Performance of the tensor power method for Gaussian mixtures

W. D. Brinda and Joseph T. Chang

Method of moments estimation procedures choose a model distribution whose moments match empirical moments of the data. For any distribution on $\mathbb{R}^d$, the first moments comprise a vector in $\mathbb{R}^d$. A particular value of the first moment may uniquely correspond to a model distribution if the model has fewer than d parameters. The second moments comprise a positive semi-definite matrix in $\mathbb{R}^{d\times d}$. Again, a particular combination of first and second moments may correspond to a unique model distribution if the model has few enough parameters. First and second moments are not sufficient to identify distributions within higher dimensional models, but one can then make reference to higher moments. Techniques have recently been developed to relate certain models' parameters to the generating distribution's tensor[1] of third moments and to efficiently find a model distribution corresponding approximately to a given set of first, second, and third moments.

The idea at the heart of the new tensor methods comes from Chang [1996] in the context of Markov models; the idea's generality and broader usefulness were not realized until Anandkumar et al. [2012] and Anandkumar et al. [2014]. Specifically, the tensor trick involves transforming a third-order tensor such that it becomes a sum of rank-one tensors built from orthonormal vectors that relate meaningfully to the model parameters. The method is best understood by example; Section 1 describes a tensor approach for estimating Gaussian mixtures adapted from explanations in Anandkumar et al. [2012] and [Hsu and Kakade, 2013, Sec 2]. Bounds on the statistical loss of this tensor method are derived in Section 2. Finally, Section 3 puts the method to the test using simulations. All proofs are at the end.

1 A tensor method for Gaussian mixtures

Let P be a Gaussian mixture with equal component weights, component means $(\mu_1, \ldots, \mu_K)$, and component covariances $\sigma^2 I_d$ with the number of components K no greater than the dimension d. Let μ denote the matrix comprising the component means as its column vectors. It will be convenient to give names to

[1] The concept of *tensor* generalizes the concepts of vectors and matrices. It means an array that can have any specified number of dimensions.

the sums of outer products of the component means:

$$\Psi := \mu\mu' = \sum_k \mu_k \mu_k' = \sum_k \mu_k \otimes \mu_k \qquad \text{and} \qquad \Gamma := \sum_k \mu_k \otimes \mu_k \otimes \mu_k.$$

It turns out that if Ψ and Γ are known, then it is possible to identify the component means within P, assuming they are linearly independent.[2] Let $Q\Lambda Q'$ be a spectral decomposition of Ψ with $\Lambda \in \mathbb{R}^{K\times K}$. Define the "whitening" matrix $W := \Lambda^{-1/2}Q'$; it transforms the component means into an orthonormal set $\{u_k := W\mu_k\}$. We verify orthonormality by checking that $W\mu$ is the inverse of its transpose.

$$\begin{aligned}(W\mu)(W\mu)' &= W(\mu\mu')W' \\ &= \Lambda^{-1/2}Q'(Q\Lambda Q')Q\lambda^{-1/2} \\ &= I_K\end{aligned}$$

Apply W' to each "side" of Γ to define

$$\begin{aligned}G :&= W\Gamma_{W'}W' \\ &= \sum_k (W\mu_k) \otimes (W\mu_k) \otimes (W\mu_k) \\ &= \sum_k u_k \otimes u_k \otimes u_k.\end{aligned}$$

Let vector subscripts denote application[3] into a tensor:

$$\begin{aligned}G_v :&= G \cdot v \\ &= \left[\sum_k u_k \otimes u_k \otimes u_k\right] \cdot v \\ &= \sum_k (v'u_k)u_k u_k'.\end{aligned}$$

As long as v is not orthogonal to any of the $\{u_k\}$, a spectral decomposition of G_v reveals the $\{u_k\}$ as its eigenvectors, up to sign.[4] To determine whether a sign should be reversed, compare the corresponding eigenvalue of the spectral decompositions proposal to what the eigenvalue should be: the inner product of v with the proposed u_k. If they agree, the proposed eigenvector is correct, otherwise its negative is correct.

At last, the original component means are recovered by undoing the whitening transformation $\mu_k = Q\Lambda^{1/2}u_k$.

[2] If desired, one can ensure that the unknown component means are linearly independent with probability 1 by randomly translating the space.

[3] In the context of this example, it is usually not important to keep track of which side of the tensor a vector is being multiplied into.

[4] It is easy to ensure that no eigenvectors are missed by applying enough vectors into G; for instance, any orthonormal basis of $\mathbb{R}^d$ suffices.

