Learning mixtures of spherical Gaussians: moment methods and spectral decompositions

Daniel Hsu and Sham M. Kakade

Microsoft Research New England

November 4, 2018

Abstract

This work provides a computationally efficient and statistically consistent moment-based estimator for mixtures of spherical Gaussians. Under the condition that component means are in general position, a simple spectral decomposition technique yields consistent parameter estimates from low-order observable moments, without additional minimum separation assumptions needed by previous computationally efficient estimation procedures. Thus computational and information-theoretic barriers to efficient estimation in mixture models are precluded when the mixture components have means in general position and spherical covariances. Some connections are made to estimation problems related to independent component analysis.

1 Introduction

The Gaussian mixture model (Pearson, 1894; Titterington et al., 1985) is one of the most well-studied and widely-used models in applied statistics and machine learning. An important special case of this model (the primary focus of this work) restricts the Gaussian components to have spherical covariance matrices; this probabilistic model is closely related to the (non-probabilistic) k-means clustering problem (MacQueen, 1967).

The mixture of spherical Gaussians model is specified as follows. Let w_i be the probability of choosing component $i \in [k] := \{1, 2, \ldots, k\}$, let $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ be the component mean vectors, and let $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2 \geq 0$ be the component variances. Define

$$w := [w_1, w_2, \ldots, w_k]^\top \in \mathbb{R}^k, \qquad A := [\mu_1 | \mu_2 | \cdots | \mu_k] \in \mathbb{R}^{d \times k};$$

so w is a probability vector, and A is the matrix whose columns are the component means. Let $x \in \mathbb{R}^k$ be the (observed) random vector given by

$$x := \mu_h + z,$$

where h is the discrete random variable with $\Pr(h = i) = w_i$ for $i \in [k]$, and z is a random vector whose conditional distribution given $h = i$ (for some $i \in [k]$) is the multivariate Gaussian $\mathcal{N}(0, \sigma_i^2 I)$ with mean zero and covariance $\sigma_i^2 I$.

E-mail: {dahsu,skakade}@microsoft.com

The estimation task is to accurately recover the model parameters (component means, variances, and mixing weights) $\{(\mu_i, \sigma_i^2, w_i) : i \in [k]\}$ from independent copies of x.

This work gives a procedure for efficiently and exactly recovering the parameters using a simple spectral decomposition of low-order moments of x, under the following condition.

Condition 1 (Non-degeneracy). The component means span a k-dimensional subspace (*i.e.*, the matrix A has column rank k), and the vector w has strictly positive entries.

The proposed estimator is based on a spectral decomposition technique (Chang, 1996; Mossel and Roch, 2006; Anandkumar et al., 2012b), and is easily stated in terms of exact population moments of the observed x. With finite samples, one can use a plug-in estimator based on empirical moments of x in place of exact moments. These empirical moments converge to the exact moments at a rate of $O(n^{-1/2})$, where n is the sample size. As discussed in Section 3, sample complexity bounds for accurate parameter estimation can be derived using matrix perturbation arguments (Anandkumar et al., 2012b). Since only low-order moments are required by the plug-in estimator, the sample complexity is polynomial in the relevant parameters of the estimation problem.

Related work. The first estimators for the Gaussian mixture models were based on the method-of-moments, as introduced by Pearson (1894) (see also Lindsay and Basak, 1993, and the references therein). Roughly speaking, these estimators are based on finding parameters under which the Gaussian mixture distribution has moments approximately matching the observed empirical moments. Finding these parameters typically involves solving systems of multivariate polynomial equations, which is typically computationally challenging. Besides this, the order of the moments of some of the early moment-based estimators were either growing with the dimension d or the number of components k, which is undesirable because the empirical estimates of such high-order moments may only be reliable when the sample size is exponential in d or k. Both the computational and sample complexity issues have been addressed in recent years, at least under various restrictions. For instance, several distance-based estimators require that the component means be well-separated in Euclidean space, by at least some large factor times the directional standard deviation of the individual component distributions (Dasgupta, 1999; Arora and Kannan, 2001; Dasgupta and Schulman, 2007; Vempala and Wang, 2002; Chaudhuri and Rao, 2008), but otherwise have polynomial computational and sample complexity. Some recent moment-based estimators avoid the minimum separation condition of distance-based estimators by requiring either computational or data resources exponential in the number of mixing components k (but not the dimension d) (Belkin and Sinha, 2010; Kalai et al., 2010; Moitra and Valiant, 2010) or by making a non-degenerate multi-view assumption (Anandkumar et al., 2012b).

By contrast, the moment-based estimator described in this work does not require a minimum separation condition, exponential computational or data resources, or non-degenerate multiple views. Instead, it relies only on the non-degeneracy condition discussed above together with a spherical noise condition. The non-degeneracy condition is much weaker than an explicit minimum separation condition because the parameters can be arbitrarily close to being degenerate, as long as the sample size grows polynomially with a natural quantity measuring this closeness to degeneracy (akin to a condition number). Like other moment-based estimators, the proposed estimator is based on solving multivariate polynomial equations, although these solutions can be found efficiently because the problems are cast as eigenvalue decompositions of symmetric matrices, which are efficient to compute.

Recent work of Moitra and Valiant (2010) demonstrates an information-theoretic barrier to estimation for general Gaussian mixture models. More precisely, they construct a pair of one-dimensional mixtures of Gaussians (with separated component means) such that the statistical distance between the two mixture distributions is exponentially small in the number of components. This implies that in the worst case, the sample size required to obtain accurate parameter estimates must grow exponentially with the number of components, even when the component distributions are non-negligibly separated. A consequence of the present work is that natural non-degeneracy conditions preclude these worst case scenarios. The non-degeneracy condition in this work is similar to one used for bypassing computational (cryptographic) barriers to estimation for hidden Markov models (Chang, 1996; Mossel and Roch, 2006; Hsu et al., 2012a; Anandkumar et al., 2012b).

Finally, it is interesting to note that similar algebraic techniques have been developed for certain models in independent component analysis (ICA) (Comon, 1994; Cardoso and Comon, 1996; Hyvärinen and Oja, 2000; Comon and Jutten, 2010; Arora et al., 2012) and other closely related problems (Frieze et al., 1996; Nguyen and Regev, 2009). In contrast to the ICA setting, handling non-spherical Gaussian noise for mixture models appears to be a more delicate issue. These connections and open problems are further discussed in Section 3.

2   Moment-based estimation

This section describes a method-of-moments estimator for the spherical Gaussian mixture model.

The following theorem is the main structural result that relates the model parameters to observable moments.

Theorem 1 (Observable moment structure). *Assume Condition 1 holds. The average variance* $\bar{\sigma}^2 := \sum_{i=1}^{k} w_i \sigma_i^2$ *is the smallest eigenvalue of the covariance matrix* $\mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^\top]$*. Let* $v \in \mathbb{R}^d$ *be any unit norm eigenvector corresponding to the eigenvalue* $\bar{\sigma}^2$*. Define*

$$M_1 := \mathbb{E}[x(v^\top(x - \mathbb{E}[x]))^2] \in \mathbb{R}^d,$$
$$M_2 := \mathbb{E}[x \otimes x] - \bar{\sigma}^2 I \in \mathbb{R}^{d \times d},$$
$$M_3 := \mathbb{E}[x \otimes x \otimes x] - \sum_{i=1}^{d}\big(M_1 \otimes e_i \otimes e_i + e_i \otimes M_1 \otimes e_i + e_i \otimes e_i \otimes M_1\big) \in \mathbb{R}^{d \times d \times d}$$

(where $\otimes$ denotes tensor product, and $\{e_1, e_2, \ldots, e_d\}$ is the coordinate basis for $\mathbb{R}^d$). Then

$$M_1 = \sum_{i=1}^{k} w_i\ \sigma_i^2 \mu_i, \qquad M_2 = \sum_{i=1}^{k} w_i\ \mu_i \otimes \mu_i, \qquad M_3 = \sum_{i=1}^{k} w_i\ \mu_i \otimes \mu_i \otimes \mu_i.$$

Remark 1. We note that in the special case where $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ (*i.e.*, the mixture components share a common spherical covariance matrix), the average variance $\bar{\sigma}^2$ is simply σ^2, and M_3 has a simpler form:

$$M_3 = \mathbb{E}[x \otimes x \otimes x] - \sigma^2 \sum_{i=1}^{d}\big(\mathbb{E}[x] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[x] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[x]\big) \in \mathbb{R}^{d \times d \times d}.$$

There is no need to refer to the eigenvectors of the covariance matrix or M_1.

Proof of Theorem 1. We first characterize the smallest eigenvalue of the covariance matrix of x, as well as all corresponding eigenvectors v. Let $\bar{\mu} := \mathbb{E}[x] = \mathbb{E}[\mu_h] = \sum_{i=1}^k w_i \mu_i$. The covariance matrix of x is

$$\mathbb{E}[(x - \bar{\mu}) \otimes (x - \bar{\mu})] = \sum_{i=1}^k w_i \left((\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu}) + \sigma_i^2 I \right)$$

$$= \sum_{i=1}^k w_i \, (\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu}) \, + \, \bar{\sigma}^2 I.$$

Since the vectors $\mu_i - \bar{\mu}$ for $i \in [k]$ are linearly dependent ($\sum_{i=1}^k w_i(\mu_i - \bar{\mu}) = 0$), the positive semidefinite matrix $\sum_{i=1}^k w_i(\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu})$ has rank $r \leq k-1$. Thus, the $d-r$ smallest eigenvalues are exactly $\bar{\sigma}^2$, while all other eigenvalues are strictly larger than $\bar{\sigma}^2$. The strict separation of eigenvalues implies that every eigenvector corresponding to $\bar{\sigma}^2$ is in the null space of $\sum_{i=1}^k w_i(\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu})$; thus $v^\top (\mu_i - \bar{\mu}) = 0$ for all $i \in [k]$.

Now we can express M_1, M_2, and M_3 in terms of the parameters w_i, μ_i, and σ_i^2. First,

$$M_1 = \mathbb{E}[x(v^\top(x - \mathbb{E}[x]))^2] = \mathbb{E}[(\mu_h + z)(v^\top(\mu_h - \bar{\mu} + z))^2] = \mathbb{E}[(\mu_h + z)(v^\top z)^2] = \mathbb{E}[\mu_h \sigma_h^2],$$

where the last step uses the fact that $z|h \sim \mathcal{N}(0, \sigma_h^2 I)$, which implies that conditioned on h, $\mathbb{E}[(v^\top z)^2|h] = \sigma_h^2$ and $\mathbb{E}[z(v^\top z)^2|h] = 0$. Next, observe that $\mathbb{E}[z \otimes z] = \sum_{i=1}^k w_i \sigma_i^2 I = \bar{\sigma}^2 I$, so

$$M_2 = \mathbb{E}[x \otimes x] - \bar{\sigma}^2 I = \mathbb{E}[\mu_h \otimes \mu_h] + \mathbb{E}[z \otimes z] - \bar{\sigma}^2 I = \mathbb{E}[\mu_h \otimes \mu_h] = \sum_{i=1}^k w_i \, \mu_i \otimes \mu_i.$$

Finally, for M_3, we first observe that

$$\mathbb{E}[x \otimes x \otimes x] = \mathbb{E}[\mu_h \otimes \mu_h \otimes \mu_h] + \mathbb{E}[\mu_h \otimes z \otimes z] + \mathbb{E}[z \otimes \mu_h \otimes z] + \mathbb{E}[z \otimes z \otimes \mu_h]$$

(terms such as $\mathbb{E}[\mu_h \otimes \mu_h \otimes z]$ and $\mathbb{E}[z \otimes z \otimes z]$ vanish because $z|h \sim \mathcal{N}(0, \sigma_h^2 I)$). We now claim that $\mathbb{E}[\mu_h \otimes z \otimes z] = \sum_{i=1}^d M_1 \otimes e_i \otimes e_i$. This holds because

$$\mathbb{E}[\mu_h \otimes z \otimes z] = \mathbb{E}\big[\mathbb{E}[\mu_h \otimes z \otimes z|h]\big]$$

$$= \mathbb{E}\left[\mathbb{E}\bigg[\sum_{i,j=1}^d z_i z_j \, \mu_h \otimes e_i \otimes e_j \Big| h\bigg]\right] = \mathbb{E}\left[\sum_{i=1}^d \sigma_h^2 \, \mu_h \otimes e_i \otimes e_i\right] = \sum_{i=1}^d M_1 \otimes e_i \otimes e_i,$$

crucially using the fact that $\mathbb{E}[z_i z_j|h] = 0$ for $i \neq j$ and $\mathbb{E}[z_i^2|h] = \sigma_h^2$. By the same derivation, we have $\mathbb{E}[z \otimes \mu_h \otimes z] = \sum_{i=1}^d e_i \otimes M_1 \otimes e_i$ and $\mathbb{E}[z \otimes z \otimes \mu_h] = \sum_{i=1}^d e_i \otimes e_i \otimes M_1$. Therefore,

$$M_3 = \mathbb{E}[x \otimes x \otimes x] - \big(\mathbb{E}[\mu_h \otimes z \otimes z] + \mathbb{E}[z \otimes \mu_h \otimes z] + \mathbb{E}[z \otimes z \otimes \mu_h]\big) = \mathbb{E}[\mu_h \otimes \mu_h \otimes \mu_h] = \sum_{i=1}^k w_i \mu_i \otimes \mu_i \otimes \mu_i$$

as claimed. $\qquad\square$

Theorem 1 shows the relationship between (some functions of) the observable moments and the desired parameters. A simple estimator based on this moment structure is given in the following theorem. For a third-order tensor $T \in \mathbb{R}^{d \times d \times d}$, we define $T(\eta) := \sum_{i_1=1}^d \sum_{i_2=1}^d \sum_{i_3=1}^d T_{i_1, i_2, i_3} \eta_{i_3} \, e_{i_1} \otimes e_{i_2} \in \mathbb{R}^{d \times d}$ for any vector $\eta \in \mathbb{R}^d$.

Theorem 2 (Moment-based estimator). *The following can be added to the results of Theorem 1. Suppose $\eta^\top\mu_1, \eta^\top\mu_2, \ldots, \eta^\top\mu_k$ are distinct and non-zero (which is satisfied almost surely, for instance, if η is chosen uniformly at random from the unit sphere in $\mathbb{R}^d$). Then the matrix*

$$M_{\mathrm{GMM}}(\eta) := M_2^{\dagger 1/2} M_3(\eta) M_2^{\dagger 1/2}$$

is diagonalizable (where $\dagger$ denotes the Moore-Penrose pseudoinverse); its non-zero eigenvalue / eigenvector pairs $(\lambda_1, v_1), (\lambda_2, v_2), \ldots, (\lambda_k, v_k)$ satisfy $\lambda_i = \eta^\top\mu_{\pi(i)}$ and $M_2^{1/2}v_i = s_i\sqrt{w_{\pi(i)}}\mu_{\pi(i)}$ for some permutation π on $[k]$ and signs $s_1, s_2, \ldots, s_k \in \{\pm 1\}$. The μ_i, σ_i^2, and w_i are recovered (up to permutation) with

$$\mu_{\pi(i)} = \frac{\lambda_i}{\eta^\top M_2^{1/2}v_i}M_2^{1/2}v_i, \qquad \sigma_i^2 = \frac{1}{w_i}e_i^\top A^\dagger M_1, \qquad w_i = e_i^\top A^\dagger \mathbb{E}[x].$$

Proof. By Theorem 1,

$$M_1 = A\operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)w, \qquad M_2 = A\operatorname{diag}(w)A^\top, \qquad M_3(\eta) = A\operatorname{diag}(w)D_1(\eta)A^\top,$$

where $D_1(\eta) := \operatorname{diag}(\eta^\top\mu_1, \eta^\top\mu_2, \ldots, \eta^\top\mu_k)$.

