detailed description and the Schur approach. Consider for simplicity a sample S of independent observations $s = [x, y, z]$ of three distinct variables taking values in the discrete set $\{1, \dots, d\}$. The empirical distribution associated to the sample S is defined as

$$\hat{P}_{i,j,k} = \frac{1}{|S|} \sum_{s \in S} \delta_{x,i} \delta_{y,j} \delta_{z,k} \tag{1}$$

where $|S|$ is the number of elements in S and $\delta_{ab} = 1$ if $a = b$ and zero otherwise. The empirical distribution $\hat{P}$ is a nonnegative order-3 tensor whose entries sum to one. Its nonnegative rank $\mathrm{rank}_+(\hat{P})$ is the minimum number of rank-1 tensors, i.e., mixture components, required to express $\hat{P}$ as a mixture of product distributions. Such a decomposition of $\hat{P}$ (exact or approximate) is always possible (Lim and Comon, 2009). Hence, for any choice of $p \leq \mathrm{rank}_+(\hat{P})$, we can hypothesize that $\hat{P}$ is generated by a model

$$\hat{P} = P + \epsilon\, \Delta P, \qquad \epsilon \geq 0 \tag{2}$$

where $P \in [0,1]^{d_x \times d_y \times d_z}$ is a nonnegative rank-p approximation of $\hat{P}$, ϵ is a model mispecification parameter, and $\Delta P \in [0,1]^{d_x \times d_y \times d_z}$ is a nonnegative tensor whose entries sum to one. The rank-p component P is interpreted as the mixture of product distributions that approximates the empirical distribution, and it can be written

$$P_{ijk} = \sum_{h=1}^{p} w_h\, X_{ih} Y_{jh} Z_{kh}\,, \tag{3}$$

where $w_h \in [0,1]$ for all $h = 1, \dots, p$, and we have defined the conditional probability matrices

$$X \in [0,1]^{d \times p}, \quad \mathbf{1}_d^T X = \mathbf{1}_p^T\,, \tag{4}$$

(and similarly for Y, Z), where $\mathbf{1}_n$ is a vector of n ones. The columns of the matrices X, Y, Z encode the conditional probabilities associated with the p mixture components, and the mixing weights satisfy $\sum_h w_h + \epsilon = 1$.

The conditional probability matrices X, Y, Z and the mixing weight w of the rank-p mixture can be estimated from the approximate eigenvalues of a set of observable matrices $\hat{M}_i$, for $i = 1, \dots p$, that are computed as follows. Let for simplicity $p = d$ and consider the matrices $[\hat{P}_i^Y]_{jk} = \hat{P}_{jik}$ (step 6) and $[\hat{P}_{xz}] = \sum_i [\hat{P}_i^Y]$ (step 8). Assuming that $\hat{P}_{xz}$ is invertible, we define (step 10)

$$\hat{M}_i = \hat{P}_i^Y\, \hat{P}_{xz}^{-1} \tag{5}$$

for $i = 1, \dots, d$. Under the model assumption $\hat{P} = P + \epsilon\, \Delta P$, it is easy to show that

$$\hat{M}_i = M_i + \Delta M_i + o(\epsilon^2) \tag{6}$$

where $\Delta M_i \in \mathbf{R}^{d \times d}$ is linear in the misspecification parameter ϵ, and

$$M_i = X\, \mathrm{diag}(Y_{i1}, \dots Y_{ip})\, X^{-1} \tag{7}$$

where $\mathrm{diag}(v_1, \dots v_d)$ denotes a diagonal matrix whose diagonal entries are $v_1, \dots v_d$. If the model is exact, i.e., $p = \mathrm{rank}_+(\hat{P})$ or equivalently $\epsilon = 0$, the matrices $\{\hat{M}_i\}$ are simultaneously diagonalizable and the entries of the conditional probability matrix Y are given (up to normalization of its columns) by

$$Y_{ij} \propto [V^{-1} \hat{M}_i V]_{jj}, \qquad i, j = 1, \dots d \tag{8}$$

where V is the matrix of the eigenvectors shared by all $\hat{M}_i$.