We have seen how knowledge of Ψ and Γ allows one to find the model GRBM. Next, we will learn how Ψ and Γ relate to the first three moments of P. The first moment is $\bar{\mu} := \mathbb{E}_{X\sim P} X = \frac{1}{K}\sum_k \mu_k$. Letting $Z \sim N(0, I_d)$, the matrix of second moments is

$$\begin{aligned}\mathbb{E}_{X\sim P} XX' &= \sum_k \tfrac{1}{K}[\mathbb{E}(\mu_k + Z)(\mu_k + Z)'] \\ &= \tfrac{1}{K}\Psi + \sigma^2 I_d,\end{aligned}$$

and the tensor of third moments has as its (d_1, d_2, d_3)-entry (with subscripts of X and Z denoting coordinates)

$$\begin{aligned}\mathbb{E}_{X\sim P} X_{d_1}X_{d_2}X_{d_3} &= \sum_k \tfrac{1}{K}[\mathbb{E}(\mu_{k,d_1} + Z_{d_1})(\mu_{k,d_2} + Z_{d_2})(\mu_{k,d_3} + Z_{d_3})] \\ &= \sum_k \tfrac{1}{K}[\mu_{k,d_1}\mu_{k,d_2}\mu_{k,d_3} + \mathbb{E}Z_{d_1}Z_{d_2}\mu_{k,d_3}\mathbb{I}_{d_1=d_2} \\ &\quad + \mathbb{E}Z_{d_1}Z_{d_3}\mu_{k,d_2}\mathbb{I}_{d_1=d_3} + \mathbb{E}Z_{d_2}Z_{d_3}\mu_{k,d_1}\mathbb{I}_{d_2=d_3}] \\ &= \sum_k \tfrac{1}{K}\mu_{k,d_1}\mu_{k,d_2}\mu_{k,d_3} \\ &\quad + \sum_k \tfrac{1}{K}\sigma^2(\mu_{k,d_3}\mathbb{I}_{d_1=d_2} + \mu_{k,d_2}\mathbb{I}_{d_1=d_3} + \mu_{k,d_1}\mathbb{I}_{d_2=d_3}) \\ &= \sum_k \tfrac{1}{K}\mu_{k,d_1}\mu_{k,d_2}\mu_{k,d_3} \\ &\quad + \sigma^2(\bar{\mu}_{d_3}\mathbb{I}_{d_1=d_2} + \bar{\mu}_{d_2}\mathbb{I}_{d_1=d_3} + \bar{\mu}_{d_1}\mathbb{I}_{d_2=d_3}).\end{aligned}$$

Notice that the first term is $1/K$ times the (d_1, d_2, d_3)-entry of Γ.

From these moment derivations, we see that with data $X_1 = (X_{1,1}, \ldots, X_{1,d}), \ldots, X_n = (X_{n,1}, \ldots, X_{n,d})$, an unbiased estimate for Ψ is

$$\widehat{\Psi} := K\left[\tfrac{1}{n}\sum_i X_i X_i' - \sigma^2 I_d\right],$$

and an unbiased estimate for the (d_1, d_2, d_3)-entry of Γ is

$$\widehat{\Gamma}_{d_1,d_2,d_3} := K\left[\tfrac{1}{n}\sum_i X_{i,d_1}X_{i,d_2}X_{i,d_3} - \sigma^2(\bar{X}_{d_3}\mathbb{I}_{d_1=d_2} + \bar{X}_{d_2}\mathbb{I}_{d_1=d_3} + \bar{X}_{d_1}\mathbb{I}_{d_2=d_3})\right]$$

with $\bar{X}$ denoting the sample mean.

One can apply the whitening and spectral decomposition procedures described above to the estimates $\widehat{\Psi}$ and $\widehat{\Gamma}$ to get estimates $\hat{\mu}_1, \ldots, \hat{\mu}_K$ of the component means. This is considered a method of moments estimator, as it corresponds to finding the model distribution whose first three moments approximately match the empirical moments.