Let $USR^\top$ be the thin SVD of $A\operatorname{diag}(w)^{1/2}$ ($U \in \mathbb{R}^{d\times k}$, $S \in \mathbb{R}^{k\times k}$, and $R \in \mathbb{R}^{k\times k}$), so $M_2 = US^2U^\top$ and $M_2^{\dagger 1/2} = US^{-1}U^\top$ since $A\operatorname{diag}(w)^{1/2}$ has rank k by assumption. Also by assumption, the diagonal entries of $D_1(\eta)$ are distinct and non-zero. Therefore, every non-zero eigenvalue of the symmetric matrix $M_{\mathrm{GMM}}(\eta) = UR^\top D_1(\eta)RU^\top$ has geometric multiplicity one. Indeed, these non-zero eigenvalues λ_i are the diagonal entries of $D_1(\eta)$ (up to some permutation π on $[k]$), and the corresponding eigenvectors v_i are the columns of $UR^\top$ up to signs:

$$\lambda_i = \eta^\top\mu_{\pi(i)} \quad\text{and}\quad v_i = s_iUR^\top e_{\pi(i)}.$$

Now, since

$$M_2^{1/2}v_i = s_i\sqrt{w_{\pi(i)}}\mu_{\pi(i)}, \qquad \frac{\lambda_i}{\eta^\top M_2^{1/2}v_i} = \frac{\eta^\top\mu_{\pi(i)}}{s_i\sqrt{w_{\pi(i)}}\eta^\top\mu_{\pi(i)}} = \frac{1}{s_i\sqrt{w_{\pi(i)}}},$$

it follows that

$$\mu_{\pi(i)} = \frac{\lambda_i}{\eta^\top M_2^{1/2}v_i}M_2^{1/2}v_i, \quad i \in [k].$$

The claims regarding σ_i^2 and w_i are also evident from the structure of M_1 and $\mathbb{E}[x] = Aw$. $\qquad\square$

An efficiently computable plug-in estimator can be derived from Theorem 2. We state one such algorithm (called LEARNGMM) in Appendix C; for simplicity, we restrict to the case where the components share the same common spherical covariance, *i.e.*, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$. The following theorem provides a sample complexity bound for accurate estimation of the component means. Since only low-order moments are used, the sample complexity is polynomial in the relevant parameters of the estimation problem (in particular, the dimension d and the number of mixing components k). It is worth noting that the polynomial is quadratic in the inverse accuracy parameter $1/\varepsilon$; this owes to the fact that the empirical moments converge to the population moments at the usual $n^{-1/2}$ rate as per the central limit theorem.

Theorem 3 (Finite sample bound). *There exists a polynomial* poly(·) *such that the following holds. Let M_2 be the matrix defined in Theorem 2, and $\varsigma_t[M_2]$ be its t-th largest singular value (for $t \in [k]$). Let $b_{\max} := \max_{i \in [k]} \|\mu_i\|_2$ and $w_{\min} := \min_{i \in [k]} w_i$. Pick any $\varepsilon, \delta \in (0,1)$. Suppose the sample size n satisfies*

$$n \geq \text{poly}\Big(d, k, 1/\varepsilon, \log(1/\delta), 1/w_{\min}, \varsigma_1[M_2]/\varsigma_k[M_2], b_{\max}^2/\varsigma_k[M_2], \sigma^2/\varsigma_k[M_2], \Big).$$

Then with probability at least $1 - \delta$ over the random sample and the internal randomness of the algorithm, there exists a permutation π on $[k]$ such that the $\{\hat{\mu}_i : i \in [k]\}$ returned by LEARNGMM *satisfy*

$$\|\hat{\mu}_{\pi(i)} - \mu_i\|_2 \leq \Big(\|\mu_i\|_2 + \sqrt{\varsigma_1[M_2]}\Big)\varepsilon$$

for all $i \in [k]$.

The proof of Theorem 3 is given in Appendix C. It is also easy to obtain accuracy guarantees for estimating σ^2 and w. The role of Condition 1 enters by observing that $\varsigma_k[M_2] = 0$ if either $\text{rank}(A) < k$ or $w_{\min} = 0$, as $M_2 = A \text{diag}(w)A^\top$. The sample complexity bound then becomes trivial in this case, as the bound grows with $1/\varsigma_k[M_2]$ and $1/w_{\min}$. Finally, we also note that LEARNGMM is just one (easy to state) way to obtain an efficient algorithm based on the structure in Theorem 1. It is also possible to use, for instance, simultaneous diagonalization techniques (Bunse-Gerstner et al., 1993) or orthogonal tensor decompositions (Anandkumar et al., 2012a) to extract the parameters from (estimates of) M_2 and M_3; these alternative methods are more robust to sampling error, and are therefore recommended for practical implementation.

3 Discussion

Multi-view methods and a simpler algorithm in higher dimensions. Some previous work of the authors on moment-based estimators for the Gaussian mixture model relies on a non-degenerate multi-view assumption (Anandkumar et al., 2012b). In this work, it is shown that if each mixture component i has an axis-aligned covariance $\Sigma_i := \text{diag}(\sigma_{1,i}^2, \sigma_{2,i}^2, \ldots, \sigma_{d,i}^2)$, then under some additional mild assumptions (which ultimately require $d > k$), a moment-based method can be used to estimate the model parameters. The idea is to partition the coordinates $[d]$ into three groups, inducing multiple "views" $x = (x_1, x_2, x_3)$ with each $x_t \in \mathbb{R}^{d_t}$ for some $d_t \geq k$ such that x_1, x_2, and x_3 are conditionally independent given h. When the matrix of conditional means $A_t := [\mathbb{E}[x_t|h = 1]|\mathbb{E}[x_t|h = 2]|\cdots|\mathbb{E}[x_t|h = k]] \in \mathbb{R}^{d_t \times k}$ for each view $t \in \{1, 2, 3\}$ has rank k, then an efficient technique similar to that described in Theorem 2 will recover the parameters. Therefore, the problem is reduced to partitioning the coordinates so that the resulting matrices A_t have rank k.

In the case where each component covariance is spherical ($\Sigma_i = \sigma_i^2 I$), we may simply apply a random rotation to x before (arbitrarily) splitting into the three views. Let $\tilde{x} := \Theta x$ for a random orthogonal matrix $\Theta \in \mathbb{R}^{d \times d}$, and partition the coordinates so that $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ with $\tilde{x}_t \in \mathbb{R}^{d_t}$ and $d_t \geq k$. By the rotational invariance of the multivariate Gaussian distribution, the distribution of $\tilde{x}$ is still a mixture of spherical Gaussians, and moreover, the matrix of conditional means $\tilde{A}_t := [\mathbb{E}[\tilde{x}_t|h = 1]|\mathbb{E}[\tilde{x}_t|h = 2]|\cdots|\mathbb{E}[\tilde{x}_t|h = k]] \in \mathbb{R}^{d_t \times k}$ for each view $\tilde{x}_t$ has rank k with probability 1. To see this, observe that a random rotation in $\mathbb{R}^d$ followed by a restriction to d_t

coordinates is simply a random projection from $\mathbb{R}^d$ to $\mathbb{R}^{d_t}$, and that a random projection of a linear subspace of dimension k (in particular, the range of A) to $\mathbb{R}^{d_t}$ is almost surely injective as long as $d_t \geq k$. Therefore it is sufficient to require $d \geq 3k$ so that it is possible to split $\tilde{x}$ into three views, each of dimension $d_t \geq k$. To guarantee that the k-th largest singular value of each $\tilde{A}_t$ is bounded below in terms of the k-th largest singular value of A (with high probability), we may require d to be somewhat larger: $O(k \log k)$ certainly works (see Appendix B), and we conjecture $c \cdot k$ for some $c > 3$ is in fact sufficient.

Spectral decomposition approaches for ICA. The Gaussian mixture model shares some similarities to a standard model for independent component analysis (ICA) (Comon, 1994; Cardoso and Comon, 1996; Hyvärinen and Oja, 2000; Comon and Jutten, 2010). Here, let $h \in \mathbb{R}^k$ be a random vector with independent entries, and let $z \in \mathbb{R}^k$ be multivariate Gaussian random vector. We think of h as an unobserved signal and z as noise. The observed random vector is

$$x := Ah + z$$

for some $A \in \mathbb{R}^{k \times k}$, where h and z are assumed to be independent. (For simplicity, we only consider square A, although it is easy to generalize to $A \in \mathbb{R}^{d \times k}$ for $d \geq k$.)

In contrast to this ICA model, the spherical Gaussian mixture model is one where h would take values in $\{e_1, e_2, \ldots, e_k\}$, and the covariance of z (given h) is spherical.

For ICA, a spectral decomposition approach related to the one described in Theorem 2 can be used to estimate the columns of A (up to scale), without knowing the noise covariance $\mathbb{E}[zz^\top]$. Such an estimator can be obtained from Theorem 4 using techniques commonplace in the ICA literature; its proof is given in Appendix A for completeness.

Theorem 4. *In the ICA model described above, assume $\mathbb{E}[h_i] = 0$, $\mathbb{E}[h_i^2] = 1$, and $\kappa_i := \mathbb{E}[h_i^4] - 3 \neq 0$ (i.e., the excess kurtosis is non-zero), and that A is non-singular. Define $f \colon \mathbb{R}^k \to \mathbb{R}$ by*

$$f(\eta) := 12^{-1}\big(m_4(\eta) - 3m_2(\eta)^2\big)$$

where $m_p(\eta) := \mathbb{E}[(\eta^\top x)^p]$. Suppose $\phi \in \mathbb{R}^k$ and $\psi \in \mathbb{R}^k$ are such that $\frac{(\phi^\top \mu_1)^2}{(\psi^\top \mu_1)^2}, \frac{(\phi^\top \mu_2)^2}{(\psi^\top \mu_2)^2}, \ldots, \frac{(\phi^\top \mu_k)^2}{(\psi^\top \mu_k)^2} \in \mathbb{R}$ are distinct. Then the matrix

$$M_{\mathrm{ICA}}(\phi, \psi) := \big(\nabla^2 f(\phi)\big)\big(\nabla^2 f(\psi)\big)^{-1}$$

is diagonalizable; the eigenvalues are $\frac{(\phi^\top \mu_1)^2}{(\psi^\top \mu_1)^2}, \frac{(\phi^\top \mu_2)^2}{(\psi^\top \mu_2)^2}, \ldots, \frac{(\phi^\top \mu_k)^2}{(\psi^\top \mu_k)^2}$ and each have geometric multiplicity one, and the corresponding eigenvectors are $\mu_1, \mu_2, \ldots, \mu_k$ (up to scaling and permutation).

Again, choosing ϕ and ψ as random unit vectors ensures the distinctness assumption is satisfied almost surely, and a finite sample analysis can be given using standard matrix perturbation techniques (Anandkumar et al., 2012b). A number of related deterministic algorithms based on algebraic techniques are discussed in the text of Comon and Jutten (2010). Recent work of Arora et al. (2012) provides a finite sample complexity analysis for an efficient estimator based on local search.

Non-degeneracy. The non-degeneracy assumption (Condition 1) is quite natural, and its has the virtue of permitting tractable and consistent estimators. Although previous work has typically tied it with additional assumptions, this work shows that they are largely unnecessary.

One drawback of Condition 1 is that it prevents the straightforward application of these techniques to certain problem domains (*e.g.*, automatic speech recognition (ASR), where the number of mixture components is typically enormous, but the dimension of observations is relatively small; alternatively, the span of the means has dimension $< k$). To compensate, one may require multiple views, which are granted by a number of models, including hidden Markov models used in ASR (Hsu et al., 2012a; Anandkumar et al., 2012b), and combining these views in a tensor product fashion (Allman et al., 2009). This increases the complexity of the estimator, but that may be inevitable as estimation for certain non-singular models is conjectured to be computationally intractable (Mossel and Roch, 2006).

Acknowledgements

References

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models, 2012a. Manuscript.

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

S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *STOC*, 2001.

S. Arora, R. Ge, A. Moitra, and S. Sachdeva. Provable ICA with unknown Gaussian noise, and implications for Gaussian mixtures and autoencoders. In *NIPS*, 2012.

M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, 2010.

A. Bunse-Gerstner, R. Byers, and V. Mehrmann. Numerical methods for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 14(4):927–949, 1993.

J.-F. Cardoso and P. Comon. Independent component analysis, a survey of some algebraic methods. In *IEEE International Symposium on Circuits and Systems Circuits and Systems Connecting the World*, 1996.

J. T. Chang. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences*, 137:51–73, 1996.

K. Chaudhuri and S. Rao. Learning mixtures of product distributions using correlations and independence. In *COLT*, 2008.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications.* Academic Press. Elsevier, 2010.

S. Dasgupta. Learning mixutres of Gaussians. In *FOCS*, 1999.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.

S. Dasgupta and L. Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.

A. M. Frieze, M. Jerrum, and R. Kannan. Learning linear transformations. In *FOCS*, 1996.

D. Hsu, S. M. Kakade, and T. Zhang. An analysis of random design linear regression, 2011. arXiv:1106.2363.

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

D. Hsu, S. M. Kakade, and T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17(14):1–13, 2012b.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, 2010.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

B. G. Lindsay and P. Basak. Multivariate normal mixtures: a fast consistent method. *Journal of the American Statistical Association*, 88(422):468–476, 1993.

A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular values of random matrices and geometry of random polytopes. *Advances in Mathematics*, 195:491–523, 2005.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, 2010.

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

P. Q. Nguyen and O. Regev. Learning a parallelepiped: Cryptanalysis of GGH and NTRU signatures. *Journal of Cryptology*, 22(2):139–160, 2009.

K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society, London, A.*, page 71, 1894.

G. Pisier. *The volume of convex bodies and Banach space geometry*. Cambridge University Press, 1989.

G. W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Wiley, 1985.

S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *FOCS*, 2002.