When $\epsilon \neq 0$, the matrices $\hat{M}_i$ are no longer simultaneously diagonalizable and an approximate simultaneous diagonalisation scheme is needed. The standard procedure consists of selecting a representative matrix $\hat{M}$, compute its eigenvectors $\hat{V}$, and use the matrix $\hat{V}$ to obtain the approximate eigenvalues of all matrices $\hat{M}_i$ and thereby estimate Y_{ij} (Mossel and Roch, 2006; Hsu et al., 2012; Anandkumar et al., 2012b). In this case, the estimation error is known to depend on the model misspecification parameter ϵ and on the inverse of an eigenvalue separation γ (see, e.g., eq. (12)). Using matrix perturbation theorems and properties of the Gaussian distribution, Mossel and Roch (2006) have shown that a certain separation $\gamma > \alpha$ is guaranteed with probability proportional to $(1-\alpha)$ if $\hat{V}$ is the matrix of the eigenvectors of some $\hat{M} = \sum_i \theta_i \hat{M}_i$, with θ sampled from a Gaussian distribution of zero mean and unit variance. In practice, however, this approach can give rise to

instabilities (such as negative or imaginary values for Y_{ij}), especially when the size of the empirical matrices grows.

Here we propose instead to triangularize the matrices $\hat{M}_i$ by means of the Schur decomposition of their linear combination $\hat{M} = \sum_i \theta_i \hat{M}_i$, for an appropriate θ (steps 12-13). The orthogonal matrix $\hat{U}$ obtained from the Schur decomposition $\hat{M} = \hat{U}\hat{T}\hat{U}^T$ is then used in place of the eigenvectors matrix of $\hat{M}$ to approximately triangularize all the observable matrices $\hat{M}_i$ and thereby recover the mixture parameters (steps 15 and 24, 25). For example, the conditional probability matrix Y is estimated as

$$\hat{Y}_{ij} \propto [\hat{U}^T \hat{M}_i \hat{U}]_{jj}, \qquad i, j = 1, \dots d \tag{9}$$

and normalized so that its columns sum to one. Let $\|\cdot\|$ denote the Frobenius norm. Our main result is the following:

Theorem 1. *Let $\hat{M}_i$ and M_i be the real $d \times d$ matrices defined in* (5) *and* (6)*. Suppose it is possible to find $\theta \in \mathbf{R}^d$ such that $\hat{M} = \sum_k \theta_k \hat{M}_k$ has real distinct eigenvalues. Then, for all $j = 1, \dots, d$, there exists a permutation π such that*

$$|\hat{Y}_{ij} - Y_{i\pi(j)}| \leq \left(a_1 \, \frac{k(X)\, \lambda_{\max}}{\hat{\gamma}} + 1 \right) E + o(E^2) \tag{10}$$

where $k(X) = \frac{\sigma_{max}(X)}{\sigma_{min}(X)}$ is the condition number of the ground-truth conditional probability matrix X, $\lambda_{\max} = \max_{i,j} Y_{i,j}$,

$$\hat{\gamma} = \min_{i \neq j} \left| \lambda_i(\hat{M}) - \lambda_j(\hat{M}) \right| > 0 \tag{11}$$

with $\lambda_i(\hat{M})$ being the ith eigenvalue of $\hat{M}$, $a_1 = \|\theta\| \sqrt{\frac{2^3 d^2}{d-1}}$, and $E = \max_i \|\Delta M_i\| = \max_i \|\hat{M}_i - M_i\|$.

Proof. See appendix. □

The analogous bound for the diagonalization approach is (Anandkumar et al., 2012c, Section B6, eq. 11)

$$|\hat{Y}_{ij} - Y_{i\pi(j)}| \leq \left(a_2 k(X)^4 \frac{\tilde{\lambda}_{\max}}{\gamma} + a_3 k(X)^2 \right) E, \tag{12}$$

where $\gamma = \min_{i \neq j} |\lambda_i(\sum_k \theta_k M_k) - \lambda_j(\sum_k \theta_k M_k)|$, $\tilde{\lambda}_{\max} = \max(\max_i [\theta^T Y]_i, \max_{i,j} Y_{i,j})$, and a_2, a_3 are constants that depend on the dimensions of the involved matrices.

When the model misspecification error E is not too large, the error bound under the Schur approach (10) is characterized by a much smoother dependence on $k(X)$ than the error bound (12). Moreover, the Schur bound depends on the eigenvalue gap $\hat{\gamma}$ of an observable matrix, and hence it can be controlled in practice by optimizing θ. The simplified dependence on $k(X)$ of the Schur bound is due to the good perturbation properties of the orthogonal matrices involved in the Schur decomposition, as compared to the eigenvector matrices of the Chang approach. The difference in the bounds suggests that, for a randomly generated true model, a spectral algorithm based on the Schur decomposition is expected to be more stable and accurate in practice that an algorithm based on matrix diagonalization. Intuitively, the key to the improved stability of the Schur approach comes from the freedom to ignore the non-unique off-diagonal parts in Schur triangulation.