Given a tensor that is a sum of no more than d rank-one tensors built of *orthonormal* vectors

$$G = \sum_k u_k \otimes u_k \otimes u_k,$$

we noted that the $\{u_k\}$ comprise the spectral decomposition of G_u for a uniformly random unit vector u. Anandkumar et al. [2014] also describes an alternative algorithm for finding $\{u_k\}$ called the *tensor power method*. It starts with a random u, then iterates

$$u^{(t+1)} \leftarrow \frac{G_{u^{(t)}} u^{(t)}}{\|G_{u^{(t)}} u^{(t)}\|}.$$

The u_k are fixed points. They are also maximizers of $u' G_u u$, which increases monotonically with the iterations, as shown in [Brinda, 2018, Thm 5.1.2]. Note that in practice, the estimated version of G does not have such an orthonormal internal structure.

Next, we consider the statistical quality of this estimator, in particular for the tensor power method.

2 Statistical loss of the tensor power method

[Hsu and Kakade, 2013, Sec C.3 through C.8] proved statements about the asymptotic behavior of the tensor method for mixtures of spherical Gaussians. We complement their analysis with exact finite-sample inequalities for GRBMs, some of which are specific to the tensor power method approach. An extraordinarily useful tool for us was [Yu et al., 2014, Thm 2], a variant of the Davis-Kahan theorem. A major advantage of their theorem is that it allows one to bound a difference in two matrices' eigenvectors while making reference to the eigenvalues of *only one* of the matrices involved; [Vu et al., 2013, Cor 3.1] is a closely related variant that shares this quality.

Throughout the remainder of this paper, the ordinary norm notation applied to any operator signifies the *operator norm*. Additionally, if T is a tensor, then we define $\|T\|_1$ to mean the sum of absolute values of its entries.

Let $\widehat{Q}\widehat{\Lambda}\widehat{Q}'$ be a truncated version of a spectral decomposition of $\hat{\Phi}$ which only includes the K largest eigenvalues and corresponding eigenvectors, discarding the rest; furthermore, if any of the K largest eigenvalues of $\widehat{\Psi}$ are negative, they should be replaced by zeros in $\widehat{\Lambda}$. In the end, each estimated component mean $\hat{\mu}_k$ is a product of an estimated "unwhitening" matrix $\widehat{Q}\widehat{\Lambda}^{1/2}$ with an estimated whitened vector $\hat{u}_k$, while the true component mean μ_k is the product of the true unwhitening matrix $Q\Lambda^{1/2}$ and true whitened vector u_k. Lemma 2.1 helps us bound the error of the component mean estimate in terms of the matrix and vector estimation errors.

Lemma 2.1. *Let u and $\hat{u}$ be vectors in a normed space $\mathcal{U}$ with $\|\hat{u}\| = 1$, and let M and $\widehat{M}$ be linear operators from $\mathcal{U}$ to a normed space $\mathcal{V}$. Then*

$$\|\widehat{M}\hat{u} - Mu\| \leq \|\widehat{M} - M\| + \|M\|\|\hat{u} - u\|.$$

For us, the relevant operator norm is $\|Q\Lambda^{1/2}\|$ which equals $\sqrt{\lambda_1}$, the square root of the largest eigenvalue of Ψ. Thus, Lemma 2.1 gives the bound

$$\|\hat{\mu}_k - \mu_k\| \leq \|\widehat{Q}\widehat{\Lambda}^{1/2} - Q\Lambda^{1/2}\| + \sqrt{\lambda_1}\|\hat{u}_k - u_k\|. \tag{1}$$

Since $\widehat{Q}'$ is a unit vector in a space of linear transformations, Lemma 2.1 can also be applied to $\widehat{\Lambda}^{1/2}\widehat{Q}' - \Lambda^{1/2}Q'$, implying

$$\begin{aligned}
\|\widehat{Q}\widehat{\Lambda}^{1/2} - Q\Lambda^{1/2}\| &= \|\widehat{\Lambda}^{1/2}\widehat{Q}' - \Lambda^{1/2}Q'\| \\
&\leq \|\widehat{\Lambda}^{1/2} - \Lambda^{1/2}\| + \|\Lambda^{1/2}\|\|\widehat{Q}' - Q'\| \\
&= \max_{k\in\{1,\dots K\}} |\sqrt{\hat{\lambda}_k} - \sqrt{\lambda_k}| + \sqrt{\lambda_1}\|\widehat{Q}' - Q'\|.
\end{aligned}$$

For any $a > 0$, the function $z \mapsto \sqrt{z}$ is Lipschitz with constant $1/2\sqrt{a}$ on $[a, \infty)$. So

$$|\sqrt{\hat{\lambda}_k} - \sqrt{\lambda_k}| \leq \frac{1}{2\sqrt{\lambda_K \wedge \hat{\lambda}_K}}|\hat{\lambda}_k - \lambda_k|$$

Weyl's inequality says that each eigenvalue difference is bounded by $\|\widehat{\Psi} - \Psi\|$; our truncated version $|\hat{\lambda}_k - \lambda_k|$ is no worse than the eigenvalue difference and thus shares the same bound.