A   Connection to independent component analysis

Proof of Theorem 4. It can be shown that

$$m_2(\eta) = \mathbb{E}[(\eta^\top Ah)^2] + \mathbb{E}[(\eta^\top z)^2], \qquad m_4(\eta) = \mathbb{E}[(\eta^\top Ah)^4] - 3\mathbb{E}[(\eta^\top Ah)^2]^2 + 3m_2(\eta)^2.$$

By the assumptions,

$$\mathbb{E}[(\eta^\top Ah)^4] = \sum_{i=1}^k (\eta^\top \mu_i)^4 \mathbb{E}[h_i^4] + 3\sum_{i\neq j}(\eta^\top \mu_i)^2(\eta^\top \mu_j)^2$$

$$= \sum_{i=1}^k \kappa_i(\eta^\top \mu_i)^4 + 3\sum_{i,j}(\eta^\top \mu_i)^2(\eta^\top \mu_j)^2$$

$$= \sum_{i=1}^k \kappa_i(\eta^\top \mu_i)^4 + 3\mathbb{E}[(\eta^\top Ah)^2]^2,$$

and therefore

$$f(\eta) = 12^{-1}\big(\mathbb{E}[(\eta^\top Ah)^4] - 3\mathbb{E}[(\eta^\top Ah)^2]^2\big) = 12^{-1}\sum_{i=1}^k \kappa_i(\eta^\top \mu_i)^4.$$

The Hessian of f is given by

$$\nabla^2 f(\eta) = \sum_{i=1}^k \kappa_i(\eta^\top \mu_i)^2 \mu_i\mu_i^\top.$$

Define the diagonal matrices

$$K := \text{diag}(\kappa_1, \kappa_2, \ldots, \kappa_k), \qquad D_2(\eta) := \text{diag}((\eta^\top \mu_1)^2, (\eta^\top \mu_2)^2, \ldots, (\eta^\top \mu_k)^2)$$

and observe that

$$\nabla^2 f(\eta) = AKD_2(\eta)A^\top.$$

By assumption, the diagonal entries of $D_2(\phi)D_2(\psi)^{-1}$ are distinct, and therefore

$$M_{\text{ICA}}(\phi, \psi) = \big(\nabla^2 f(\phi)\big)\big(\nabla^2 f(\psi)\big)^{-1} = AD_2(\phi)D_2(\psi)^{-1}A^{-1}$$

is diagonalizable, and every eigenvalue has geometric multiplicity one. □

B Incoherence and random rotations

The multi-view technique of Anandkumar et al. (2012b) can be used to estimate mixtures of product distributions, which include, as special cases, mixtures of Gaussians with axis-aligned covariances $\Sigma_i = \mathrm{diag}(\sigma_{1,i}^2, \sigma_{2,i}^2, \ldots, \sigma_{d,i}^2)$. Spherical covariances $\Sigma_i = \sigma_i^2 I$ are, of course, also axis-aligned. The idea is to randomly partition the coordinates $[d]$ into three groups, inducing multiple "views" $x = (x_1, x_2, x_3)$ with each $x_t \in \mathbb{R}^{d_t}$ for some $d_t \geq k$ such that x_1, x_2, and x_3 are conditionally independent given h. When the matrix of conditional means $A_t := [\mathbb{E}[x_t | h = 1] | \mathbb{E}[x_t | h = 2] | \cdots | \mathbb{E}[x_t | h = k]] \in \mathbb{R}^{d_t \times k}$ for each view $t \in \{1, 2, 3\}$ has rank k, then an efficient technique similar to that described in Theorem 2 will recover the parameters (for details, see Anandkumar et al., 2012b,a).

Anandkumar et al. (2012b) show that if A has rank k and also satisfies a mild incoherence condition, then a random partitioning guarantees that each A_t has rank k, and lower-bounds the k-th largest singular value of each A_t by that of A. The condition is similar to the spreading condition of Chaudhuri and Rao (2008).

Define coherence$(A) := \max_{i \in [d]} \{e_i^\top \Pi_A e_i\}$ to be the largest diagonal entry of the ortho-projector Π_A to the range of A. When A has rank k, we have coherence$(A) \in [k/d, 1]$; it is maximized when range$(A) = \mathrm{span}\{e_1, e_2, \ldots, e_k\}$ and minimized when the range is spanned by a subset of the Hadamard basis of cardinality k. Roughly speaking, if the matrix of conditional means has low coherence, then its full-rank property is witnessed by many partitions of $[d]$; this is made formal in the following lemma.

Lemma 1. *Assume A has rank k and that* coherence$(A) \leq (\varepsilon^2/6)/\ln(3k/\delta)$ *for some $\varepsilon, \delta \in (0, 1)$. With probability at least $1 - \delta$, a random partitioning of the dimensions $[d]$ into three groups (for each $i \in [d]$, independently pick $t \in \{1, 2, 3\}$ uniformly at random and put i in group t) has the following property. For each $t \in \{1, 2, 3\}$, the matrix A_t obtained by selecting the rows of A in group t has full column rank, and the k-th largest singular value of A_t is at least $\sqrt{(1 - \varepsilon)/3}$ times that of A.*

For a mixture of spherical Gaussians, one can randomly rotate x before applying the random coordinate partitioning. This is because if $\Theta \in \mathbb{R}^{d \times d}$ is an orthogonal matrix, then the distribution of $\tilde{x} := \Theta x$ is also a mixture of spherical Gaussians. Its matrix of conditional means is given by $\tilde{A} := \Theta A$. The following lemma implies that multiplying a tall matrix A by a random rotation Θ causes the product to have low coherence.

Lemma 2 (Hsu et al., 2011). *Let $A \in \mathbb{R}^{d \times k}$ be a fixed matrix with rank k, and let $\Theta \in \mathbb{R}^{d \times d}$ be chosen uniformly at random among all orthogonal $d \times d$ matrices. For any $\eta \in (0, 1)$, with probability at least $1 - \eta$, the matrix $\tilde{A} := \Theta A$ satisfies*

$$\mathrm{coherence}(\tilde{A}) \leq \frac{k + \sqrt{2k \ln(d/\eta)} + 2\ln(d/\eta)}{d\big(1 - 1/(4d) - 1/(360d^3)\big)^2}.$$

Take η from Lemma 2 and ε, δ from Lemma 1 to be constants. Then the incoherence condition of Lemma 1 is satsified provided that $d \geq c \cdot (k \log k)$ for some positive constant c.

C Learning algorithm and finite sample analysis

In this section, we state and analyze a learning algorithm based on the estimator from Theorem 2, which assumed availability of exact moments of x. The proposed algorithm only uses a finite sample

to estimate moments, and also explicitly deals with the eigenvalue separation condition assumed in Theorem 2 via internal randomization.

C.1 Notation

For a matrix $X \in \mathbb{R}^{m \times m}$, we use $\varsigma_t[X]$ to denote the t-th largest singular value of a matrix X, and $\|X\|_2$ to denote its spectral norm (so $\|X\|_2 = \varsigma_1[X]$).

For a third-order tensor $Y \in \mathbb{R}^{m \times m \times m}$ and $U, V, W \in \mathbb{R}^{m \times n}$, we use the notation $Y[U, V, W] \in \mathbb{R}^{n \times n \times n}$ to denote the third-order tensor given by

$$Y[U, V, W]_{j_1, j_2, j_3} = \sum_{1 \leq i_1, i_2, i_3 \leq m} U_{i_1, j_1} V_{i_2, j_2} W_{i_3, j_3} Y_{i_1, i_2, i_3}, \quad \forall j_1, j_2, j_3 \in [n].$$

Note that this is the analogue of $U^\top X V \in \mathbb{R}^{n \times n}$ for a matrix $X \in \mathbb{R}^{m \times m}$ and $U, V \in \mathbb{R}^{m \times n}$. For $Y \in \mathbb{R}^{m \times m \times m}$, we use $\|Y\|_2$ to denote its operator (or supremum) norm $\|Y\|_2 := \sup\{|Y[u, v, w]| : u, v, w \in \mathbb{R}^m, \|u\|_2 = \|v\|_2 = \|w\|_2 = 1\}$.

C.2 Algorithm

The proposed algorithm, called LEARNGMM, is described in Figure 1. The algorithm essentially implements the decomposition strategy in Theorem 2 using plug-in moments. To simplify the analysis, we split our sample (say, initially of size $2n$) in two: we use the first half for empirical moments ($\hat{\mu}$ and $\widehat{\mathcal{M}}_2$) used in constructing $\hat{\sigma}^2$, $\widehat{M}_2$, $\widehat{W}$, and $\widehat{B}$; and we use the second half for empirical moments ($\widehat{W}^\top \underline{\hat{\mu}}$ and $\underline{\widehat{\mathcal{M}}_3}[\widehat{W}, \widehat{W}, \widehat{W}]$) used in constructing $\widehat{M}_3[\widehat{W}, \widehat{W}, \widehat{W}]$. Observe that this ensures $\widehat{M}_3$ is independent of $\widehat{W}$.

Let $\{(x_i, h_i) : i \in [n]\}$ be n i.i.d. copies of (x, h), and write $\mathcal{S} := \{x_1, x_2, \ldots, x_n\}$. Let $\underline{\mathcal{S}}$ be an independent copy of $\mathcal{S}$. Furthermore, define the following moments and empirical moments:

$$\mu := \mathbb{E}[x], \qquad \mathcal{M}_2 := \mathbb{E}[xx^\top], \qquad \mathcal{M}_3 := \mathbb{E}[x \otimes x \otimes x],$$

$$\hat{\mu} := \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} x, \qquad \widehat{\mathcal{M}}_2 := \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} xx^\top, \qquad \underline{\widehat{\mathcal{M}}_3} := \frac{1}{|\underline{\mathcal{S}}|} \sum_{x \in \underline{\mathcal{S}}} x \otimes x \otimes x, \qquad \underline{\hat{\mu}} := \frac{1}{|\underline{\mathcal{S}}|} \sum_{x \in \underline{\mathcal{S}}} x.$$

So $\mathcal{S}$ represents the first half of the sample, and $\underline{\mathcal{S}}$ represents the second half of the sample.

C.3 Structure of the moments

We first recall the basic structure of the moments μ, $\mathcal{M}_2$, and $\mathcal{M}_3$ as established in Theorem 2; for simplicity, we restrict to the special case where $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$.

Lemma 3 (Structure of moments)**.**

$$\mu = \sum_{i=1}^{k} w_i \mu_i,$$

$$\mathcal{M}_2 = \sum_{i=1}^{k} w_i \mu_i \mu_i^\top + \sigma^2 I,$$

$$\mathcal{M}_3 = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i + \sigma^2 \sum_{j=1}^{d} \Big(\mu \otimes e_j \otimes e_j + e_j \otimes \mu \otimes e_j + e_j \otimes e_j \otimes \mu \Big).$$

LEARNGMM

1. Using the first half of the sample, compute empirical mean $\hat{\mu}$ and empirical second-order moments $\widehat{\mathcal{M}}_2$.

2. Let $\hat{\sigma}^2$ be the k-th largest eigenvalue of the empirical covariance matrix $\widehat{\mathcal{M}}_2 - \hat{\mu}\hat{\mu}^\top$.

3. Let $\widehat{M}_2$ be the best rank-k approximation to $\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I$

$$\widehat{M}_2 := \arg\min_{X \in \mathbb{R}^{d \times d}: \mathrm{rank}(X) \leq k} \|(\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I) - X\|_2$$

   which can be obtained via the singular value decomposition.

4. Let $\widehat{U} \in \mathbb{R}^{d \times k}$ be the matrix of left orthonormal singular vectors of $\widehat{M}_2$.

5. Let $\widehat{W} := \widehat{U}(\widehat{U}^\top \widehat{M}_2 \widehat{U})^{\dagger 1/2}$, where $X^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix X.
   Also define $\widehat{B} := \widehat{U}(\widehat{U}^\top \widehat{M}_2 \widehat{U})^{1/2}$.

6. Using the second half of the sample, compute whitened empirical averages $\widehat{W}^\top \underline{\hat{\mu}}$ and third-order moments $\underline{\widehat{\mathcal{M}}_3}[\widehat{W}, \widehat{W}, \widehat{W}]$.

7. Let $\widehat{M}_3[\widehat{W}, \widehat{W}, \widehat{W}] := \underline{\widehat{\mathcal{M}}_3}[\widehat{W}, \widehat{W}, \widehat{W}] - \hat{\sigma}^2 \sum_{i=1}^d \big((\widehat{W}^\top \underline{\hat{\mu}}) \otimes (\widehat{W}^\top e_i) \otimes (\widehat{W}^\top e_i) + (\widehat{W}^\top e_i) \otimes (\widehat{W}^\top \underline{\hat{\mu}}) \otimes (\widehat{W}^\top e_i) + (\widehat{W}^\top e_i) \otimes (\widehat{W}^\top e_i) \otimes (\widehat{W}^\top \underline{\hat{\mu}})\big)$.

8. Repeat the following steps t times (where $t := \lceil \log_2(1/\delta) \rceil$ for confidence $1 - \delta$):

   (a) Choose $\theta \in \mathbb{R}^k$ uniformly at random from the unit sphere in $\mathbb{R}^k$.
   (b) Let $\{(\hat{v}_i, \hat{\lambda}_i) : i \in [k]\}$ be the eigenvector/eigenvalue pairs of $\widehat{M}_3[\widehat{W}, \widehat{W}, \widehat{W}\theta]$.

   Retain the results for which $\min\big(\{|\hat{\lambda}_i - \hat{\lambda}_j| : i \neq j\} \cup \{|\hat{\lambda}_i| : i \in [k]\}\big)$ is largest.

9. Return the parameter estimates $\hat{\sigma}^2$,

$$\hat{\mu}_i := \frac{\hat{\lambda}_i}{\theta^\top \hat{v}_i} \widehat{B} \hat{v}_i, \quad i \in [k],$$
$$\hat{w} := [\hat{\mu}_1 | \hat{\mu}_2 | \cdots | \hat{\mu}_k]^\dagger \hat{\mu}.$$

Figure 1: Algorithm for learning mixtures of Gaussians with common spherical covariance.

C.4 Concentration behavior of empirical quantities

In this subsection, we prove concentration properties of empirical quantities based on $\mathcal{S}$; clearly the same properties hold for $\underline{\mathcal{S}}$.

Let $\mathcal{S}_i := \{x_j \in \mathcal{S} : h_j = i\}$ and $\hat{w}_i := |\mathcal{S}_i|/|\mathcal{S}|$ for $i \in [k]$. Also, define the following (empirical) conditional moments:

$$\mu_i := \mathbb{E}[x|h=i], \qquad \mathcal{M}_{2,i} := \mathbb{E}[xx^\top|h=i], \qquad \mathcal{M}_{3,i} := \mathbb{E}[x \otimes x \otimes x|h=i],$$

$$\hat{\mu}_i := \frac{1}{|\mathcal{S}_i|} \sum_{x \in \mathcal{S}_i} x, \qquad \widehat{\mathcal{M}}_{2,i} := \frac{1}{|\mathcal{S}_i|} \sum_{x \in \mathcal{S}_i} xx^\top, \qquad \widehat{\mathcal{M}}_{3,i} := \frac{1}{|\mathcal{S}_i|} \sum_{x \in \mathcal{S}_i} x \otimes x \otimes x.$$

Lemma 4 (Concentration of proportions). *Pick any $\delta \in (0, 1/2)$. With probability at least $1 - 2\delta$,*

$$|\hat{w}_i - w_i| \le \sqrt{\frac{2w_i(1-w_i)\ln(2k/\delta)}{n}} + \frac{2\ln(2k/\delta)}{3n}, \quad \forall i \in [k];$$

$$\left(\sum_{i=1}^{k} (\hat{w}_i - w_i)^2 \right)^{1/2} \le \frac{1 + \sqrt{\ln(1/\delta)}}{\sqrt{n}}.$$

Proof. The first inequality follows from Bernstein's inequality and a union bound. The second inequality follows from a simple application of McDiarmid's inequality (see Hsu et al., 2012a, Proposition 19). □

Lemma 5 (Concentration of per-component empirical moments). *Pick any $\delta \in (0, 1)$ and any matrix $R \in \mathbb{R}^{d \times r}$ of rank r.*

1. *First-order moments: with probability at least $1 - \delta$,*

$$\|R^\top(\hat{\mu}_i - \mu_i)\|_2 \le \sigma\|R\|_2 \sqrt{\frac{r + 2\sqrt{r\ln(k/\delta)} + 2\ln(k/\delta)}{\hat{w}_i n}}, \quad \forall i \in [k].$$

2. *Second-order moments: with probability at least $1 - \delta$,*

$$\|R^\top(\widehat{\mathcal{M}}_{2,i} - \mathcal{M}_{2,i})R\|_2 \le \sigma^2\|R\|_2^2 \left(\sqrt{\frac{128(r\ln 9 + \ln(2k/\delta))}{\hat{w}_i n}} + \frac{4(r\ln 9 + \ln(2k/\delta))}{\hat{w}_i n} \right)$$

$$+ 2\sigma\|R^\top\mu_i\|_2\|R\|_2 \sqrt{\frac{r + 2\sqrt{r\ln(2k/\delta)} + 2\ln(2k/\delta)}{\hat{w}_i n}}, \quad \forall i \in [k].$$

3. *Third-order moments: with probability at least $1 - \delta$,*

$$\|(\widehat{\mathcal{M}}_{3,i} - \mathcal{M}_{3,i})[R, R, R]\|_2 \le \sigma^3\|R\|_2^3 \sqrt{\frac{108e^3\lceil r\ln 13 + \ln(3k/\delta)\rceil^3}{\hat{w}_i n}}$$

$$+ 3\sigma^2\|R^\top\mu_i\|_2\|R\|_2^2 \left(\sqrt{\frac{128(r\ln 9 + \ln(3k/\delta))}{\hat{w}_i n}} + \frac{4(r\ln 9 + \ln(3k/\delta))}{\hat{w}_i n} \right)$$

$$+ 3\sigma\|R^\top\mu_i\|_2^2\|R\|_2 \sqrt{\frac{r + 2\sqrt{r\ln(3k/\delta)} + 2\ln(3k/\delta)}{\hat{w}_i n}}, \quad \forall i \in [k].$$

Proof. We separately consider first-, second-, and third-order moments. Throughout, we let the thin SVD of R be given by $R = USV^\top$, where $U \in \mathbb{R}^{d \times r}$ has orthonormal columns, and $\|VS\|_2 = \|R\|_2$.