During the reviewing process we were made aware of the work of Kuleshov et al. (2015), who propose computing a tensor factorization from the simultaneous diagonalization of a set of matrices obtained by projections of the tensor along random directions. Kuleshov et al. (2015) establish an error bound that is independent of the eigenvalue gap, but their approach does not come with global optimality guarantees (but the authors report good results in practice). It would be of interest to see whether such random projections combined with a simultaneous Schur decomposition (see, e.g., De Lathauwer et al. (2004)) could offer improved bounds.

3 EXPERIMENTS

We have compared the performance of the proposed spectral algorithm based on Schur decomposition with the classical spectral method based on eigenvalue decomposition. The two algorithms that we tested are equivalent except for line 13 of Algorithm 1, which in the classical spectral approach should be "find V such that $V^{-1}MV = D$, with D diagonal". In all experiments we used the same code with decompositions performed via the two Matlab functions `schur(M)` and `eig(M)` respectively. We tested the two algorithms on simulated real multi-view and Hidden Markov Model data. In what follows we denote by 'schur' the algorithm based on the Schur decomposition and by 'eig' the algorithm based on the eigenvalues-eigenvector decomposition.

In the first set of experiments we generated multi-view data from a mixture of product distributions of p mixture components in d dimensions. For each experiment, we created two different datasets $\mathcal{N} = \{[x_n y_n z_n] \in [1, \dots, d]^3\}$, and $\mathcal{N}_{test}$, one for training and one for testing, the latter containing the labels $L \in [1, \dots, p]^{|\mathcal{N}_{test}|}$ of the mixture components that generated each instance. The output was evaluated by measuring the distance between the estimated conditional probability distributions $\hat{X}, \hat{Y}, \hat{Z}$ and the corresponding ground-truth values X, Y, Z:

$$E = \|\hat{X} - X\|^2 + \|\hat{Y} - Y\|^2 + \|\hat{Z} - Z\|^2. \tag{13}$$

Since the order of the columns in $\hat{X}, \hat{Y}, \hat{Z}$ may be different from X, Y, Z, the norms were computed after obtaining the best permutation. We also tested according to a

$\|\mathcal{N}\|(d=10,p=5)$	E_{schur}	E_{eig}	S_{schur}	S_{eig}	$\|\hat{T}_{schur}-T\|$	$\|\hat{T}_{eig}-T\|$
1000	**0.057** (0.013)	0.066 (0.0167)	**0.364** (0.135)	0.360 (0.135)	**0.016** (0.004)	0.026 (0.009)
2000	**0.039** (0.008)	0.120 (0.227)	**0.415** (0.068)	0.356 (0.138)	**0.011** (0.004)	0.019 (0.008)
5000	**0.043** (0.012)	0.046 (0.013)	**0.387** (0.114)	0.386 (0.084)	**0.013** (0.004)	0.021 (0.004)
10000	**0.036** (0.014)	0.047 (0.009)	0.390 (0.130)	**0.402** (0.085)	**0.013** (0.006)	0.022 (0.007)
20000	**0.032** (0.014)	0.113 (0.230)	0.431 (0.124)	**0.341** (0.157)	**0.011** (0.007)	0.025 (0.007)
50000	**0.019** (0.015)	0.025 (0.010)	**0.475** (0.1887)	0.434 (0.143)	**0.007** (0.006)	0.015 (0.009)

Table 1: Columns recovery error E, classification score S, and distance of the approximate distribution $\|\hat{T}-T\|$ for multi-view datasets of increasing size. The algorithm based on Schur decomposition obtained the best scores on almost all datasets.

classification rule where the estimated conditional probability matrices $[\hat{X},\hat{Y},\hat{Z}]$ were used to assign each triple in the test dataset to one of the mixture components. For every run, we obtained a classification score by counting the number of successful predictions divided by the number of elements in the test dataset:

$$S=\frac{1}{|\mathcal{N}_{test}|}\sum_{n\in\mathcal{N}_{test}} f(n), \tag{14}$$

$$f(n)=\begin{cases} 0 & \arg\max_i \hat{X}_{x_n i}\hat{Y}_{y_n i}\hat{Z}_{z_n i}\neq L(n) \\ 1 & \arg\max_i \hat{X}_{x_n i}\hat{Y}_{y_n i}\hat{Z}_{z_n i}= L(n)\end{cases}. \tag{15}$$

