A Method of Moments for Mixture Models and Hidden Markov Models

Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
University of California, Irvine

Daniel Hsu DAHSU@MICROSOFT.COM
Microsoft Research

Sham M. Kakade SKAKADE@MICROSOFT.COM
Microsoft Research

Editor: Shie Mannor, Nathan Srebro, Bob Williamson

Abstract

Mixture models are a fundamental tool in applied statistics and machine learning for treating data taken from multiple subpopulations. The current practice for estimating the parameters of such models relies on local search heuristics (*e.g.*, the EM algorithm) which are prone to failure, and existing consistent methods are unfavorable due to their high computational and sample complexity which typically scale exponentially with the number of mixture components. This work develops an efficient *method of moments* approach to parameter estimation for a broad class of high-dimensional mixture models with many components, including multi-view mixtures of Gaussians (such as mixtures of axis-aligned Gaussians) and hidden Markov models. The new method leads to rigorous unsupervised learning results for mixture models that were not achieved by previous works; and, because of its simplicity, it offers a viable alternative to EM for practical deployment.

1. Introduction

Mixture models are a fundamental tool in applied statistics and machine learning for treating data taken from multiple subpopulations (Titterington et al., 1985). In a mixture model, the data are generated from a number of possible sources, and it is of interest to identify the nature of the individual sources. As such, estimating the unknown parameters of the mixture model from sampled data—especially the parameters of the underlying constituent distributions—is an important statistical task. For most mixture models, including the widely used mixtures of Gaussians and hidden Markov models (HMMs), the current practice relies on the Expectation-Maximization (EM) algorithm, a local search heuristic for maximum likelihood estimation. However, EM has a number of well-documented drawbacks regularly faced by practitioners, including slow convergence and suboptimal local optima (Redner and Walker, 1984).

An alternative to maximum likelihood and EM, especially in the context of mixture models, is the *method of moments* approach. The method of moments dates back to the origins of mixture models with Pearson's solution for identifying the parameters of a mixture of two univariate Gaussians (Pearson, 1894). In this approach, model parameters are chosen to specify a distribution whose p-th order moments, for several values of p, are equal to the corresponding empirical moments observed in the data. Since Pearson's work, the method of moments has been studied and adapted for a variety of problems; their intuitive appeal is also complemented with a guarantee of statistical consistency under mild conditions. Unfortunately, the method often runs into trouble with large mixtures of high-dimensional distributions. This is because the equations determining the parameters are typically based on moments of order equal to the number of model parameters, and high-order moments are exceedingly difficult to estimate accurately due to their large variance.

This work develops a computationally efficient method of moments based on only low-order moments that can be used to estimate the parameters of a broad class of high-dimensional mixture models with many components. The resulting estimators can be implemented with standard numerical linear algebra routines (singular value and eigenvalue decompositions), and the estimates have low variance because they only involve low-order moments. The class of models covered by the method includes certain multivariate Gaussian mixture models and HMMs, as well as mixture models with no explicit likelihood equations. The method exploits the availability of multiple indirect "views" of a model's underlying latent variable that determines the source distribution, although the notion of a "view" is rather general. For instance, in an HMM, the past, present, and future observations can be thought of as different noisy views of the present hidden state; in a mixture of product distributions (such as axis-aligned Gaussians), the coordinates in the output space can be partitioned (say, randomly) into multiple non-redundant "views". The new method of moments leads to unsupervised learning guarantees for mixture models under mild rank conditions that were not achieved by previous works; in particular, the sample complexity of accurate parameter estimation is shown to be polynomial in the number of mixture components and other relevant quantities. Finally, due to its simplicity, the new method (or variants thereof) also offers a viable alternative to EM and maximum likelihood for practical deployment.