First-order moments. Observe that $(\hat{w}_i n / \sigma^2) \| U^\top (\hat{\mu}_i - \mu_i) \|_2^2$ is distributed as the sum of r independent χ^2 random variables, each with one degree of freedom. Thus, Lemma 18 and union bounds imply

$$\Pr\left[\exists i \in [k] \centerdot \|U^\top(\hat{\mu}_i - \mu_i)\|_2^2 > \sigma^2 \left(\frac{r + 2\sqrt{r \ln(k/\delta)} + 2\ln(k/\delta)}{\hat{w}_i n} \right)\right] \leq \delta.$$

Second-order moments. Since $\mathcal{M}_{2,i} = \sigma^2 I + \mu_i \mu_i^\top$, it follows by the triangle and Cauchy-Schwarz inequalities that

$$\|R^\top(\widehat{\mathcal{M}}_{2,i} - \mathcal{M}_{2,i})R\|_2 \leq \|R\|_2^2 \left\| \frac{1}{\hat{w}_i n} \sum_{j \in [n] : x_j \in \mathcal{S}_i} \left(U^\top(x_j - \mu_i)(x_j - \mu_i)^\top U - \sigma^2 I \right) \right\|_2$$
$$+ 2\|R^\top \mu_i\|_2 \|R^\top(\hat{\mu}_i - \mu_i)\|_2.$$

A tail bound for the first term follows from Lemma 19, combined with a union bound:

$$\Pr\left[\exists i \in [k] \centerdot \left\| \frac{1}{\hat{w}_i n} \sum_{j \in [n]: x_j \in \mathcal{S}_i} \left(U^\top(x_j - \mu_i)(x_j - \mu_i)^\top U - \sigma^2 I \right) \right\|_2 \right.$$
$$\left. > \sigma^2 \left(\sqrt{\frac{128(r \ln 9 + \ln(k/\delta))}{\hat{w}_i n}} + \frac{4(r \ln 9 + \ln(k/\delta))}{\hat{w}_i n} \right) \right] \leq \delta.$$

The second term is handled as above.

Third-order moments. It can be checked that

$$\mathcal{M}_{3,i} = \mu_i \otimes \mu_i \otimes \mu_i + \sigma^2 \sum_{\iota=1}^{d} \left(\mu_i \otimes e_\iota \otimes e_\iota + e_\iota \otimes \mu_i \otimes e_\iota + e_\iota \otimes e_\iota \otimes \mu_i \right)$$

(similar to Lemma 3) and

$$\widehat{\mathcal{M}}_{3,i} - \mathcal{M}_{3,i}$$
$$= \frac{1}{\hat{w}_i n} \left(\sum_{j \in [n]:x_j \in \mathcal{S}_i} (x_j - \mu_i) \otimes (x_j - \mu_i) \otimes (x_j - \mu_i) \right.$$
$$+ \sum_{j \in [n]:x_j \in \mathcal{S}_i} \left(\mu_i \otimes (x_j - \mu_i) \otimes (x_j - \mu_i) - \sigma^2 \sum_{\iota=1}^{d} \mu_i \otimes e_\iota \otimes e_\iota \right)$$
$$+ \sum_{j \in [n]:x_j \in \mathcal{S}_i} \left((x_j - \mu_i) \otimes \mu_i \otimes (x_j - \mu_i) - \sigma^2 \sum_{\iota=1}^{d} e_\iota \otimes \mu_i \otimes e_\iota \right)$$
$$+ \sum_{j \in [n]:x_j \in \mathcal{S}_i} \left((x_j - \mu_i) \otimes (x_j - \mu_i) \otimes \mu_i - \sigma^2 \sum_{\iota=1}^{d} e_\iota \otimes e_\iota \otimes \mu_i \right)$$
$$+ \left. \sum_{j \in [n]:x_j \in \mathcal{S}_i} \mu_i \otimes \mu_i \otimes (x_j - \mu_i) + \sum_{j \in [n]:x_j \in \mathcal{S}_i} \mu_i \otimes (x_j - \mu_i) \otimes \mu_i + \sum_{j \in [n]:x_j \in \mathcal{S}_i} (x_j - \mu_i) \otimes \mu_i \otimes \mu_i \right).$$

Therefore, by the triangle and Cauchy-Schwarz inequalities,

$$\|(\widehat{\mathcal{M}}_{3,i} - \mathcal{M}_{3,i})[R, R, R]\|_2 \leq \|R\|_2^3 \left\|\frac{1}{\hat{w}_i n} \sum_{j \in [n]: x_j \in \mathcal{S}_i} U^\top(x_j - \mu_i) \otimes U^\top(x_j - \mu_i) \otimes U^\top(x_j - \mu_i)\right\|_2$$

$$+ 3\|R^\top \mu_i\|_2 \left\|\frac{1}{\hat{w}_i n} \sum_{j \in [n]: x_j \in \mathcal{S}_i} R^\top\left((x_j - \mu_i)(x_j - \mu_i)^\top - \sigma^2 I\right)R\right\|_2$$

$$+ 3\|R^\top \mu_i\|_2^2 \|R^\top(\hat{\mu}_i - \mu_i)\|_2.$$

A tail bound for the first term is given by Lemma 21, combined with a union bound:

$$\Pr\left[\exists i \in [k] \centerdot \left\|\frac{1}{\hat{w}_i n} \sum_{j \in [n]: x_j \in \mathcal{S}_i} \frac{1}{\sigma^3} U^\top(x_j - \mu_i) \otimes U^\top(x_j - \mu_i) \otimes U^\top(x_j - \mu_i)\right\|_2 \right.$$

$$\left. > \sigma^3 \sqrt{\frac{108e^3\lceil r \ln 13 + \ln(k/\delta)\rceil^3}{\hat{w}_i n}}\right] \leq \delta.$$

The other terms are handled as per above. □

Lemma 6 (Accuracy of empirical moments). *Fix a matrix* $R \in \mathbb{R}^{d \times r}$. *Define* $\mathcal{B}_{1,R} := \max_{i \in [k]} \|R^\top \mu_i\|_2$, $\mathcal{B}_{2,R} := \max_{i \in [k]} \|R^\top \mathcal{M}_{2,i} R\|_2$, $\mathcal{B}_{3,R} := \max_{i \in [k]} \|\mathcal{M}_{3,i}[R, R, R]\|_2$, $\mathcal{E}_{1,R} := \max_{i \in [k]} \|R^\top(\hat{\mu}_i - \mu_i)\|_2$, $\mathcal{E}_{2,R} := \max_{i \in [k]} \|R^\top(\widehat{\mathcal{M}}_{2,i} - \mathcal{M}_{2,i})R\|_2$, $\mathcal{E}_{3,R} := \max_{i \in [k]} \|\widehat{\mathcal{M}}_{3,i}[R, R, R] - \mathcal{M}_{3,i}[R, R, R]\|_2$, *and* $\mathcal{E}_w := (\sum_{i=1}^k (\hat{w}_i - w_i)^2)^{1/2}$. *Then*

$$\|R^\top(\hat{\mu} - \mu)\|_2 \leq (1 + \sqrt{k}\mathcal{E}_w)\mathcal{E}_{1,R} + \sqrt{k}\mathcal{B}_{1,R}\mathcal{E}_w;$$
$$\|R^\top(\widehat{\mathcal{M}}_2 - \mathcal{M}_2)R\|_2 \leq (1 + \sqrt{k}\mathcal{E}_w)\mathcal{E}_{2,R} + \sqrt{k}\mathcal{B}_{2,R}\mathcal{E}_w;$$
$$\|(\widehat{\mathcal{M}}_3 - \mathcal{M}_3)[R, R, R]\|_2 \leq (1 + \sqrt{k}\mathcal{E}_w)\mathcal{E}_{3,R} + \sqrt{k}\mathcal{B}_{3,R}\mathcal{E}_w.$$

Proof. We just show the third claimed inequalitiy, as the others are similar. Write as shorthand $\widehat{T}_i := \widehat{\mathcal{M}}_{3,i}[R, R, R]$ and $T_i := \mathcal{M}_{3,i}[R, R, R]$. Then

$$\left\|\sum_{i=1}^k \hat{w}_i \widehat{T}_i - \sum_{i=1}^k w_i T_i\right\|_2 \leq \left\|\sum_{i=1}^k w_i(\widehat{T}_i - T_i)\right\|_2 + \left\|\sum_{i=1}^k (\hat{w}_i - w_i)T_i\right\|_2 + \left\|\sum_{i=1}^k (\hat{w}_i - w_i)(\widehat{T}_i - T_i)\right\|_2$$

$$\leq \sum_{i=1}^k w_i\|\widehat{T}_i - T_i\|_2 + \sum_{i=1}^k |\hat{w}_i - w_i|\|T_i\|_2 + \sum_{i=1}^k |\hat{w}_i - w_i|\|\widehat{T}_i - T_i\|_2$$

$$\leq \max_{i \in [k]} \|\widehat{T}_i - T_i\|_2 + \sqrt{k}\|\hat{w} - w\|_2 \max_{i \in [k]} \|T_i\|_2 + \sqrt{k}\|\hat{w} - w\|_2 \max_{i \in [k]} \|\widehat{T}_i - T_i\|_2$$

$$= \mathcal{E}_{3,R} + \sqrt{k}\mathcal{E}_w\mathcal{B}_{3,R} + \sqrt{k}\mathcal{E}_w\mathcal{E}_{3,R}$$

where the first and second steps use the triangle inequality, and the second step uses Hölder's inequality. □

16

C.5   Estimation of σ^2, M_2, and M_3

The covariance matrix can be written as $\mathcal{M}_2 - \mu\mu^\top$, and the empirical covariance matrix can be written as $\widehat{\mathcal{M}}_2 - \hat{\mu}\hat{\mu}^\top$. Recall that the estimate of of σ^2, denoted by $\hat{\sigma}^2$, is given by the k-th largest eigenvalue of the empirical covariance matrix $\widehat{\mathcal{M}}_2 - \hat{\mu}\hat{\mu}^\top$; and that the estimate of M_2, denoted by $\widehat{M_2}$, is the best rank-k approximation to $\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I$. Of course, the singular values of a positive semi-definite matrix are the same as its eigenvalues; in particular, $\hat{\sigma}^2 = \varsigma_k[\widehat{\mathcal{M}}_2 - \hat{\mu}\hat{\mu}^\top]$.

Lemma 7 (Accuracy of $\hat{\sigma}^2$ and $\widehat{M_2}$)**.**

1. $|\hat{\sigma}^2 - \sigma^2| \leq \|\widehat{\mathcal{M}}_2 - \mathcal{M}_2\|_2 + 2\|\mu\|_2\|\hat{\mu} - \mu\|_2 + \|\hat{\mu} - \mu\|_2^2$.

2. $\|\widehat{M_2} - M_2\|_2 \leq 4\|\widehat{\mathcal{M}}_2 - \mathcal{M}_2\|_2 + 4\|\mu\|_2\|\hat{\mu} - \mu\|_2 + 2\|\hat{\mu} - \mu\|_2^2$.

Proof. Using Weyl's inequality (Stewart and Sun, 1990, Theorem 4.11, p. 204), we obtain $|\varsigma_k[\widehat{\mathcal{M}}_2 - \hat{\mu}\hat{\mu}^\top] - \varsigma_k[\mathcal{M}_2 - \mu\mu^\top]| \leq \|(\widehat{\mathcal{M}}_2 - \hat{\mu}\hat{\mu}^\top) - (\mathcal{M}_2 - \mu\mu^\top)\|_2 \leq \|\widehat{\mathcal{M}}_2 - \mathcal{M}_2\|_2 + 2\|\mu\|_2\|\hat{\mu} - \mu\|_2 + \|\hat{\mu} - \mu\|_2^2$. The first claim then follows by observing that $\varsigma_k[\mathcal{M}_2 - \mu\mu^\top] = \sigma^2$ as per Theorem 1.

For the second claim, observe that $\varsigma_{k+1}(\mathcal{M}_2 - \sigma^2 I) = 0$ as $\mathcal{M}_2 - \sigma^2 I$ has rank k. Therefore $\varsigma_{k+1}(\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I) = |\varsigma_{k+1}(\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I) - \varsigma_{k+1}(\mathcal{M}_2 - \sigma^2 I)| \leq \|(\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I) - (\mathcal{M}_2 - \sigma^2 I)\|_2$, again using Weyl's inequality. Since $\widehat{M_2}$ is the best rank-k approximation to $\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I$, it follows that $\|\widehat{M_2} - (\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I)\|_2 \leq \varsigma_{k+1}(\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I) \leq \|(\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I) - (\mathcal{M}_2 - \sigma^2 I)\|_2$. Therefore $\|\widehat{M_2} - M_2\|_2 \leq \|(\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I) - (\mathcal{M}_2 - \sigma^2 I)\|_2 + \|\widehat{M_2} - (\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I)\|_2 \leq 2\|(\widehat{\mathcal{M}}_2 - \hat{\sigma}^2 I) - (\mathcal{M}_2 - \sigma^2 I)\|_2 \leq 2\|\widehat{\mathcal{M}}_2 - \mathcal{M}_2\|_2 + 2|\hat{\sigma}^2 - \sigma^2| \leq 4\|\widehat{\mathcal{M}}_2 - \mathcal{M}_2\|_2 + 4\|\mu\|_2\|\hat{\mu} - \mu\|_2 + 2\|\hat{\mu} - \mu\|_2^2$. □

Recall that the estimate of M_3, denoted by $\widehat{M_3}$, is given by $\underline{\widehat{\mathcal{M}}_3} - \hat{\sigma}^2 \sum_{i=1}^d (\hat{\mu} \otimes e_i \otimes e_i + e_i \otimes \hat{\mu} \otimes e_i + e_i \otimes e_i \otimes \hat{\mu})$.