Finally we computed the distance in norm between the recovered tensor

$$\hat{T}_{ijk}=\sum_r \hat{w}_r[\hat{X}]_{ir}[\hat{Y}]_{jr}[\hat{Z}]_{kr} \tag{16}$$

and the original tensor

$$T_{ijk}=\sum_r [w]_r[X]_{ir}[Y]_{jr}[Z]_{kr} \tag{17}$$

as follows

$$\|\hat{T}-T\|=\sqrt{\sum_{i,j,k}[\hat{T}-T]^2_{ijk}}\,. \tag{18}$$

In Table 1 we show the results obtained by the two algorithms for $d=10$, $p=5$ and increasing size of the training dataset $|\mathcal{N}|$. In the table we report the average score over 10 analogous runs and the corresponding standard deviation in brackets. When the recovered matrices contained infinite values we have set $E=\|X\|^2+\|Y\|^2+\|Z\|^2$, and $\|\hat{T}-T\|=\|T\|$. The proposed algorithm based on Schur decomposition obtained the best scores on almost all datasets.

In the second set of experiments we tested the two algorithms on datasets generated by a d-dimensional Hidden Markov Model with p hidden states. For each experiment we randomly picked a model $M_{true}=M_{true}(O_{true},R_{true},h_{true})$, where $O_{true}\in[0,1]^{d\times p}$ is the observation matrix, $R_{true}\in[0,1]^{p\times p}$ is the transition matrix, and $h_{true}\in[0,1]^p$ is the starting distribution, and we generated two sample datasets, one for training and one for testing. All sequences s_n were simulated starting from an initial hidden state drawn from h_{true} and following the dynamics of the model according to R_{true} and O_{true}. The length of the sequences in the training and testing datasets was set to 20. We evaluated the two algorithms based on the columns recovery error E as in the previous set of experiments. Also, letting $O_{true}=[o_{true1},\ldots,o_{truep}]$ and $O=[o_1,\ldots,o_p]$, we considered a recovery ratio $R(M)=\frac{r}{p}$, where r is the number of columns satisfying

$$\|o_{truei}-o_i\|^2<\xi,\qquad \xi=0.05^2*d\,. \tag{19}$$

In Table 2 we show the results for recovering a $d=\{5,10,20,30\}$ HMM with $p=5$ hidden states. All values are computed by averaging over 10 experiments and the corresponding standard variation is reported between brackets. The Schur algorithm is in general better than the classical approach. We note that, for a fixed number of hidden states, the inference of the HMM parameters becomes harder as the dimensionality of the space decreases. As the recovery ratio R shows, in the limit situation $d=p$ both algorithms fail (values $R=0$ imply unstable solutions).

4 CONCLUSIONS

We have presented a new spectral algorithm for learning multi-view mixture models that is based on the Schur decomposition of an observable matrix. Our main result is a theoretical bound on the estimation error (Theorem 1), which is shown to depend very mildly (and much more smoothly than in previous results) on the condition number of the ground-truth conditional probability matrix, and inversely on the eigengap of an observable matrix. Numerical experiments show that the proposed method is more stable, and performs better in general, than the classical spectral approach using direct matrix diagonalization.

Appendix - Proof of Theorem 1

Theorem 1. *Let $\hat{M}_i$ and M_i be the real $d\times d$ matrices defined in* (5) *and* (6)*. Suppose it is possible to find $\theta\in\mathbf{R}^d$ such that $\hat{M}=\sum_k\theta_k\hat{M}_k$ has real distinct eigenvalues.*