The operator norm of a matrix is the same as the operator norm of its transpose, so the following lemma applies to $\|\widehat{Q}' - Q'\|$.

Lemma 2.2. *Let $\Psi \in \mathbb{R}^d$ be a symmetric positive semidefinite matrix of rank K. Let $\widehat{\Psi} \in \mathbb{R}^{d\times d}$ be a symmetric matrix with K leading eigenvectors $\widehat{Q} := (\hat{q}_1, \dots, \hat{q}_K)$. There exists a matrix $Q := (q_1, \dots, q_K)$ of ordered eigenvectors of Ψ such that*

$$\|\widehat{Q} - Q\| \leq \frac{3\sqrt{K}\|\widehat{\Psi} - \Psi\|}{\delta \wedge \lambda_K}$$

where δ is the smallest gap between consecutive positive eigenvalues of Ψ and λ_K is the Kth eigenvalue of Ψ.

Lemma 2.3. *Let $\Gamma \in \mathbb{R}^{d\times d\times d}$ and $W \in \mathbb{R}^{d\times K}$ with $K \leq d$. Suppose $W\Gamma_{W'}W' = \sum_k u_k \otimes u_k \otimes u_k$ for some orthonormal $\{u_1, \dots, u_K\}$. Let $\widehat{\Gamma} \in \mathbb{R}^{d\times d\times d}$ and $\widehat{W} \in \mathbb{R}^{d\times K}$, and suppose that $\hat{u}$ is a fixed point of the tensor power method on $\widehat{W}\widehat{\Gamma}_{\widehat{W}'}\widehat{W}'$. Then there exists $u \in \{u_1, \dots, u_K\}$ such that*

$$\begin{aligned}
&\|\hat{u} - u\| \wedge \|-\hat{u} - u\| \\
&\quad\leq \frac{12(1 + \|W\|^3)(1 + \|\Gamma\|_1)}{\|[\widehat{W}\widehat{\Gamma}_{\widehat{W}'}\widehat{W}'] \cdot \hat{u}^2\|}\left[\|\widehat{W} - W\| + \|\widehat{W} - W\|^3 + \|\widehat{\Gamma} - \Gamma\|_1(1 + \|\widehat{W} - W\|^3)\right].
\end{aligned}$$

We invoke Lemma 2.3 with $\widehat{W} = \widehat{\Lambda}^{-1/2}Q'$, and the other symbols as previously defined. $\|W\|$ is $\lambda_K^{-1/2}$. Lemma 2.1 applies to $\widehat{W} - W$ to give

$$\begin{aligned}\|\widehat{\Lambda}^{-1/2}\widehat{Q}' - \Lambda^{-1/2}Q'\| &\leq \|\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2}\| + \|\Lambda^{-1/2}\|\|\widehat{Q}' - Q'\| \\ &= \max_k |\hat{\lambda}_k^{-1/2} - \lambda_k^{-1/2}| + \lambda_K^{-1/2}\|\widehat{Q}' - Q'\|\end{aligned}$$

The norm of $\widehat{Q}' - Q'$ was addressed in Lemma 2.2. For any $a > 0$, the function $z \mapsto z^{-1/2}$ is Lipschitz with constant $1/2a^{3/2}$ on $[a, \infty)$. We use this with Weyl's inequality to bound the largest possible difference of reciprocal square root eigenvalues.

$$\begin{aligned}|\hat{\lambda}_k^{-1/2} - \lambda_k^{-1/2}| &\leq \frac{1}{2(\lambda_K \wedge \hat{\lambda}_K)^{3/2}}|\hat{\lambda}_k - \lambda_k| \\ &\leq \frac{1}{2(\lambda_K \wedge \hat{\lambda}_K)^{3/2}}\|\widehat{\Psi} - \Psi\|\end{aligned}$$

Gathering together the pieces established in this section, we can say the following.