Lemma 8 (Accuracy of $\widehat{M_3}$)**.** *For any matrix $R \in \mathbb{R}^{d \times r}$,*

$$
\begin{aligned}
\|\widehat{M_3}[R, R, R] - M_3[R, R, R]\|_2 \leq &\ \|\underline{\widehat{\mathcal{M}}_3}[R, R, R] - \mathcal{M}_3[R, R, R]\|_2 \\
&+ 3\|R\|_2^2 (\|R^\top(\hat{\mu} - \mu)\|_2 + \|R^\top \mu\|_2) \\
&\quad (\|\widehat{\mathcal{M}}_2 - \mathcal{M}_2\|_2 + 2\|\mu\|_2\|\hat{\mu} - \mu\|_2 + \|\hat{\mu} - \mu\|_2^2) \\
&+ \sigma^2 \|R\|_2^2 \|R^\top(\hat{\mu} - \mu)\|_2 .
\end{aligned}
$$

Proof. Let $\widehat{G} := \sum_{i=1}^d (\hat{\mu} \otimes e_i \otimes e_i + e_i \otimes \hat{\mu} \otimes e_i + e_i \otimes e_i \otimes \hat{\mu})$ and $G := \sum_{i=1}^d (\mu \otimes e_i \otimes e_i + e_i \otimes \mu \otimes e_i + e_i \otimes e_i \otimes \mu)$. Then

$\|(\widehat{M_3} - M_3)[R, R, R]\|_2$

$= \|(\underline{\widehat{\mathcal{M}}_3} - \mathcal{M}_3)[R, R, R] - (\hat{\sigma}^2 - \sigma^2)(\widehat{G} - G)[R, R, R] - (\hat{\sigma}^2 - \sigma^2)G[R, R, R] - \sigma^2(\widehat{G} - G)[R, R, R]\|_2$

$\leq \|(\underline{\widehat{\mathcal{M}}_3} - \mathcal{M}_3)[R, R, R]\|_2 + |\hat{\sigma}^2 - \sigma^2|\|(\widehat{G} - G)[R, R, R]\|_2 + |\hat{\sigma}^2 - \sigma^2|\|G[R, R, R]\|_2$

$\quad + \sigma^2 \|(\widehat{G} - G)[R, R, R]\|_2$.

Observe that by the triangle inequality, $\|G[R, R, R]\|_2 \leq 3\|R\|_2^2\|R^\top \mu\|_2$ and $\|(\widehat{G} - G)[R, R, R]\|_2 \leq 3\|R\|_2^2\|R^\top(\hat{\mu} - \mu)\|_2$. Furthermore, by Lemma 7, we have $|\hat{\sigma}^2 - \sigma^2| \leq \|\widehat{\mathcal{M}}_2 - \mathcal{M}_2\|_2 + 2\|\mu\|_2\|\hat{\mu} - \mu\|_2 + \|\hat{\mu} - \mu\|_2^2$. Therefore the claim follows. □

C.6 Properties of projection and whitening operators

Recall that $\widehat{U} \in \mathbb{R}^{d \times k}$ is the matrix of left orthonormal singular vectors of $\widehat{M}_2$, and let $\widehat{S} \in \mathbb{R}^{k \times k}$ be the diagonal matrix of corresponding singular values. Analogously define U and S relative to M_2.

Define $\mathcal{E}_{M_2} := \|\widehat{M}_2 - M_2\|_2 / \varsigma_k[M_2]$.

Lemma 9 (Properties of projection operators). *Assume $\mathcal{E}_{M_2} \leq 1/3$. Then*

1. $(1 + \mathcal{E}_{M_2})S \succeq \widehat{U}^\top \widehat{M}_2 \widehat{U} = \widehat{S} \succeq (1 - \mathcal{E}_{M_2})S \succ 0$.

2. $\varsigma_k[\widehat{U}^\top U] \geq \sqrt{1 - (9/4)\mathcal{E}_{M_2}^2} > 0$.

3. $\varsigma_k[\widehat{U}^\top M_2 \widehat{U}] \geq (1 - (9/4)\mathcal{E}_{M_2}^2)\varsigma_k[M_2] > 0$.

4. $\|(I - \widehat{U}\widehat{U}^\top)UU^\top\|_2 \leq (3/2)\mathcal{E}_{M_2}$.

Proof. By the assumptions that $\mathcal{E}_{M_2} \leq 1/3$ and M_2 is symmetric positive definite, we have

$$|\varsigma_t[\widehat{M}_2] - \varsigma_t[M_2]| \leq \|\widehat{M}_2 - M_2\|_2 \leq \mathcal{E}_{M_2}\varsigma_k[M_2], \quad \forall t \in [k]$$

by Weyl's inequality. Therefore $\widehat{M}_2$ is symmetric positive definite, and

$$(1 + \mathcal{E}_{M_2})S \succeq \widehat{U}^\top \widehat{M}_2 \widehat{U} = \widehat{S} \succeq (1 - \mathcal{E}_{M_2})S,$$

which proves the first claim.

Now let $\widehat{U}_\perp \in \mathbb{R}^{d \times (d-k)}$ be a matrix with orthonormal columns spanning the orthogonal complement of the range of $\widehat{M}_2$. By Wedin's theorem (Stewart and Sun, 1990, Theorem 4.4, p. 262) and the assumption that $\mathcal{E}_{M_2} \leq 1/3$,

$$\|\widehat{U}_\perp^\top U\|_2 \leq \frac{\|\widehat{M}_2 - M_2\|_2}{\varsigma_k[\widehat{M}_2]} \leq \frac{\mathcal{E}_{M_2}}{1 - \mathcal{E}_{M_2}} \leq \frac{3}{2}\mathcal{E}_{M_2} \leq \frac{1}{2}.$$

Therefore $\widehat{U}^\top U$ is non-singular, and for any $v \in \mathbb{R}^k$, $\|\widehat{U}^\top Uv\|_2^2 = 1 - \|\widehat{U}_\perp^\top Uv\|_2^2 \geq 1 - (9/4)\mathcal{E}_{M_2}^2$, which in turn implies the second claim. The second claim then implies that $\varsigma_k[\widehat{U}^\top M_2 \widehat{U}] \geq \varsigma_k[\widehat{U}^\top U]^2 \varsigma_k[M_2] \geq (1 - (9/4)\mathcal{E}_{M_2}^2)\varsigma_k[M_2]$, which gives the third claim. For the final claim, observe that $\|(I - \widehat{U}\widehat{U}^\top)UU^\top\|_2 = \|\widehat{U}_\perp \widehat{U}_\perp^\top UU^\top\|_2 = \|\widehat{U}_\perp^\top U\|_2 \leq (3/2)\mathcal{E}_{M_2}$, using the fact that $\widehat{U}_\perp$ and U have orthonormal columns, and the above displayed inequality. $\square$

Recall that $\widehat{W} = \widehat{U}(\widehat{U}^\top \widehat{M}_2 \widehat{U})^{\dagger 1/2}$.

Lemma 10 (Properties of whitening operators). *Define $W := \widehat{W}(\widehat{W}^\top M_2 \widehat{W})^{\dagger 1/2}$. Assume $\mathcal{E}_{M_2} \leq 1/3$. Then*

1. $\widehat{W}^\top M_2 \widehat{W}$ *is symmetric positive definite, $W^\top M_2 W = I$, and $W^\top A \operatorname{diag}(w)^{1/2}$ is orthogonal.*

2. $\|\widehat{W}\|_2 \leq \frac{1}{\sqrt{(1 - \mathcal{E}_{M_2})\varsigma_k[M_2]}}$.

3. $\|(\widehat{W}^\top M_2 \widehat{W})^{1/2} - I\|_2 \leq (3/2)\mathcal{E}_{M_2}$,
   $\|(\widehat{W}^\top M_2 \widehat{W})^{-1/2} - I\|_2 \leq (3/2)\mathcal{E}_{M_2}$,
   $\|\widehat{W}^\top A \operatorname{diag}(w)^{1/2}\|_2 \leq \sqrt{1 + (3/2)\mathcal{E}_{M_2}}$,
   $\|(\widehat{W} - W)^\top A \operatorname{diag}(w)^{1/2}\|_2 \leq (3/2)\sqrt{1 + (3/2)\mathcal{E}_{M_2}}\,\mathcal{E}_{M_2}$.

Proof. By Lemma 9 (first and third claims), the matrices $\widehat{U}^\top \widehat{M_2}\widehat{U}$ and $\widehat{U}^\top M_2 \widehat{U}$ are symmetric positive definite. Therefore

$$\widehat{W}^\top M_2 \widehat{W} = (\widehat{U}^\top \widehat{M_2}\widehat{U})^{-1/2}(\widehat{U}^\top M_2 \widehat{U})(\widehat{U}^\top \widehat{M_2}\widehat{U})^{-1/2} \succ 0,$$
$$W = \widehat{W}(\widehat{W}^\top M_2 \widehat{W})^{-1/2},$$
$$W^\top M_2 W = (\widehat{W}^\top M_2 \widehat{W})^{-1/2}(\widehat{W}^\top M_2 \widehat{W})(\widehat{W}^\top M_2 \widehat{W})^{-1/2} = I.$$

Since $M_2 = A \operatorname{diag}(w)A^\top$, it follows from the third equation above that $W^\top A \operatorname{diag}(w)^{1/2}$ is orthogonal. Thus the first claim is established.

For the second claim, note that

$$\|\widehat{W}\|_2 \leq \|(\widehat{U}^\top \widehat{M_2}\widehat{U})^{-1/2}\|_2 = \varsigma_k[\widehat{U}^\top \widehat{M_2}\widehat{U}]^{-1/2} \leq ((1 - \mathcal{E}_{M_2})\varsigma_k[M_2])^{-1/2}$$

where the last inequality follows from Lemma 9 (first claim).

To show the third claim, we first bound $\|\widehat{W}^\top M_2 \widehat{W} - I\|_2$ as

$$\|\widehat{W}^\top M_2 \widehat{W} - I\|_2 = \|\widehat{W}^\top(M_2 - \widehat{M_2})\widehat{W}\|_2$$
$$\leq \|\widehat{W}\|_2^2 \|M_2 - \widehat{M_2}\|_2$$
$$\leq \frac{\mathcal{E}_{M_2}}{1 - \mathcal{E}_{M_2}} \leq \frac{3}{2}\mathcal{E}_{M_2} \leq \frac{1}{2}$$

where the second inequality follows from the first claim. This implies that every eigenvalue of $\widehat{W}^\top M_2 \widehat{W}$ is contained in the interval of radius $(3/2)\mathcal{E}_{M_2}$ around 1. Because $|(1 + x)^{-1/2} - 1| \leq |x|$ for all $|x| \leq 1/2$, the same is true of the eigenvalues of $(\widehat{W}^\top M_2 \widehat{W})^{-1/2}$:

$$\|(\widehat{W}^\top M_2 \widehat{W})^{-1/2} - I\|_2 \leq \frac{3}{2}\mathcal{E}_{M_2}.$$

Furthermore,

$$\|\widehat{W}^\top A \operatorname{diag}(w)^{1/2}\|_2^2 = \|\widehat{W}^\top M_2 \widehat{W}\|_2$$
$$= \|I + \widehat{W}^\top M_2 \widehat{W} - I\|_2$$
$$\leq 1 + \|\widehat{W}^\top M_2 \widehat{W} - I\|_2 \leq 1 + \frac{3}{2}\mathcal{E}_{M_2},$$

so

$$\|(\widehat{W} - W)^\top A \operatorname{diag}(w)^{1/2}\|_2 = \|(I - (\widehat{W}^\top M_2 \widehat{W})^{-1/2})\widehat{W}^\top A \operatorname{diag}(w)^{1/2}\|_2$$
$$\leq \|I - (\widehat{W}^\top M_2 \widehat{W})^{-1/2}\|_2 \|\widehat{W}^\top A \operatorname{diag}(w)^{1/2}\|_2 \leq \frac{3}{2}\mathcal{E}_{M_2}\sqrt{1 + \frac{3}{2}\mathcal{E}_{M_2}}.$$

This establishes the third claim. $\qquad\square$

Define $\widehat{T} := \widehat{M}_3[\widehat{W}, \widehat{W}, \widehat{W}]$ and $T := M_3[W, W, W]$, both symmetric tensors in $\mathbb{R}^{k \times k \times k}$. Also, define $\widehat{T}[u] := \widehat{M}_3[\widehat{W}, \widehat{W}, \widehat{W}u]$ and $T[u] := M_3[W, W, Wu]$, both symmetric matrices in $\mathbb{R}^{k \times k}$.

Lemma 11 (Tensor structure). *The tensor T can be written as*

$$T = \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} (W^\top A \operatorname{diag}(w)^{1/2} e_i) \otimes (W^\top A \operatorname{diag}(w)^{1/2} e_i) \otimes (W^\top A \operatorname{diag}(w)^{1/2} e_i)$$

where the vectors $\{W^\top A \operatorname{diag}(w)^{1/2} e_i : i \in [k]\}$ are orthonormal. Furthermore, the eigenvectors of $T[u]$ are $\{W^\top A \operatorname{diag}(w)^{1/2} e_i : i \in [k]\}$ and the corresponding eigenvalues are $\{u^\top W^\top \mu_i : i \in [k]\}$.