$\lvert\mathcal{N}\rvert(d=30,p=5)$	$E(T_{schur})$	$E(T_{eig})$	$R(T_{schur})$	$R(T_{eig})$
100	**0.011** (0.001)	0.014 (0.005)	1 (0)	1 (0)
500	**0.011** (0.001)	0.012 (0.002)	1 (0)	1 (0)
1000	**0.011** (0.001)	0.013 (0.002)	1 (0)	1 (0)
2000	0.011 (0.001)	0.011 (0.001)	1 (0)	1 (0)
5000	0.010 (0.001)	0.010 (0.001)	1 (0)	1 (0)
$\lvert\mathcal{N}\rvert(d=20,p=5)$	$E(T_{schur})$	$E(T_{eig})$	$R(T_{schur})$	$R(T_{eig})$
100	**0.019** (0.002)	0.026 (0.010)	**1** (0)	0.880 (0.168)
500	**0.018** (0.003)	0.044 (0.080)	**1** (0)	0.900 (0.316)
1000	**0.019** (0.004)	0.021 (0.002)	1 (0)	1 (0)
2000	0.017 (0.002)	0.017 (0.002)	1 (0)	1 (0)
5000	**0.015** (0.002)	0.018 (0.006)	**1** (0)	0.960(0.126)
$\lvert\mathcal{N}\rvert(d=10,p=5)$	$E(T_{schur})$	$E(T_{eig})$	$R(T_{schur})$	$R(T_{eig})$
100	**0.047** (0.011)	0.050 (0.011)	0.200 (0.188)	**0.220** (0.175)
500	**0.044** (0.008)	0.097 (0.154)	**0.240** (0.157)	0.200 (0.188)
1000	**0.046** (0.017)	0.051 (0.018)	**0.260** (0.211)	0.120 (0.139)
2000	**0.043** (0.016)	0.048 (0.011)	**0.180** (0.220)	0.120 (0.139)
5000	**0.040** (0.013)	0.089 (0.153)	**0.380** (0.257)	0.200 (0.133)
$\lvert\mathcal{N}\rvert(d=5,p=5)$	$E(T_{schur})$	$E(T_{eig})$	$R(T_{schur})$	$R(T_{eig})$
100	**0.163** (0.089)	0.164 (0.074)	0 (0)	0 (0)
500	**0.173** (0.063)	0.228 (0.294)	0 (0)	**0.040** (0.084)
1000	**0.191** (0.068)	0.251 (0.273)	0.020 (0.063)	**0.040** (0.084)
2000	0.205 (0.070)	**0.166** (0.052)	**0.060** (0.096)	0 (0)
5000	**0.142** (0.063)	0.164 (0.069)	0 (0)	**0.020** (0.063)

Table 2: Columns recovery error E and recovery ratio R for recovering HMMs of various dimensionality and hidden states using the Schur and the standard spectral approach. See text for details.

Then, for all $j = 1, \ldots, d$, there exists a permutation π such that

$$|\hat{Y}_{ij} - Y_{i\pi(j)}| \leq \left(a_1 \, \frac{k(X)\,\lambda_{\max}}{\hat{\gamma}} + 1 \right) E + o(E^2) \quad (20)$$

with $\hat{Y}$ estimated from (9)*, and where $k(X) = \frac{\sigma_{max}(X)}{\sigma_{min}(X)}$ is the condition number of the ground-truth conditional probability matrix X, $\lambda_{\max} = \max_{i,j} Y_{i,j}$,*

$$\hat{\gamma} = \min_{i \neq j} \left| \lambda_i(\hat{M}) - \lambda_j(\hat{M}) \right| > 0 \quad (21)$$

with $\lambda_i(\hat{M})$ being the ith eigenvalue of $\hat{M}$, $a_1 = \|\theta\| \sqrt{\frac{2^3 d^2}{d-1}}$, and $E = \max_i \|\Delta M_i\| = \max_i \|\hat{M}_i - M_i\|$.

Proof. Consider the set of real commuting matrices M_i, $i = 1, \ldots d$, and their random perturbations $\hat{M}_i = M_i + \Delta M_i$ defined in (5) and (6). Assume that $\|\Delta M_i\| < E$ for all $i = 1, \ldots, d$ and that $\theta \in \mathbf{R}^d$ is such that the eigenvalues of $\hat{M} = M + \Delta M = \sum_i \theta_i (M_i + \Delta M_i)$ are real and non-degenerate. Let $\hat{U}$ be the orthogonal matrix defined by the Schur decomposition of $\hat{M} = \hat{U}^T \hat{T} \hat{U}$ computed by the matrix decomposition subroutine in Algorithm 1. Note that $\hat{U}$ may not be unique and different choices of $\hat{U}$ lead to different entries in the strictly upper-diagonal part of $\hat{T}$. However, for any given $\hat{U}$ such that $\hat{U}^T \hat{T} \hat{U}$ is upper triangular, there exists an orthogonal matrix U and a real matrix $\Delta U \in \mathbf{R}^{d,d}$ such that $U = \hat{U} + \Delta U$ and

$$\begin{aligned} U^T M U &= (\hat{U} + \Delta U)^T (\hat{M} - \Delta M)(\hat{U} + \Delta U) && (22) \\ &= \hat{T} - \Delta T && (23) \\ &= T && (24) \end{aligned}$$