Lemma 2.4. *Let $\mu_1, \ldots, \mu_K$ be linearly independent vectors in $\mathbb{R}^d$, and define $\Psi := \sum_k \mu_k \mu_k'$ and $\Gamma := \sum_k \mu_k \otimes \mu_k \otimes \mu_k$; let $\lambda_1, \ldots, \lambda_K$ denote the leading eigenvalues of Ψ. Let $\widehat{\Gamma} \in \mathbb{R}^{d \times d \times d}$, and let $\widehat{\Psi}$ be a symmetric matrix. Let each $\hat{\lambda}_k$ be the kth eigenvalue of $\widehat{\Psi}$ or zero if that eigenvalue is negative, and let $\widehat{\Lambda}$ denote the $K \times K$ diagonal matrix of $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$. Let $\widehat{Q} := (\hat{q}_1, \ldots, \hat{q}_K)$ comprise K leading eigenvectors of $\widehat{\Psi}$. Then for any fixed point $\hat{u}$ of the tensor power method applied to $\widehat{\Lambda}^{-1/2}\widehat{Q}'\widehat{\Gamma}_{\widehat{Q}\widehat{\Lambda}^{-1/2}}\widehat{Q}\widehat{\Lambda}^{-1/2}$, there exists $\mu \in \{\mu_1, \ldots, \mu_K\}$ such that $\hat{\mu} := \widehat{Q}\widehat{\Lambda}^{1/2}\hat{u}$ has*

$$\begin{aligned}&\|\hat{\mu} - \mu\| \wedge \|-\hat{\mu} - \mu\| \\ &\quad \leq \eta(1+\gamma)\|\widehat{\Psi} - \Psi\| + \eta^3\gamma\|\widehat{\Psi} - \Psi\|^3 + \gamma\|\widehat{\Gamma} - \Gamma\|_1 + \eta^3\gamma\|\widehat{\Gamma} - \Gamma\|_1\|\widehat{\Psi} - \Psi\|^3\end{aligned}$$

where

$$\eta := \frac{4\sqrt{K}\sqrt{\lambda_1 \vee 1}}{(\lambda_K \wedge \hat{\lambda}_K \wedge \delta \wedge 1)^{3/2}}, \qquad \gamma := \frac{12\sqrt{\lambda_1}(1 + \lambda_K^{-3/2})(1 + \|\Gamma\|_1)}{\|\widehat{\Lambda}^{-1/2}\widehat{Q}'\widehat{\Gamma}_{\widehat{Q}\widehat{\Lambda}^{-1/2}\hat{u}}\widehat{Q}\widehat{\Lambda}^{-1/2}\hat{u}\|},$$

and δ is the smallest gap between consecutive positive eigenvalues of Ψ.

Unfortunately, this approach does not address the question of whether $\hat{\mu}$ points closer to the direction of μ or $-\mu$.

Each entry of $\widehat{\Psi} - \Psi$ is the deviation of a sample average from its expectation:

$$\widehat{\Psi}_{d_1,d_2} - \Psi_{d_1,d_2} = K\left[\tfrac{1}{n}\sum_i X_{i,d_1}X_{i,d_2} - \mathbb{E}X_{d_1}X_{d_2}\right]$$

with X_{d_1} and X_{d_2} denoting the d_1 and d_2 coordinates of $X \sim P$. The operator norm of a matrix is bounded by the Frobenius norm, so a consequence of Lemma 2.4 can be stated in terms of the sum of the squared deviations. The expected Frobenius norm $\mathbb{E}\|\widehat{\Psi} - \Psi\|_F$ equals K/n times the square root of the sum of the variances of $X_{d_1}X_{d_2}$.

Similarly, the entries of $\|\widehat{\Gamma} - \Gamma\|_1$ are bounded in terms of deviations of sample averages from their expectation:

$$\widehat{\Gamma}_{d_1,d_2,d_3} - \Gamma_{d_1,d_2,d_3} = K[\tfrac{1}{n}\sum_i X_{i,d_1}X_{i,d_2}X_{i,d_3} - \mathbb{E}X_{d_1}X_{d_2}X_{d_3} \\ - \sigma^2[(\bar{X}_{d_3} - \mu_{d_3})\mathbb{I}_{d_1=d_2} + (\bar{X}_{d_2} - \mu_{d_2})\mathbb{I}_{d_1=d_3} + (\bar{X}_{d_1} - \mu_{d_1})\mathbb{I}_{d_2=d_3}]].$$

$\mathbb{E}\|\widehat{\Gamma} - \Gamma\|_1$ also has a bound with order $1/n$.

An exact overall risk bound is still challenging, however, because there is an error cross-term and there are also the random quantities $\hat{\lambda}_d$ and $\widehat{G}_{\hat{u}}\hat{u}$. While random, both are known by the practitioner, so loss bounds involving them can still be of interest.