Proof. The structure of T follows from Lemma 3, and the orthogonality of $\{W^\top A \operatorname{diag}(w)^{1/2} e_i : i \in [k]\}$ follows from Lemma 10 (first claim). The eigendecomposition of $T[u]$ is then readily seen from the structure of T. $\qquad\square$

Lemma 12 (Tensor accuracy). *Assume $\mathcal{E}_{M_2} \leq 1/3$. Then*

$$\|\widehat{T} - T\|_2 \leq \|\widehat{M}_3[\widehat{W}, \widehat{W}, \widehat{W}] - M_3[\widehat{W}, \widehat{W}, \widehat{W}]\|_2 + \frac{6}{\sqrt{w_{\min}}} \mathcal{E}_{M_2}.$$

Proof. By Lemma 11, $T = \sum_{i=1}^{k} w_i^{-1/2} v_i \otimes v_i \otimes v_i$ for some orthonormal vectors $\{v_i : i \in [k]\}$, so $\|T\|_2 \leq 1/\sqrt{w_{\min}}$. By Lemma 10 (first and third claims), $\widehat{W} = W(\widehat{W}^\top M_2 \widehat{W})^{1/2}$ and $\|(\widehat{W}^\top M_2 \widehat{W})^{1/2} - I\|_2 \leq (3/2)\mathcal{E}_{M_2}$. Therefore

$$\|M_3[\widehat{W}, \widehat{W}, \widehat{W}] - M_3[W, W, W]\|_2$$

$$\leq \|M_3[\widehat{W} - W, \widehat{W}, \widehat{W}]\|_2 + \|M_3[W, \widehat{W} - W, \widehat{W}]\|_2 + \|M_3[W, W, \widehat{W} - W]\|_2$$

$$\leq \|M_3[W, W, W]\|_2 \Big(\|(\widehat{W}^\top M_2 \widehat{W})^{1/2} - I\|_2 \|(\widehat{W}^\top M_2 \widehat{W})^{1/2}\|_2^2$$

$$+ \|(\widehat{W}^\top M_2 \widehat{W})^{1/2} - I\|_2 \|(\widehat{W}^\top M_2 \widehat{W})^{1/2}\|_2 + \|(\widehat{W}^\top M_2 \widehat{W})^{1/2} - I\|_2 \Big)$$

$$\leq \|M_3[W, W, W]\|_2 \Big((1 + (3/2)\mathcal{E}_{M_2})^2 (3/2)\mathcal{E}_{M_2} + (1 + (3/2)\mathcal{E}_{M_2})(3/2)\mathcal{E}_{M_2} + (3/2)\mathcal{E}_{M_2} \Big)$$

$$\leq 6\|M_3[W, W, W]\|_2 \mathcal{E}_{M_2} \leq \frac{6}{\sqrt{w_{\min}}} \mathcal{E}_{M_2}.$$

Thus we can bound $\|\widehat{T} - T\|_2$ using the triangle inequality and the above bound:

$$\|\widehat{T} - T\|_2 = \|\widehat{M}_3[\widehat{W}, \widehat{W}, \widehat{W}] - M_3[W, W, W]\|_2$$

$$\leq \|\widehat{M}_3[\widehat{W}, \widehat{W}, \widehat{W}] - M_3[\widehat{W}, \widehat{W}, \widehat{W}]\|_2 + \|M_3[\widehat{W}, \widehat{W}, \widehat{W}] - M_3[W, W, W]\|_2$$

$$\leq \|\widehat{M}_3[\widehat{W}, \widehat{W}, \widehat{W}] - M_3[\widehat{W}, \widehat{W}, \widehat{W}]\|_2 + \frac{6}{\sqrt{w_{\min}}} \mathcal{E}_{M_2}. \qquad\square$$

C.7 Eigendecomposition analysis

Define

$$\gamma := \frac{1}{2\sqrt{w_{\max}}\sqrt{ek}\binom{k+1}{2}} \tag{1}$$

where $w_{\max} := \max_{i \in [k]} w_i$.

Lemma 13 (Random separation). *Let $\theta \in \mathbb{R}^k$ be a random vector distributed uniformly over the unit sphere in $\mathbb{R}^k$. Let $Q := \{e_i - e_j : \{i,j\} \in \binom{k}{2}\} \cup \{e_i : i \in [k]\}$. Then*

$$\Pr\left[\min_{q \in Q} |\theta^\top W^\top Aq| > \gamma\right] \geq \frac{1}{2}$$

where the probability is taken with respect to the distribution of θ.

Proof. By Lemma 17, with probability at least $1/2$,

$$\min_{q \in Q} |\theta^\top W^\top Aq| > \frac{\min_{q \in Q} \|W^\top Aq\|_2}{2\sqrt{ek}\binom{k+1}{2}}.$$

Now fix any $i \neq j$. By Lemma 10 (first claim), $W^\top A \operatorname{diag}(w)^{1/2}$ is orthogonal, so $\|W^\top A(e_i - e_j)\|_2 = \|\operatorname{diag}(w)^{-1/2}(e_i - e_j)\|_2 = \|e_i/\sqrt{w_i} - e_j/\sqrt{w_j}\|_2 = \sqrt{1/w_i + 1/w_j}$. Similarly, for any $i \in [k]$, $\|W^\top Ae_i\|_2 = \sqrt{1/w_i}$. Therefore $\min_{q \in Q} \|W^\top Aq\|_2 = \min_{i \in [k]} \sqrt{1/w_i}$. $\square$

Let $\mathcal{E}_T := \|\widehat{T} - T\|_2/\gamma$. Let $\theta_1, \theta_2, \ldots, \theta_t$ be the random unit vectors in $\mathbb{R}^k$ drawn by the algorithm. Define $\widehat{T}[\theta_{t'}] := \widehat{M_3}[\widehat{W}, \widehat{W}, \widehat{W}\theta_{t'}]$ and $T[\theta_{t'}] := M_3[W, W, W\theta_{t'}]$. Also, let $\Delta(t') := \min\{|\lambda_i - \lambda_j| : i \neq j\} \cup \{|\lambda_i| : i \in [k]\}$ for the eigenvalues $\{\lambda_i : i \in [k]\}$ of $T[\theta_{t'}]$, and let $\widehat{\Delta}(t') := \min\{|\hat{\lambda}_i - \hat{\lambda}_j| : i \neq j\} \cup \{|\hat{\lambda}_i| : i \in [k]\}$ for the eigenvalues $\{\hat{\lambda}_i : i \in [k]\}$ of $\widehat{T}[\theta_{t'}]$.

Lemma 14 (Eigenvalue gap). *Pick any $\delta \in (0,1)$. If $t \geq \log_2(1/\delta)$, then with probability at least $1 - \delta$, the trial $\hat{\tau} := \arg\max_{t' \in [t]} \widehat{\Delta}(t')$ satisfies*

$$\widehat{\Delta}(\hat{\tau}) \geq \gamma - 2\mathcal{E}_T\gamma.$$

Proof. For each $t' \in [t]$, the eigenvalues of $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_k$ of $\widehat{T}[\theta_{t'}]$ (arranged in non-increasing order) satisfy

$$|\hat{\lambda}_i - \hat{\lambda}_j| \geq |\lambda_i - \lambda_j| - 2\|\widehat{T}[\theta_{t'}] - T[\theta_{t'}]\|_2 \geq |\lambda_i - \lambda_j| - 2\mathcal{E}_T\gamma, \quad i \neq j \tag{2}$$

where the second inequality follows from Weyl's inequality. Similarly,

$$|\hat{\lambda}_i| \geq |\lambda_i| - |\hat{\lambda}_i - \lambda_i| \geq |\lambda_i| - \mathcal{E}_T\gamma. \tag{3}$$

By Lemma 11, the eigenvalues of $T[v]$ are $v^\top W^\top Ae_i$ for $i \in [k]$. Thus, by Lemma 13, the probability that some $\tau \in [t]$ has $\Delta(\tau) > \gamma$ is at least $1 - \delta$. In this event, (2) implies that trial τ has $\widehat{\Delta}(\tau) \geq \Delta(\tau) - 2\mathcal{E}_T\gamma$, and hence $\widehat{\Delta}(\hat{\tau}) = \max_{t' \in [t]} \widehat{\Delta}(t') \geq \gamma - 2\mathcal{E}_T\gamma$. $\square$

We now just consider the trial $\hat{\tau}$ retained by the algorithm. Let $\{(v_i, \lambda_i) : i \in [k]\}$ be the eigenvector/eigenvalue pairs of $T[\theta_{\hat{\tau}}]$, and let $\{(\hat{v}_i, \hat{\lambda}_i) : i \in [k]\}$ be the eigenvector/eigenvalue pairs of $\widehat{T}[\theta_{\hat{\tau}}]$.

Lemma 15 (Accuracy of eigendecomposition). *Assume the $1 - \delta$ probability event in Lemma 14 holds, and also assume that $\mathcal{E}_T \leq 1/4$. Then there exists a permutation π on $[k]$ and signs $s_1, s_2, \ldots, s_k \in \{\pm 1\}$ such that, for all $i \in [k]$,*

$$\|v_i - s_i \hat{v}_{\pi(i)}\|_2 \leq 4\sqrt{2}\mathcal{E}_T$$
$$|\lambda_i - \hat{\lambda}_{\pi(i)}| \leq \mathcal{E}_T \gamma.$$

Proof. To simplify notation, assume the eigenvalues of $\widehat{T}[\theta]$ and $T[\theta]$ are already sorted in non-increasing order. Observe that for all $i \neq j$,

$$\begin{aligned}
|\hat{\lambda}_i - \lambda_j| = |\hat{\lambda}_i - \hat{\lambda}_j + \hat{\lambda}_j - \lambda_j| \\
\geq |\hat{\lambda}_i - \hat{\lambda}_j| - |\hat{\lambda}_j - \lambda_j| \\
\geq (\gamma - 2\mathcal{E}_T \gamma) - \mathcal{E}_T \gamma \\
\geq \gamma/4
\end{aligned}$$

where the second-to-last inequality follows by the assumption $\widehat{\Delta}(\hat{\tau}) \geq \gamma - 2\mathcal{E}_T\gamma$ and by Weyl's inequality. Therefore, the interval of radius $\gamma/4$ surrounding each eigenvalue $\hat{\lambda}_i$ of $\widehat{T}[\theta_{\hat{\tau}}]$ contains only one eigenvalue λ_i of $T[\theta_{\hat{\tau}}]$. By the Davis-Kahan $\sin(\Theta)$ theorem (Stewart and Sun, 1990, Theorem 3.4, p. 250), we have that

$$\sqrt{1 - (v_i^\top \hat{v}_i)^2} \leq 4\mathcal{E}_T.$$

Therefore, for $s_i := \text{sign}(v_i^\top \hat{v}_i)$,

$$\|v_i - s_i \hat{v}_i\|_2^2 = 2\big(1 - s_i v_i^\top \hat{v}_i\big) = 2\big(1 - |v_i^\top \hat{v}_i|\big) \leq 2\big(1 - \sqrt{1 - (4\mathcal{E}_T)^2}\big) \leq 32\mathcal{E}_T^2.$$

The bound $|\lambda_i - \hat{\lambda}_i| \leq \mathcal{E}_T \gamma$ follows simply from Weyl's inequality. $\qquad\square$

C.8   Overall error analysis

Define

$$\begin{aligned}
\kappa[M_2] &:= \varsigma_1[M_2]/\varsigma_k[M_2], \\
\epsilon_0 &:= \Big(5.5\mathcal{E}_{M_2} + 7\mathcal{E}_T\Big)/\sqrt{w_{\min}}, \\
\epsilon_1 &:= \left(\frac{1.25\|M_2\|_2^{1/2}\epsilon_0/\sqrt{w_{\min}}}{\varsigma_k[M_2]^{1/2}} + 2\mathcal{E}_{M_2} + \gamma\sqrt{w_{\min}}\mathcal{E}_T\right)/\big(\gamma\sqrt{w_{\min}}\big) \\
&= \left(\Big(6.875\kappa[M_2]^{1/2} + 2\Big)\mathcal{E}_{M_2} + \Big(8.75\kappa[M_2]^{1/2} + \gamma\sqrt{w_{\min}}\Big)\mathcal{E}_T\right)/\big(\gamma\sqrt{w_{\min}}\big).
\end{aligned}$$

Lemma 16 (Error bound). *Assume the $1 - \delta$ probability event of Lemma 14 holds, and also assume that $\mathcal{E}_{M_2} \leq 1/3$, $\mathcal{E}_T \leq 1/4$, and $\epsilon_1 \leq 1/3$. Then there exists a permutation π on $[k]$ such that*

$$\|\hat{\mu}_{\pi(i)} - \mu_i\|_2 \leq 3\|\mu_i\|_2\epsilon_1 + 2\|M_2\|_2^{1/2}\epsilon_0, \quad i \in [k].$$

Proof. To simplify notation, we assume throughout that the permutation π from Lemma 15 is the identity permutation. Let $V := [v_1|v_2|\cdots|v_k]$. We first bound $\widehat{B}s_i\hat{v}_i - \sqrt{w_i}\mu_i$. This quantity can be split into two parts: the part in the range of $\widehat{U}$, and the rest. The part in the range of $\widehat{U}$ is bounded as

$$
\begin{aligned}
\|\widehat{B}s_i\hat{v}_i - \widehat{U}\widehat{U}^\top\sqrt{w_i}\mu_i\|_2 &= \|(\widehat{U}^\top\widehat{M_2}\widehat{U})^{1/2}s_i\hat{v}_i - \widehat{U}^\top A\operatorname{diag}(w)^{1/2}e_i\|_2\\
&= \|(\widehat{U}^\top\widehat{M_2}\widehat{U})^{1/2}(s_i\hat{v}_i - v_i) + ((\widehat{U}^\top\widehat{M_2}\widehat{U})^{1/2} - \widehat{U}^\top A\operatorname{diag}(w)^{1/2}V^\top)v_i\|_2\\
&\leq \|(\widehat{U}^\top\widehat{M_2}\widehat{U})^{1/2}\|_2\|s_i\hat{v}_i - v_i\|_2 + \|(\widehat{U}^\top\widehat{M_2}\widehat{U})^{1/2} - \widehat{U}^\top A\operatorname{diag}(w)^{1/2}V^\top\|_2\\
&\leq \|\widehat{M_2}\|_2^{1/2}\|s_i\hat{v}_i - v_i\|_2 + \|(\widehat{U}^\top\widehat{M_2}\widehat{U})^{1/2} - \widehat{U}^\top A\operatorname{diag}(w)^{1/2}V^\top\|_2\\
&\leq \left((1 + \mathcal{E}_{M_2})\|M_2\|_2\right)^{1/2}4\sqrt{2}\mathcal{E}_T + \|(\widehat{U}^\top\widehat{M_2}\widehat{U})^{1/2} - \widehat{U}^\top A\operatorname{diag}(w)^{1/2}V^\top\|_2
\end{aligned}
$$

where the third inequality follows from Lemma 9 and Lemma 15. To bound the second term in the last step, recall that $W = \widehat{W}(\widehat{W}^\top M_2\widehat{W})^{-1/2}$ (using Lemma 10 to guarantee the positive definiteness of $\widehat{W}^\top M_2\widehat{W}$), so we may write $\widehat{U}^\top AV^\top$ as

$$
\begin{aligned}
\widehat{U}^\top A\operatorname{diag}(w)^{1/2}V^\top &= \widehat{U}^\top A\operatorname{diag}(w)^{1/2}(W^\top A\operatorname{diag}(w)^{1/2})^\top\\
&= \widehat{U}^\top M_2\widehat{W}(\widehat{W}^\top M_2\widehat{W})^{-1/2}\\
&= \left((\widehat{U}^\top M_2\widehat{U})^{1/2} + \widehat{U}^\top(M_2 - \widehat{M_2})\widehat{U}(\widehat{U}^\top\widehat{M_2}\widehat{U})^{-1/2}\right)(\widehat{W}^\top M_2\widehat{W})^{-1/2}.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
&\|(\widehat{U}^\top\widehat{M_2}\widehat{U})^{1/2} - \widehat{U}^\top A\operatorname{diag}(w)^{1/2}V^\top\|_2\\
&\leq \|(\widehat{U}^\top\widehat{M_2}\widehat{U})^{1/2}\left(I - (\widehat{W}^\top M_2\widehat{W})^{-1/2}\right)\|_2 + \|\widehat{U}^\top(M_2 - \widehat{M_2})\widehat{U}\|_2\|(\widehat{U}^\top\widehat{M_2}\widehat{U})^{-1/2}\|_2\|(\widehat{W}^\top M_2\widehat{W})^{-1/2}\|_2\\
&\leq \|\widehat{M_2}\|_2^{1/2}\|I - (\widehat{W}^\top M_2\widehat{W})^{-1/2}\|_2 + \frac{1}{\varsigma_k[\widehat{U}^\top\widehat{M_2}\widehat{U}]^{1/2}}\|M_2 - \widehat{M_2}\|_2\left(1 + \|I - (\widehat{W}^\top M_2\widehat{W})^{-1/2}\|_2\right)\\
&\leq (1 + \mathcal{E}_{M_2})^{1/2}\|M_2\|_2^{1/2}(3/2)\mathcal{E}_{M_2} + \frac{(1 + (3/2)\mathcal{E}_{M_2})\mathcal{E}_{M_2}\varsigma_k[M_2]}{(1 - \mathcal{E}_{M_2})^{1/2}\varsigma_k[M_2]^{1/2}}\\
&= (1 + \mathcal{E}_{M_2})^{1/2}\|M_2\|_2^{1/2}(3/2)\mathcal{E}_{M_2} + \frac{(1 + (3/2)\mathcal{E}_{M_2})}{(1 - \mathcal{E}_{M_2})^{1/2}}\varsigma_k[M_2]^{1/2}\mathcal{E}_{M_2}
\end{aligned}
$$

where the last inequality uses Lemma 9 and Lemma 10. Thus

$$
\begin{aligned}
\|\widehat{B}s_i\hat{v}_i - \widehat{U}\widehat{U}^\top\sqrt{w_i}\mu_i\|_2 \leq\ &(1 + \mathcal{E}_{M_2})^{1/2}\|M_2\|_2^{1/2}4\sqrt{2}\mathcal{E}_T\\
&+ (1 + \mathcal{E}_{M_2})^{1/2}\|M_2\|_2^{1/2}(3/2)\mathcal{E}_{M_2} + \frac{(1 + (3/2)\mathcal{E}_{M_2})}{(1 - \mathcal{E}_{M_2})^{1/2}}\varsigma_k[M_2]^{1/2}\mathcal{E}_{M_2}.
\end{aligned}
$$