with T upper triangular. Let Y_{ij} be the ground-truth matrix defined in (3) and $\hat{Y}$ the estimation output by Algorithm 1. Then, assuming that ΔM and ΔU are small, there exists a permutation of the indexes π such that, for all $i, j = 1, \ldots d$

$$\begin{aligned} \delta_y = \; & |\hat{Y}_{ij} - Y_{i\pi(j)}| && (25) \\ = \; & |[\hat{U}^T \hat{M}_i \hat{U}]_{jj} - [U^T M_i U]_{\pi(j)\pi(j)}| && (26) \\ \leq \; & \|(U - \Delta U)^T (M_i + \Delta M_i)(U - \Delta U) - T_i\| && (27) \\ = & \|\Delta U^T U T_i + T_i U^T \Delta U - U^T \Delta M_i U + o(\Delta^2)\| && (28) \\ = \; & \|x T_i - T_i x + U^T \Delta M_i U + o(\Delta^2)\| && (29) \\ \leq \; & 2\,\|x\| \|T_i\| + \|\Delta M_i\| + o(\|\Delta^2\|) && (30) \\ \leq \; & 2\,\|x\| \mu + E + o(\|\Delta^2\|) && (31) \end{aligned}$$

where we have defined $x = U^T \Delta U$, $\mu = \max_i \|M_i\|$ and used $1 = (U + \Delta U)^T (U + \Delta U) = 1 + x^T + x + o(\Delta^2)$ where $o(\Delta^2) = o(x^2) + o(\Delta M x)$.

Following (Konstantinov et al., 1994), a linear bound of $\|x\|$ can be estimated as follows. First, observe that the Schur decomposition of M in (24) implies

$$\mathrm{low}(\hat{T}\hat{x} - \hat{x}\hat{T}) = \mathrm{low}(\hat{U}^T \Delta M \hat{U}) + o(\Delta^2) \quad (32)$$

where $\mathrm{low}(A)$ denotes the strictly lower diagonal part of A and $\hat{x} = \hat{U}^T \Delta U$. Since $\hat{T}$ is upper triangular, one has $\mathrm{low}(\hat{T}\hat{x} - \hat{x}\hat{T}) = \mathrm{low}(\hat{T}\mathrm{low}(\hat{x}) - \mathrm{low}(\hat{x})\hat{T})$, *i.e.* the linear operator defined by $\mathscr{L}_{\hat{T}}(\hat{x}) = \mathrm{low}(\hat{T}\hat{x} - \hat{x}\hat{T})$ maps strictly lower-triangular matrices to strictly lower-triangular matrices. Let $\widetilde{\mathscr{L}_{\hat{T}}}(\cdot)$ be the restriction of $\mathscr{L}_{\hat{T}}(\cdot)$ to the subspace

of lower-triangular matrices, then from (32) one has

$$\widetilde{\mathscr{L}_{\hat{T}}}(\mathrm{low}(\hat{x})) = \mathrm{low}(\hat{U}^T \Delta M \hat{U}) + o(\Delta^2) \qquad (33)$$

and the operator $\widetilde{\mathscr{L}_{\hat{T}}}$ is invertible. The invertibility of $\widetilde{\mathscr{L}_{\hat{T}}}$ follows form the non-singularity of its matrix representation $\mathrm{mat}(\widetilde{\mathscr{L}_{\hat{T}}})$ defined by

$$\mathrm{vec}\left(\widetilde{\mathscr{L}_{\hat{T}}}(\mathrm{low}(\hat{x}))\right) = \mathrm{mat}(\widetilde{\mathscr{L}_{\hat{T}}}) L\, \mathrm{vec}(\mathrm{low}(\hat{x})) \qquad (34)$$

where $\mathrm{vec}(A)$ is the columnwise vector representation of A and $L = [L_{ij}] \in [0,1]^{\frac{d(d-1)}{2} \times d^2}$ the projector to the subspace of vectorized lower-triangular matrices

$$L_{ij} \in [0,1]^{d-i \times d}, \qquad i,j = 1, \ldots, d-1 \qquad (35)$$

$$L_{ij} = \begin{cases} 0_{d-i,d} & i \neq j \\ [0_{d-i,i}, 1_{d-i}] & i = j \end{cases} \qquad (36)$$