Lemmas and ideas from [Hsu and Kakade, 2013, Sec C.4 through C.8] could be employed to devise probabilistic loss bounds that replace $\|\widehat{\Psi} - \Psi\|_F^3$ and $\|\widehat{\Gamma} - \Gamma\|_1$ with functions of the true parameters of P.

3 Simulations

DO SIMULATIONS AND PUT THEM HERE

Proofs

We first establish two handy general-purpose lemmas about tensors.[5]

Lemma 3.1. *Let $T := a \otimes b \otimes c$ with vectors a, b, c each belonging to an inner product space. Then any of $T(v, \cdot, \cdot)$, $T(\cdot, v, \cdot)$, or $T(\cdot, \cdot, v)$ that are well-defined all have operator norms bounded by $\|a\|\|b\|\|c\|\|v\|$.*

Proof. Assume the dimension of v matches that of a, and observe that $T(v, \cdot, \cdot) = (a'v)bc'$. Using Cauchy-Schwarz,

$$\begin{aligned}\|T(v, \cdot, \cdot)\| &= \|(a'v)bc'\| \\ &= |a'v|\|bc'\| \\ &\leq \|a\|\|v\|\|bc'\| \\ &= \|a\|\|v\|\|b\|\|c\|.\end{aligned}$$

To justify the last step, consider multiplying $bc'u$ where u is a unit vector. By Cauchy-Schwarz, $|c'u| \leq \|c\|$, so the Euclidean norm of $bc'u$ is bounded by

[5] In fact, all of the lemmas in this section are written abstractly and may potentially apply to other tensor method problems with little or no modification.

$\|c\|\|b\|$. And letting $u = c$ shows that this bound is achievable, so $\|bc'\|_2 = \|b\|\|c\|$.

Because the bound has v related in the same way to each of a, b, and c, we see that the same bound must result no matter which side of T we apply v into. □

We derive a simple consequence of Lemma 3.1 using the fact that any tensor in $\mathbb{R}^{d_1 \times d_2 \times d_3}$ can be expressed as a sum of canonical rank-one basis tensors.

Lemma 3.2. *Let T be a tensor in $\mathbb{R}^{d_1 \times d_2 \times d_3}$. Then any well-defined application of a vector v into T has operator norm bounded by $\|v\|\|T\|_1$.*

Proof. We can express T in terms of the canonical basis unit vectors as $\sum_{i,j,k} T_{i,j,k}\, e_i \otimes e_j \otimes e_k$.

$$\begin{aligned}
\|T \cdot v\| &= \left\| \left[\sum_{i,j,k} T_{i,j,k}\, e_i \otimes e_j \otimes e_k \right] \cdot v \right\| \\
&= \left\| \sum_{i,j,k} [T_{i,j,k}\, e_i \otimes e_j \otimes e_k] \cdot v \right\| \\
&\leq \sum_{i,j,k} \|[T_{i,j,k}\, e_i \otimes e_j \otimes e_k] \cdot v\| \\
&= \sum_{i,j,k} |T_{i,j,k}| \|[e_i \otimes e_j \otimes e_k] \cdot v\| \\
&\leq \sum_{i,j,k} |T_{i,j,k}| \|e_i\| \|e_j\| \|e_k\| \|v\| \qquad \text{by Lemma 3.1} \\
&= \|v\| \sum_{i,j,k} |T_{i,j,k}|
\end{aligned}$$

□

Proof of Lemma 2.1. First, expand $\widehat{M}\hat{u}$.

$$\begin{aligned}
\widehat{M}\hat{u} &= [M + (\widehat{M} - M)]\hat{u} \\
&= M\hat{u} + (\widehat{M} - M)\hat{u} \\
&= Mu + M(\hat{u} - u) + (\widehat{M} - M)\hat{u}.
\end{aligned}$$

Use the triangle inequality to bound the difference between $\widehat{M}\hat{u}$ and Mu by the sum of the norms of the two right-most terms. Bound those norms by the operator norms times vector norms, and use $\|\hat{u}\| = 1$. □

Proof of Lemma 2.2. First, consider each $\hat{q}_k - q_k$ individually in light of [Yu et al., 2014, Cor 1]. If $\hat{q}_k' q_k \geq 0$, then

$$\|\hat{q}_k - q_k\| \leq \frac{3\|\widehat{\Psi} - \Psi\|}{\delta \wedge \lambda_K}$$

Let the kth column of Q be the eigenvector pointing in the "right" direction.