Now consider the part not in the range of $\widehat{U}$. This is simply bounded as

$$
\begin{aligned}
\|(I - \widehat{U}\widehat{U}^\top)\sqrt{w_i}UU^\top\mu_i\|_2 &\leq \|I - \widehat{U}\widehat{U}^\top UU^\top\|_2\sqrt{w_i}\|\mu_i\|_2\\
&\leq (3/2)\mathcal{E}_{M_2}\sqrt{w_i}\|\mu_i\|_2
\end{aligned}
$$

using Lemma 9. Therefore, overall, we have

$$
\begin{aligned}
\|\widehat{B}\hat{v}_i - s_i\sqrt{w_i}\mu_i\|_2 &= \|\widehat{B}s_i\hat{v}_i - \sqrt{w_i}\mu_i\|_2 \\
&\leq (1+\mathcal{E}_{M_2})^{1/2}\|M_2\|_2^{1/2}4\sqrt{2}\mathcal{E}_T + (1+\mathcal{E}_{M_2})^{1/2}\|M_2\|_2^{1/2}(3/2)\mathcal{E}_{M_2} \\
&\quad + \frac{(1+(3/2)\mathcal{E}_{M_2})}{(1-\mathcal{E}_{M_2})^{1/2}}\varsigma_k[M_2]^{1/2}\mathcal{E}_{M_2} + (3/2)\mathcal{E}_{M_2}\sqrt{w_i}\|\mu_i\|_2 \leq \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}}.
\end{aligned}
$$

Since the actual estimate of μ_i is $(\hat{\lambda}_i/\theta^\top\hat{v}_i)\widehat{B}\hat{v}_i$, we need to show that $\theta^\top\hat{v}_i$ is approximately $s_i\sqrt{w_i}\hat{\lambda}_i$. Indeed,

$$
\begin{aligned}
|\theta^\top\hat{v}_i - s_i\sqrt{w_i}\hat{\lambda}_i| &= |\theta^\top\widehat{W}^\top(\widehat{B}\hat{v}_i - s_i\sqrt{w_i}\mu_i) + s_i\sqrt{w_i}\theta^\top(\widehat{W}-W)^\top\mu_i + s_i\sqrt{w_i}(\lambda_i-\hat{\lambda}_i)| \\
&\leq \|\widehat{W}\theta\|_2\|\widehat{B}\hat{v}_i - s_i\sqrt{w_i}\mu_i\|_2 + \|(\widehat{W}-W)^\top A\operatorname{diag}(w)^{1/2}e_i\|_2 + \sqrt{w_i}|\lambda_i-\hat{\lambda}_i| \\
&\leq \frac{\|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}}}{(1-\mathcal{E}_{M_2})^{1/2}\varsigma_k[M_2]^{1/2}} + (3/2)\sqrt{1+(3/2)\mathcal{E}_{M_2}}\mathcal{E}_{M_2} + \sqrt{w_i}\mathcal{E}_T\gamma \leq \epsilon_1\gamma\sqrt{w_{\min}}
\end{aligned}
$$

where the last inequality uses Lemma 10 and Lemma 15. Therefore

$$
\begin{aligned}
&\sqrt{w_i}\|(\hat{\lambda}_i/\theta^\top\hat{v}_i)\widehat{B}\hat{v}_i - \mu_i\|_2 \\
&= \|(\hat{\lambda}_i/\theta^\top\hat{v}_i)s_i\sqrt{w_i}\widehat{B}\hat{v}_i - s_i\sqrt{w_i}\mu_i\|_2 \\
&\leq |(\hat{\lambda}_i/\theta^\top\hat{v}_i)s_i\sqrt{w_i} - 1|\|\widehat{B}\hat{v}_i\|_2 + \|\widehat{B}\hat{v}_i - s_i\sqrt{w_i}\mu_i\|_2 \\
&\leq |(\hat{\lambda}_i s_i\sqrt{w_i}/\theta^\top\hat{v}_i) - 1|(\sqrt{w_i}\|\mu_i\|_2 + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}}) + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}} \\
&\leq \frac{|\hat{\lambda}_i s_i\sqrt{w_i} - \theta^\top\hat{v}_i|}{|\hat{\lambda}_i\sqrt{w_i}| - |\theta^\top\hat{v}_i - \hat{\lambda}_i s_i\sqrt{w_i}|}(\sqrt{w_i}\|\mu_i\|_2 + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}}) + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}} \\
&\leq \frac{\epsilon_1\gamma\sqrt{w_{\min}}}{|\lambda_i\sqrt{w_i}| - |\hat{\lambda}_i - \lambda_i|\sqrt{w_i} - \epsilon_1\gamma\sqrt{w_{\min}}}(\sqrt{w_i}\|\mu_i\|_2 + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}}) + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}} \\
&\leq \frac{\epsilon_1\gamma\sqrt{w_{\min}}}{\gamma\sqrt{w_i}(1-\mathcal{E}_T) - \epsilon_1\gamma\sqrt{w_{\min}}}(\sqrt{w_i}\|\mu_i\|_2 + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}}) + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}} \\
&= \frac{\epsilon_1\gamma}{\gamma(1-\mathcal{E}_T) - \epsilon_1\gamma}(\sqrt{w_i}\|\mu_i\|_2 + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}}) + \|M_2\|_2^{1/2}\epsilon_0\sqrt{w_{\min}}
\end{aligned}
$$

where the fourth inequality uses Lemma 15. We thus conclude that

$$
\begin{aligned}
\|\hat{\mu}_i - \mu_i\|_2 = \|(\hat{\lambda}_i/\theta^\top\hat{v}_i)\widehat{B}\hat{v}_i - \mu_i\|_2 &\leq \frac{\epsilon_1\gamma}{\gamma(1-\mathcal{E}_T)-\epsilon_1\gamma}(\|\mu_i\|_2 + \|M_2\|_2^{1/2}\epsilon_0) + \|M_2\|_2^{1/2}\epsilon_0 \\
&\leq 3\epsilon_1(\|\mu_i\|_2 + \|M_2\|_2^{1/2}\epsilon_0) + \|M_2\|_2^{1/2}\epsilon_0 \\
&\leq 3\|\mu_i\|_2\epsilon_1 + 2\|M_2\|_2^{1/2}\epsilon_0. \qquad \square
\end{aligned}
$$

We can now prove Theorem 3, stated below with the explicit polynomial sample complexity bound (up to constants).

Theorem 3 restated (Finite sample bound)**.** *There exists a constant $C > 0$ such that the following*

holds. Let $b_{\max} := \max_{i\in[k]} \|\mu_i\|_2$. *Pick any* $\varepsilon, \delta \in (0,1)$. *Suppose the sample size* $2n$ *satisfies*

$$n \geq C \cdot \frac{d + \log(k/\delta)}{w_{\min}} \cdot \left(\left[\frac{\kappa[M_2]^{1/2}(\sigma^2 + b_{\max}^2)}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon} \right]^2 + \left[\frac{\kappa[M_2]^{1/2}(\sigma^2 + b_{\max}^2)}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon} \right] \right)$$

$$+ C \cdot \frac{(k + \log(k/\delta))^3}{w_{\min}} \cdot \left[\frac{\kappa[M_2]^{1/2}\sigma^3}{\gamma^2\sqrt{w_{\min}}\varsigma_k[M_2]^{3/2}\varepsilon} \right]^2$$

$$+ C \cdot \frac{k + \log(k/\delta)}{w_{\min}} \cdot \left(\left[\frac{\kappa[M_2]^{1/2}\sigma^2}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon} \right]^2 + \left[\frac{\kappa[M_2]^{1/2}\sigma^2}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon} \right] + \left[\frac{\kappa[M_2]^{1/2}\sigma}{\gamma^2 w_{\min}^{3/2}\varsigma_k[M_2]^{1/2}\varepsilon} \right]^2 \right.$$

$$\left. + \left[\frac{\kappa[M_2]^{1/2}}{\gamma^2\sqrt{w_{\min}}\varepsilon} \cdot \frac{\sigma^2}{\varsigma_k[M_2]^{1/2}} \cdot \max\left\{1, \sigma^2/\varsigma_k[M_2]\right\} \right]^2 \right)$$

$$+ C \cdot \frac{k \log(1/\delta)}{w_{\min}} \left(\left[\frac{\kappa[M_2]^{1/2}}{\gamma^2\sqrt{w_{\min}}\varepsilon} \cdot \max\left\{1, \sigma^2/\varsigma_k[M_2]\right\} \right]^2 + \left[\frac{\kappa[M_2]^{1/2}}{\gamma^2 w_{\min}^2\varepsilon} \right]^2 + \left[\frac{\kappa[M_2]^{1/2}\sigma^2}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon} \right]^2 \right)$$

where

$$\gamma = \frac{1}{2\sqrt{w_{\max}}\sqrt{ek}\binom{k+1}{2}}$$

(as defined in Lemma 13). Then with probability at least $1 - 3\delta$ *over the random sample and the internal randomness of the algorithm, there exists a permutation* π *on* $[k]$ *such that*

$$\|\hat{\mu}_{\pi(i)} - \mu_i\|_2 \leq \left(\|\mu_i\|_2 + \|M_2\|_2^{1/2} \right)\varepsilon$$

for all $i \in [k]$.

Proof. Throughout, $C_1, c_1, C_2, c_2, \dots$ will represent absolute positive constants. First, observe that the sample size bound

$$n \geq C \cdot k \log(1/\delta)$$

and Lemma 4 ensure that $\mathcal{E}_w \leq 1$ (where $\mathcal{E}_w$ is defined in Lemma 6). Therefore, from Lemma 5 and Lemma 6 (together with union bounds), with probability at least $1 - \delta$,

$$\|\hat{\mu} - \mu\|_2 \leq C_1\sigma\sqrt{\frac{d + \log(k/\delta)}{w_{\min}n}} + C_1 b_{\max}\sqrt{\frac{k\log(1/\delta)}{n}},$$

$$\|\widehat{\mathcal{M}}_2 - \mathcal{M}_2\|_2 \leq C_1 \left(\sigma^2\sqrt{\frac{d + \log(k/\delta)}{w_{\min}n}} + \sigma^2\frac{d + \log(k/\delta)}{w_{\min}n} + \sigma b_{\max}\sqrt{\frac{d + \log(k/\delta)}{w_{\min}n}} \right)$$

$$+ C_1 \left(b_{\max}^2 + \sigma^2 \right)\sqrt{\frac{k\log(1/\delta)}{n}}$$

$$\leq C_1 \left(1.7\left(\sigma^2 + b_{\max}^2 \right)\sqrt{\frac{d + \log(k/\delta)}{w_{\min}n}} + \sigma^2\frac{d + \log(k/\delta)}{w_{\min}n} \right).$$

Therefore, by Lemma 7,

$$\max\{|\hat{\sigma}^2 - \sigma^2|,\ \|\widehat{M_2} - M_2\|_2\} \le 4C_1\left(1.7\left(\sigma^2 + b_{\max}^2\right)\sqrt{\frac{d + \log(k/\delta)}{w_{\min}n}} + \sigma^2\frac{d + \log(k/\delta)}{w_{\min}n}\right)$$

$$+ 4C_1\|\mu\|_2\left(\sigma\sqrt{\frac{d + \log(k/\delta)}{w_{\min}n}} + b_{\max}\sqrt{\frac{k\log(1/\delta)}{n}}\right)$$

$$+ 2C_1^2\left(\sigma\sqrt{\frac{d + \log(k/\delta)}{w_{\min}n}} + b_{\max}\sqrt{\frac{k\log(1/\delta)}{n}}\right)^2$$

$$\le C_2(\sigma^2 + b_{\max}^2)\left(\sqrt{\frac{d + \log(k/\delta)}{w_{\min}n}} + \frac{d + \log(k/\delta)}{w_{\min}n}\right).$$

The sample size bound

$$n \ge C \cdot \frac{d + \log(k/\delta)}{w_{\min}} \cdot \left(\left[\frac{\kappa[M_2]^{1/2}(\sigma^2 + b_{\max}^2)}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon}\right]^2 + \left[\frac{\kappa[M_2]^{1/2}(\sigma^2 + b_{\max}^2)}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon}\right]\right),$$

ensures that

$$\max\left\{\frac{|\hat{\sigma}^2 - \sigma^2|}{\varsigma_k[M_2]}, \mathcal{E}_{M_2}\right\} \le c_1\frac{\gamma^2 w_{\min}}{\kappa[M_2]^{1/2}}\varepsilon \le 1/3. \tag{4}$$

Now condition on the above event and take $\widehat{W}$ as given. By Lemma 10,

$$\|\widehat{W}\|_2 \le \sqrt{1.5/\varsigma_k[M_2]},$$

$$\max_{i\in[k]}\|\widehat{W}^\top\mu_i\|_2 \le \|\widehat{W}^\top A\operatorname{diag}(w)^{1/2}\|_2/\sqrt{w_{\min}} \le \sqrt{1.5/w_{\min}},$$

$$\max_{i\in[k]}\|\widehat{W}^\top\mathcal{M}_{2,i}\widehat{W}\|_2 \le 1.5/\varsigma_k[M_2] + 1.5/w_{\min},$$

$$\max_{i\in[k]}\|\mathcal{M}_{3,i}[\widehat{W},\widehat{W},\widehat{W}]\|_2 \le (1.5/w_{\min})^{3/2} + 3\sigma^2\sqrt{1.5/w_{\min}}1.5/\varsigma_k[M_2].$$