More explicitly, $\mathrm{mat}(\widetilde{\mathscr{L}_{\hat{T}}}) = L\ (1 \otimes \hat{T} - \hat{T}^T \otimes 1)\ L^T$ is a block lower-triangular matrix $\mathrm{mat}(\widetilde{\mathscr{L}_{\hat{T}}}) = [\mathscr{M}_{ij}] \in \mathbf{R}^{\frac{d(d-1)}{2} \times \frac{d(d-1)}{2}}$ where

$$[\mathscr{M}_{ij}] \in \mathbf{R}^{d-i \times d-j}, \qquad i,j = 1, \ldots, d-1 \qquad (37)$$

$$\mathscr{M}_{ij} = \begin{cases} [0_{d-i,i-j}, 1_{d-j}] & i > j \\ [m_i] & i = j \\ 0 & i < j \end{cases} \qquad (38)$$

$$[m_i]_{jk} = \begin{cases} \hat{T}_{j+i-1,k+i} & j < k \\ \hat{T}_{j+i,j+i} - T_{i,i} & j = k \\ 0 & j > k \end{cases} \qquad (39)$$

for $i,j = 1, \ldots d-1$. The determinant of $\mathscr{M}$ is the product of the determinants of its diagonal blocks, *i.e.*

$$\det(M) = \prod_{i>j} (\hat{T}_{ii} - \hat{T}_{jj}) \qquad (40)$$

and is not null provided that the eigenvalues of $\hat{T}$ are real separated. In this case, the matrix $\mathscr{M}$ and hence the operator $\widetilde{\mathscr{L}_{\hat{T}}}(\cdot)$ are invertible. From (32) one has

$$\mathrm{low}(\hat{x}) = \widetilde{\mathscr{L}_{\hat{T}}^{-1}} \mathrm{low}\ (\hat{U}^T \Delta M \hat{U}) + o(\Delta^2) \qquad (41)$$

and in particular

$$\|\hat{x}\| = \sqrt{2} \|\mathrm{low}(\hat{x})\| \qquad (42)$$

$$= \sqrt{2} \|\widetilde{\mathscr{L}_T^{-1}}\|_F \|\Delta M\| + o(\|\Delta\|^2) \qquad (43)$$

where the first equality is obtained using the linear approximation $\hat{x} = -\hat{x}^T$ and $\|A\|^2 = \|\mathrm{low}(A)\|^2 + \|\mathrm{diag}(A)\|^2 + \|\mathrm{up}(A)\|^2$, with $\mathrm{diag}(A)$ and $\mathrm{up}(A)$ denoting the diagonal and upper-diagonal parts of A. The norm of the inverse operator can be bound using its matrix realization, *i.e.*

$$\|\widetilde{\mathscr{L}_{\hat{T}}^{-1}}\|_F = \|\mathscr{M}^{-1}\| \leq \frac{1}{\sigma_{\min}(\mathscr{M})} \qquad (44)$$

where $\sigma_{\min}(A)$ is the smallest singular value of A. We can estimate $\sigma_{\min}(\mathscr{M})$ by using the following lemma

$$\sigma_{\min}(A) = \min_{\mathrm{rank}(B)<n} \|A - B\|, \qquad \mathrm{rank}(A) = n \quad (45)$$

and observing that the rank deficient matrix closest to $\mathscr{M}$ is obtained by setting $\hat{T} = T_{\mathrm{singular}}$ in (38), where T_{singular} is defined by

$$[T_{\mathrm{singular}}]_{i,j} = \begin{cases} \hat{T}_{j^*,j^*} & \text{if} \quad i = j = i^* \\ \hat{T}_{i,j} & \text{otherwise} \end{cases} \qquad (46)$$

with $(i^*, j^*) = \arg\min_{i \neq j} |\hat{T}_{ii} - \hat{T}_{jj}|$. One has

$$\sigma_{\min}(\mathscr{M}) = \|\mathscr{M} - \mathscr{M}_{\mathrm{singular}}\| \qquad (47)$$

$$= \sqrt{\sum_{i,j} (\mathscr{M} - \mathscr{M}_{\mathrm{singular}})_{i,j}^2} \qquad (48)$$

$$= \sqrt{d-1}\, \hat{\gamma} \qquad (49)$$

$$\hat{\gamma} = |\hat{T}_{i^*i^*} - \hat{T}_{j^*j^*}| = \min_{i \neq j} |\hat{T}_{ii} - \hat{T}_{jj}| \qquad (50)$$

where the last equality is obtained by noting that, for all $i = 1, \ldots, d$, the element $\hat{T}_{ii}$ appears $d-1$ times in $\mathscr{M}$.