We can bound $\|\widehat{Q} - Q\|$ by bounding the norm of $(\widehat{Q} - Q)v$ for an arbitrary unit vector $v \in \mathbb{R}^K$.

$$\begin{aligned}
\|(\widehat{Q} - Q)v\| &= \|\sum_k v_k(\hat{q}_k - q_k)\| \\
&\leq \sum_k \|v_k(\hat{q}_k - q_k)\| \\
&= \sum_k |v_k| \|\hat{q}_k - q_k\| \\
&\leq \sum_k |v_k| \left(\frac{3\|\widehat{\Psi} - \Psi\|}{\delta \wedge \lambda_K} \right) \\
&= \frac{3\|\widehat{\Psi} - \Psi\|}{\delta \wedge \lambda_K} \|v\| \sqrt{K}
\end{aligned}$$

Note that if $K = d$, then our use of [Yu et al., 2014, Cor 1] does not need λ_K in the denominator. □

Proof of Lemma 2.3. Because $\hat{u}$ is a fixed point of the tensor power method on $\widehat{G}$, we know that $\widehat{G}_{\hat{u}}$ must be proportional to $\hat{u}\hat{u}'$; its principal eigenvalue is $\|\widehat{G}_{\hat{u}}\hat{u}\|$ and all other eigenvalues are zero. Define u to be the leading eigenvector of $G_{\hat{u}} := W\Gamma_{W'\hat{u}}W' = \sum_k (\hat{u}'u_k)u_k u_k'$; it is one of the $\{u_1, \ldots, u_K\}$. By [Yu et al., 2014, Cor 1],

$$\|\hat{u} - u\| \wedge \| - \hat{u} - u\| \leq \frac{3\|\widehat{G}_{\hat{u}} - G_{\hat{u}}\|}{\|\widehat{G}_{\hat{u}}\hat{u}\|}.$$

It remains to bound the operator norm of $\widehat{G}_{\hat{u}} - G_{\hat{u}}$. Expand $\widehat{G}_{\hat{u}}$ after writing $\widehat{W}$ as $W + (\widehat{W} - W)$ and $\widehat{\Gamma}$ as $\Gamma + (\widehat{\Gamma} - \Gamma)$.

$$\begin{aligned}
\widehat{G}_{\hat{u}} &= \widehat{W}\widehat{\Gamma}_{\widehat{W}'\hat{u}}\widehat{W}' \\
&= G_{\hat{u}} + (\widehat{W} - W)\Gamma_{W'\hat{u}}W' + W\Gamma_{(\widehat{W}-W)'\hat{u}}W' + W\Gamma_{W'\hat{u}}(\widehat{W} - W)' \\
&\quad + (\widehat{W} - W)\Gamma_{(\widehat{W}-W)'\hat{u}}W' + (\widehat{W} - W)\Gamma_{W'\hat{u}}(\widehat{W} - W)' + W\Gamma_{(\widehat{W}-W)'\hat{u}}(\widehat{W} - W)' \\
&\quad + (\widehat{W} - W)\Gamma_{(\widehat{W}-W)'\hat{u}}(\widehat{W} - W)' + W(\Gamma - \widehat{\Gamma})_{W'\hat{u}}W' + (\widehat{W} - W)(\Gamma - \widehat{\Gamma})_{W'\hat{u}}W' \\
&\quad + W(\Gamma - \widehat{\Gamma})_{(\widehat{W}-W)'\hat{u}}W' + W(\Gamma - \widehat{\Gamma})_{W'\hat{u}}(\widehat{W} - W)' \\
&\quad + (\widehat{W} - W)(\Gamma - \widehat{\Gamma})_{(\widehat{W}-W)'\hat{u}}W' + (\widehat{W} - W)(\Gamma - \widehat{\Gamma})_{W'\hat{u}}(\widehat{W} - W)' \\
&\quad + W(\Gamma - \widehat{\Gamma})_{(\widehat{W}-W)'\hat{u}}(\widehat{W} - W)' + (\widehat{W} - W)(\Gamma - \widehat{\Gamma})_{(\widehat{W}-W)'\hat{u}}(\widehat{W} - W)'
\end{aligned}$$