Therefore, Lemma 5 and Lemma 6 imply that with probability at least $1 - \delta$,

$$\|\widehat{W}^\top(\hat{\underline{\mu}} - \mu)\|_2 \le C_3\frac{\sigma}{\varsigma_k[M_2]^{1/2}}\sqrt{\frac{k + \log(k/\delta)}{w_{\min}n}} + C_3\frac{1}{\sqrt{w_{\min}}}\sqrt{\frac{k\log(1/\delta)}{n}},$$

$$\|\underline{\widehat{\mathcal{M}_3}}[\widehat{W},\widehat{W},\widehat{W}] - \mathcal{M}_3[\widehat{W},\widehat{W},\widehat{W}]\|_2 \le C_3\frac{\sigma^3}{\varsigma_k[M_2]^{3/2}}\sqrt{\frac{(k + \log(k/\delta))^3}{w_{\min}n}}$$

$$+ C_3\frac{\sigma^2}{\sqrt{w_{\min}}\varsigma_k[M_2]}\left[\sqrt{\frac{k + \log(k/\delta)}{w_{\min}n}} + \frac{k + \log(k/\delta)}{w_{\min}n}\right]$$

$$+ C_3\frac{\sigma}{w_{\min}\varsigma_k[M_2]^{1/2}}\sqrt{\frac{k + \log(k/\delta)}{w_{\min}n}}$$

$$+ C_3\left(\frac{1}{w_{\min}^{3/2}} + \frac{\sigma^2}{\sqrt{w_{\min}}\varsigma_k[M_2]}\right)\sqrt{\frac{k\log(1/\delta)}{n}}.$$

Therefore, by Lemma 8 and Lemma 12,

$$\|\widehat{T} - T\|_2 \leq \|\underline{\widehat{\mathcal{M}_3}}[\widehat{W}, \widehat{W}, \widehat{W}] - \mathcal{M}_3[\widehat{W}, \widehat{W}, \widehat{W}]\|_2$$

$$+ \frac{4.5}{\varsigma_k[M_2]}\left(\|\widehat{W}^\top(\underline{\hat{\mu}} - \mu)\|_2 + \sqrt{1.5/w_{\min}}\right)|\hat{\sigma}^2 - \sigma| + \frac{1.5\sigma^2}{\varsigma_k[M_2]}\|\widehat{W}^\top(\underline{\hat{\mu}} - \mu)\|_2. \qquad (5)$$

The sample size bound

$$n \geq C \cdot \frac{k + \log(k/\delta)}{w_{\min}} \cdot \left[\frac{\kappa[M_2]^{1/2}}{\gamma^2\sqrt{w_{\min}}\varepsilon} \cdot \frac{\sigma^2}{\varsigma_k[M_2]^{1/2}} \cdot \max\left\{1, \sigma^2/\varsigma_k[M_2]\right\}\right]^2$$

$$+ C \cdot \frac{k\log(1/\delta)}{w_{\min}} \cdot \left[\frac{\kappa[M_2]^{1/2}}{\gamma^2\sqrt{w_{\min}}\varepsilon} \cdot \max\left\{1, \sigma^2/\varsigma_k[M_2]\right\}\right]^2$$

ensures

$$\max\left\{1, \sigma^2/\varsigma_k[M_2]^{1/2}\right\}\frac{\|\widehat{W}^\top(\underline{\hat{\mu}} - \mu)\|_2}{\gamma} \leq c_2\frac{\gamma\sqrt{w_{\min}}}{\kappa[M_2]^{1/2}}\varepsilon \leq 1. \qquad (6)$$

Furthermore, the sample size bound

$$n \geq C \cdot \frac{(k + \log(k/\delta))^3}{w_{\min}} \cdot \left[\frac{\kappa[M_2]^{1/2}\sigma^3}{\gamma^2\sqrt{w_{\min}}\varsigma_k[M_2]^{3/2}\varepsilon}\right]^2$$

$$+ C \cdot \frac{k + \log(k/\delta)}{w_{\min}} \cdot \left(\left[\frac{\kappa[M_2]^{1/2}\sigma^2}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon}\right]^2 + \left[\frac{\kappa[M_2]^{1/2}\sigma^2}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon}\right]\right)$$

$$+ C \cdot \frac{k + \log(k/\delta)}{w_{\min}} \cdot \left[\frac{\kappa[M_2]^{1/2}\sigma}{\gamma^2 w_{\min}^{3/2}\varsigma_k[M_2]^{1/2}\varepsilon}\right]^2$$

$$+ C \cdot k\log(1/\delta) \cdot \left(\left[\frac{\kappa[M_2]^{1/2}}{\gamma^2 w_{\min}^2\varepsilon}\right]^2 + \left[\frac{\kappa[M_2]^{1/2}\sigma^2}{\gamma^2 w_{\min}\varsigma_k[M_2]\varepsilon}\right]^2\right)$$

ensures

$$\frac{\|\underline{\widehat{\mathcal{M}_3}}[\widehat{W}, \widehat{W}, \widehat{W}] - \mathcal{M}_3[\widehat{W}, \widehat{W}, \widehat{W}]\|_2}{\gamma} \leq c_2\frac{\gamma\sqrt{w_{\min}}}{\kappa[M_2]^{1/2}}\varepsilon. \qquad (7)$$

Using the inequalities (6), (7), and (4) with (5) gives

$$\mathcal{E}_T = \frac{\|\widehat{T} - T\|_2}{\gamma} \leq c_3\frac{\gamma\sqrt{w_{\min}}}{\kappa[M_2]^{1/2}}\varepsilon. \qquad (8)$$

Since $t = \log_2(1/\delta)$, the inequality from Lemma 14 holds with probability at least $1 - \delta$ over the internal randomness of the algorithm. Therefore, using (4) and (8) with Lemma 16 gives in this event:

$$\|\hat{\mu}_{\pi(i)} - \mu_i\|_2 \leq C_4 \cdot \|\mu_i\|_2 \cdot \left(\mathcal{E}_{M_2} + \mathcal{E}_T\right) \cdot \frac{\kappa[M_2]^{1/2}}{\gamma\sqrt{w_{\min}}}$$

$$+ C_4 \cdot \|M_2\|_2^{1/2} \cdot \left(\mathcal{E}_{M_2} + \mathcal{E}_T\right) \cdot \frac{1}{\sqrt{w_{\min}}}$$

$$\leq \left(\|\mu_i\|_2 + \|M_2\|_2^{1/2}\right)\varepsilon$$

for all $i \in [k]$, for some permutation π on $[k]$. $\qquad\square$

D   Probability tail inequalities

We recall and derive some probability tail inequalities used in the analysis.

Lemma 17 (Dasgupta and Gupta, 2003; Anandkumar et al., 2012b). *Pick any $\delta \in (0,1)$, matrix $X \in \mathbb{R}^{p \times p}$, and finite subset $Q \subseteq \mathbb{R}^p$. If $\theta \in \mathbb{R}^p$ be a random vector distributed uniformly over the unit sphere in $\mathbb{R}^p$, then*

$$\Pr\left[\min_{q \in Q} |\theta^\top X q| > \frac{\min_{q \in Q} \|Xq\|_2 \cdot \delta}{\sqrt{ep}|Q|}\right] \geq 1 - \delta.$$

Lemma 18 (Laurent and Massart, 2000). *Let $z_1^2, z_2^2, \ldots, z_m^2$ be i.i.d. χ^2 random variables, each with one degree of freedom. Then for any $\delta \in (0,1)$,*

$$\Pr\left[\sum_{i=1}^{m} z_i^2 > m + 2\sqrt{m \ln(1/\delta)} + 2\ln(1/\delta)\right] \leq \delta.$$

Lemma 19 (Litvak et al., 2005; Hsu et al., 2012b). *Let $y_1, y_2, \ldots, y_m$ be i.i.d. $N(0, I)$ random vectors in $\mathbb{R}^p$. Then for any $\epsilon_0 \in (1, 1/2)$ and $\delta \in (0,1)$,*

$$\Pr\left[\left\|\frac{1}{m}\sum_{i=1}^{m} y_i y_i^\top - I\right\|_2 > \frac{1}{1 - 2\epsilon_0}\left(\sqrt{\frac{32\ln((1 + 2/\epsilon_0)^p/\delta)}{m}} + \frac{2\ln((1 + 2/\epsilon_0)^p/\delta)}{m}\right)\right] \leq \delta.$$

Lemma 20 (Sums of cubes of normal random variables). *Let $z_1, z_2, \ldots, z_m$ be i.i.d. $N(0,1)$ random variables. Then for any $\delta \in (0,1)$,*

$$\Pr\left[\left|\sum_{i=1}^{m} z_i^3\right| > \sqrt{27e^3 m \lceil \ln(1/\delta)\rceil^3}\right] \leq \delta.$$

Proof. We use Markov's inequality via the p-th moment to derive the tail inequality. Pick some even integer $p \geq 2$, and observe that

$$\mathbb{E}\left[\left(\sum_{i=1}^{m} z_i^3\right)^p\right] = \sum_{i_1, i_2, \ldots, i_p \in [m]} \mathbb{E}\left[\prod_{j=1}^{p} z_{i_j}^3\right].$$

By the independence of the z_i's, a term in the summation is zero if any index $i \in [m]$ is selected an odd number of times (*i.e.*, $|\{j \in [p] : i_j = i\}|$ is odd, for any $i \in [m]$). Therefore the summation can be written as

$$\sum_{p_1 + \cdots + p_m = p/2} \mathbb{E}\left[\prod_{i=1}^{m} z_i^{6p_i}\right] = \sum_{p_1 + \cdots + p_m = p/2} \prod_{i=1}^{m} \mathbb{E}\left[z_i^{6p_i}\right] = \sum_{p_1 + \cdots + p_m = p/2} \prod_{i=1}^{m} (6p_i - 1)!!$$

where the summations are over non-negative integers $p_1, p_2, \ldots, p_m$ that sum to $p/2$, and $n!!$ is the product of all odd integers between 1 and n; the last step uses the well-known fact that $\mathbb{E}[z^k] = (k-1)!!$ for a standard normal random variable z. As $p_i \leq p/2$ for each $i \in [m]$, the product can be crudely bounded by $(3p)^{3p/2}$, and hence the sum is bounded by

$$(3p)^{3p/2}\binom{p/2 + m - 1}{p/2} \leq (27p^3)^{p/2}\left(\frac{(p/2 + m - 1)e}{p/2}\right)^{p/2} \leq (27ep^3 m)^{p/2}.$$

By Markov's inequality, for any $t > 0$,

$$\Pr\left[\left|\sum_{i=1}^m z_i^3\right| > t\right] \leq t^{-p}\mathbb{E}\left[\left(\sum_{i=1}^m z_i^3\right)^p\right] \leq \left(\frac{\sqrt{27ep^3m}}{t}\right)^p.$$

The bound is at most δ for $t := e\sqrt{27em\lceil\ln(1/\delta)\rceil^3}$ and $p := \lceil\ln(1/\delta)\rceil$. $\qquad\square$

Lemma 21 (Third-order tensor of normal random vectors). *Let $y_1, y_2, \ldots, y_m$ be i.i.d. $N(0, I)$ random vectors in $\mathbb{R}^p$. Then for any $\epsilon_0 \in (1, 1/3)$ and $\delta \in (0, 1)$,*

$$\Pr\left[\left\|\frac{1}{m}\sum_{i=1}^m y_i \otimes y_i \otimes y_i\right\|_2 > \frac{1}{1 - 3\epsilon_0}\sqrt{\frac{27e^3\lceil\ln((1 + 2/\epsilon_0)^p/\delta)\rceil^3}{m}}\right] \leq \delta.$$

Proof. We follow the covering approach of Litvak et al. (2005). Let $Q \subseteq \{x \in \mathbb{R}^p : \|x\|_2 = 1\}$ be an ϵ_0-cover of the unit sphere in $\mathbb{R}^p$ of cardinality at most $(1 + 2/\epsilon_0)^p$, which can be shown to exist by a standard volume argument (Pisier, 1989). Let $Y := m^{-1}\sum_{i=1}^m y_i \otimes y_i \otimes y_i$ and $\epsilon := (27e^3\lceil\ln(|Q|/\delta)\rceil^3/m)^{1/2}$. Since $y_i^\top q$ is distributed as $N(0, 1)$ for any $q \in Q$, it follows from Lemma 20 and a union bound that $\Pr[\exists q \in Q \bullet |Y[q, q, q]| > \epsilon] \leq \delta$. Henceforth we assume $\forall q \in Q \bullet |Y[q, q, q] \leq \epsilon$. Now pick any unit vector u such that $|Y[u, u, u]|$ is maximized (*i.e.*, $\|Y\|_2 = |Y[u, u, u]|$), choose $q \in Q$ such that $\|q - u\|_2 \leq \epsilon_0$, and set $\Delta := u - q$ and $\bar{\Delta} := \Delta/\|\Delta\|_2$. Then

$$\begin{aligned}
\|Y\|_2 &= |Y[u, u, u]| \\
&= |Y[\Delta, u, u] + Y[q, \Delta, u] + Y[q, q, \Delta] + Y[q, q, q]| \\
&\leq \epsilon_0(Y[\bar{\Delta}, u, u] + Y[q, \bar{\Delta}, u] + Y[q, q, \bar{\Delta}]) + \epsilon
\end{aligned}$$

by the triangle inequality and facts $\|\Delta\|_2 \leq \epsilon_0$ and $|Y[q, q, q]| \leq \epsilon$. Since Y has the form $Y = \sum_{j=1}^m \tilde{y}_j \otimes \tilde{y}_j \otimes \tilde{y}_j$ for vectors $\tilde{y}_j := m^{-1/3}y_j \in \mathbb{R}^m$, it follows that

$$\begin{aligned}
\sup_{\|u\|_2=\|v\|_2=\|w\|_2=1} |Y[u, v, w]| &= \sup_{\|u\|_2=\|v\|_2=\|w\|_2=1}\left|\sum_{j=1}^r (u^\top\tilde{y}_j)(v^\top\tilde{y}_j)(w^\top\tilde{y}_j)\right| \\
&= \sup_{\|u\|_2=\|v\|_2=\|w\|_2=1}\left|u^\top\tilde{Y}\,\mathrm{diag}(w^\top\tilde{Y})\tilde{Y}^\top v\right| \\
&= \sup_{\|w\|_2=1}\|\tilde{Y}\,\mathrm{diag}(w^\top\tilde{Y})\tilde{Y}^\top\|_2 \\
&= \sup_{\|u\|_2=\|w\|_2=1} |Y[u, u, w]|
\end{aligned}$$

where $\tilde{Y} = [\tilde{y}_1|\tilde{y}_2|\cdots|\tilde{y}_m] \in \mathbb{R}^{p \times m}$—*i.e.*, we can take the unit vectors u and v achieving $|Y[u, v, w]| = \|Y\|_2$ to be the same. By symmetry, $\sup_{\|u\|_2=1} |Y[u, u, u]| = \|Y\|_2$. Therefore

$$\|Y\|_2 \leq \epsilon_0(Y[\bar{\Delta}, u, u] + Y[q, \bar{\Delta}, u] + Y[q, q, \bar{\Delta}]) + \epsilon \leq 3\epsilon_0\|Y\|_2 + \epsilon$$

which implies $\|Y\|_2 \leq \epsilon/(1 - 3\epsilon_0)$. This proves that

$$\Pr[\|Y\|_2 \leq \epsilon/(1 - 3\epsilon_0)] \geq \Pr[\forall q \in Q \bullet |Y[q, q, q]| \leq \epsilon] \geq 1 - \delta$$

as required. $\qquad\square$