The norm upper bound μ in (31) obeys

$$\mu = \max_i \|M_i\| \qquad (51)$$

$$= \max_i \|X\, \mathrm{diag}(Y_{i1}, \ldots Y_{id})\, X^{-1}\| \qquad (52)$$

$$\leq \|X\| \|X^{-1}\| \max_i \|\mathrm{diag}(Y_{i1}, \ldots, Y_{id})\| \qquad (53)$$

$$\leq k(X) \sqrt{d} \max_{i,j} Y_{ij} \qquad (54)$$

$$= k(X) \sqrt{d}\, \lambda_{\max}\,. \qquad (55)$$

where X is the ground-truth matrix defined in (3) and $k(X) = \frac{\sigma_{\max}(X)}{\sigma_{\min}(X)}$ is the condition number of X.

Finally, the statement (10) follows from (31), (43), $\|\hat{x}\| = \|x\|$, (44), (49), (55) and

$$\|\Delta M\|^2 = \|\sum_i^d \theta_i \Delta M_i\|^2 \qquad (56)$$

$$= \sum_{j,k}^d \left| \sum_i^d \theta_i [\Delta M_i]_{jk} \right|^2 \qquad (57)$$

$$\leq \sum_{j,k}^d \|\theta\|^2 \sum_i^d [\Delta M_i]_{jk}^2 \qquad (58)$$

$$= \|\theta\|^2 \| \sum_i^d \Delta M_i\|^2 \qquad (59)$$

$$\leq d \|\theta\|^2 E^2 \qquad (60)$$

where we have used the Cauchy-Schwarz inequality and the definition of E. In particular, for all higher orders terms contained in Δ^2, one has $o(\|\Delta^2\|) = o(\|x\|^2) + o(\|\Delta M\|^2) = o(E^2)$. □

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2012a). Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559.

Anandkumar, A., Hsu, D., Huang, F., and Kakade, S. M. (2012b). Learning high-dimensional mixtures of graphical models. *arXiv preprint arXiv:1203.0697*.

Anandkumar, A., Hsu, D., and Kakade, S. M. (2012c). A method of moments for mixture models and hidden Markov models. *CoRR*, abs/1203.0683.

Arora, S., Ge, R., and Moitra, A. (2012). Learning topic models-going beyond SVD. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE.

Balle, B., Hamilton, W., and Pineau, J. (2014). Methods of moments for learning stochastic languages: Unified presentation and empirical comparison. In Jebara, T. and Xing, E. P., editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1386–1394. JMLR Workshop and Conference Proceedings.

Buesing, L., Sahani, M., and Macke, J. H. (2012). Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. In *Advances in neural information processing systems*, pages 1682–1690.

Chang, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math Biosci*, 137(1):51–73.

Corless, R. M., Gianni, P. M., and Trager, B. M. (1997). A reordered Schur factorization method for zero-dimensional polynomial systems with multiple roots. pages 133–140. ACM Press.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2004). Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition. *SIAM Journal on Matrix Analysis and Applications*, 26(2):295–327.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1 – 38.

Favoreel, W., De Moor, B., and Van Overschee, P. (2000). Subspace state space system identification for industrial processes. *Journal of Process Control*, 10(2):149–155.

Hsu, D. and Kakade, S. M. (2012). Learning gaussian mixture models: Moment methods and spectral decompositions. *CoRR*, abs/1206.5766.

Hsu, D., Kakade, S. M., and Zhang, T. (2012). A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460 – 1480.

Konstantinov, M. M., Petkov, P. H., and Christov, N. D. (1994). Nonlocal perturbation analysis of the Schur system of a matrix. *SIAM Journal on Matrix Analysis and Applications*, 15(2):383–392.

Kuleshov, V., Chaganty, A., and Liang, P. (2015). Tensor factorization via matrix factorization. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Lim, L. H. and Comon, P. (2009). Nonnegative approximations of nonnegative tensors. *Journal of Chemometrics*, 23(7-8):432–441.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages 1–163. JSTOR.

Mossel, E. and Roch, S. (2006). Learning nonsingular phylogenies and hidden Markov models. *The Annals of Applied Probability*, 16(2):583–614.

Song, J. and Chen, K. C. (2014). Spectacle: Faster and more accurate chromatin state annotation using spectral learning. *bioRxiv*.

Titterington, D. M. (1985). *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics. Wiley, Chichester ; New York.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. (2014). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems*, pages 1260–1268.

Zou, J. Y., Hsu, D., Parkes, D. C., and Adams, R. P. (2013). Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, pages 2238–2246.