Lemma 3.2 can be used to bound each term. The first error term, for in-

stance, has norm

$$\begin{aligned}\|(\widehat{W}-W)\Gamma_{W'\hat{u}}W'\| &\le \|\widehat{W}-W\|\|\Gamma_{W'\hat{u}}\|\|W'\| \\ &\le \|\widehat{W}-W\|\|W\|\|W'\hat{u}\|\|\Gamma\|_1 \\ &\le \|\widehat{W}-W\|\|W\|^2\|\Gamma\|_1.\end{aligned}$$

Repeat this process on each of the fifteen error terms. An upper bound that works for the sum of norms is

$$\begin{aligned}&(1+\|W\|^3)(1+\|\Gamma\|_1)[3\|\widehat{W}-W\|+3\|\widehat{W}-W\|^2+\|\widehat{W}-W\|^3 \\ &\quad+\|\widehat{\Gamma}-\Gamma\|_1+3\|\widehat{W}-W\|\|\widehat{\Gamma}-\Gamma\|_1+3\|\widehat{W}-W\|^2\|\widehat{\Gamma}-\Gamma\|_1+\|\widehat{W}-W\|^3\|\widehat{\Gamma}-\Gamma\|_1].\end{aligned}$$

The sum of the first three terms in the square brackets can be upper bounded using the inequality $3z+3z^2+z^3 \le 4(z+z^3)$ for $z \ge 0$, which one can easily verify. For the remaining four terms in the square brackets, use the fact that $1+3z+3z^2+z^3 \le 4(1+z^3)$ for $z \ge 0$. These steps simplify the bound while loosening it. □

Proof of Lemma 2.4. From (1) and Lemma 2.2 along with the surrounding discussion,

$$\begin{aligned}\|\hat{\mu}-\mu\| &\le \frac{1}{2\sqrt{\lambda_K \wedge \hat{\lambda}_K}}\|\widehat{\Psi}-\Psi\|+\sqrt{\lambda_1}\|\widehat{Q}'-Q'\|+\sqrt{\lambda_1}\|\hat{u}-u\| \\ &\le \frac{1}{2\sqrt{\lambda_K \wedge \hat{\lambda}_K}}\|\widehat{\Psi}-\Psi\|+\frac{3\sqrt{K\lambda_1}}{\delta \wedge \lambda_K}\|\widehat{\Psi}-\Psi\|+\sqrt{\lambda_1}\|\hat{u}-u\| \qquad (2)\end{aligned}$$

where Λ is the $K \times K$ diagonal eigenvalue matrix of Ψ, Q is an accompanying eigenvector matrix with directions chosen such that every $q_k'\hat{q}_k \ge 0$, and vectors $\hat{u}$ and u are the whitened versions of $\hat{\mu}$ and μ. The same inequality can be stated replacing $\hat{\mu}$ and $\hat{u}$ with $-\hat{\mu}$ and $-\hat{u}$. Lemma 2.3 and the comments afterward apply to one of these two inequalities, bounding the error in the estimation of u in terms of $W := \Lambda^{-1/2}Q'$ and $\widehat{W} := \widehat{\Lambda}^{-1/2}\widehat{Q}'$. Again using Lemma 2.2 for $\|\widehat{Q}'-Q'\|$,

$$\begin{aligned}\|\widehat{W}-W\| &\le \frac{1}{2(\lambda_K \wedge \hat{\lambda}_K)^{3/2}}\|\widehat{\Psi}-\Psi\|+\frac{1}{\sqrt{\lambda_K}}\frac{3\sqrt{K\lambda_1}}{\delta \wedge \lambda_K}\|\widehat{\Psi}-\Psi\| \\ &\le \eta\|\widehat{\Psi}-\Psi\|\end{aligned}$$

since the coefficients of $\|\widehat{\Psi}-\Psi\|$ are bounded by $\eta/4$ and $3\eta/4$. The first two terms in (2) also have coefficients of $\|\widehat{\Psi}-\Psi\|$ bounded by $\eta/4$ and $3\eta/4$. These bounds are loose but allow for a relatively straight-forward overall statement.

Finally, γ is the product of the $\sqrt{\lambda_1}$ from (1) with the first factor from Lemma 2.3. □

References

Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pages 33–1, 2012.

Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

W. D. Brinda. *Adaptive Estimation with Gaussian Radial Basis Mixtures*. PhD thesis, Yale University, 2018.

Joseph T Chang. Full reconstruction of markov models on evolutionary trees: identifiability and consistency. *Mathematical biosciences*, 137(1):51–73, 1996.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in neural information processing systems*, pages 2670–2678, 2013.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2014.