1.1. Related work

Gaussian mixture models. The statistical literature on mixture models is vast (a more thorough treatment can be found in the texts of Titterington et al. (1985) and Lindsay (1995)), and many advances have been made in computer science and machine learning over the past decade or so, in part due to their importance in modern applications. The use of mixture models for clustering data comprises a large part of this work, beginning with the work of Dasgupta (1999) on learning mixtures of k well-separated d-dimensional Gaussians. This and subsequent work (Arora and Kannan, 2001; Dasgupta and Schulman, 2007; Vempala and Wang, 2002; Kannan et al., 2005; Achlioptas and McSherry, 2005; Chaudhuri and Rao, 2008; Brubaker and Vempala, 2008; Chaudhuri et al., 2009) have focused on efficient algorithms that provably recover the parameters of the constituent Gaussians from data generated by such a mixture distribution, provided that the distance between each pair of means is sufficiently large (roughly either d^c or k^c times the standard deviation of the Gaussians, for some $c > 0$). Such separation conditions are natural to expect in many clustering applications, and a number of spectral projection techniques have been shown to enhance the separation (Vempala and Wang, 2002; Kannan et al., 2005; Brubaker and Vempala, 2008; Chaudhuri et al., 2009). More recently, techniques have been developed for learning mixtures of Gaussians without any separation condition (Kalai et al., 2010; Belkin and Sinha, 2010; Moitra and Valiant, 2010), although the computational and sample complexities of these methods grow exponentially with the number of mixture components k. This dependence has also been shown to be inevitable without further assumptions (Moitra and Valiant, 2010).

Method of moments. The latter works of Belkin and Sinha (2010), Kalai et al. (2010), and Moitra and Valiant (2010) (as well as the algorithms of Feldman et al. (2005, 2006) for a related but different learning objective) can be thought of as modern implementations of the method of moments, and their exponential dependence on k is not surprising given the literature on other moment methods for mixture models. In particular, a number of moment methods for both discrete and continuous mixture models have been developed using techniques such as the Vandermonde decompositions of Hankel matrices (Lindsay, 1989; Lindsay and Basak, 1993; Boley et al., 1997; Gravin et al., 2012). In these methods, following the spirit of Pearson's original solution, the model parameters are derived from the roots of polynomials whose coefficients are based on moments up to the $\Omega(k)$-th order. The accurate estimation of such moments generally has computational and sample complexity exponential in k.

Spectral approach to parameter estimation with low-order moments. The present work is based on a notable exception to the above situation, namely Chang's spectral decomposition technique for discrete Markov models of evolution (Chang, 1996) (see also Mossel and Roch (2006) and Hsu et al. (2009) for adaptations to other discrete mixture models such as discrete HMMs). This spectral technique depends only on moments up to the third-order; consequently, the resulting algorithms have computational and sample complexity that scales only polynomially in the number of mixture components k. The success of the technique depends on a certain rank condition of the transition matrices; but this condition is much milder than separation conditions of clustering works, and it remains sufficient even when the dimension of the observation space is very large (Hsu et al., 2009). In this work, we extend Chang's spectral technique to develop a general method of moments approach to parameter estimation, which is applicable to a large class of mixture models and HMMs with both discrete and continuous component distributions in high-dimensional spaces. Like the moment methods of Moitra and Valiant (2010) and Belkin and Sinha (2010), our algorithm does not require a separation condition; but unlike those previous methods, the algorithm has computational and sample complexity polynomial in k.

Some previous spectral approaches for related learning problems only use second-order moments, but these approaches can only estimate a subspace containing the parameter vectors and not the parameters themselves (McSherry, 2001). Indeed, it is known that the parameters of even very simple discrete mixture models are not generally identifiable from only second-order moments (Chang, 1996)[1]. We note that moments beyond the second-order (specifically, fourth-order moments) have been exploited in the methods of Frieze et al. (1996) and Nguyen and Regev (2009) for the problem of learning a parallelepiped from random samples, and that these methods are very related to techniques used for

1. See Appendix G for an example of Chang (1996) demonstrating the non-identifiability of parameters from only second-order moments in a simple class of Markov models.

independent component analysis (Hyvärinen and Oja, 2000). Adapting these techniques for other parameter estimation problems is an enticing possibility.

Multi-view learning. The spectral technique we employ depends on the availability of multiple views, and such a multi-view assumption has been exploited in previous works on learning mixtures of well-separated distributions (Chaudhuri and Rao, 2008; Chaudhuri et al., 2009). In these previous works, a projection based on a *canonical correlation analysis* (Hotelling, 1935) between two views is used to reinforce the separation between the mixture components, and to cancel out noise orthogonal to the separation directions. The present work, which uses similar correlation-based projections, shows that the availability of a third view of the data can remove the separation condition entirely. The multi-view assumption substantially generalizes the case where the component distributions are product distributions (such as axis-aligned Gaussians), which has been previously studied in the literature (Dasgupta, 1999; Vempala and Wang, 2002; Chaudhuri and Rao, 2008; Feldman et al., 2005, 2006); the combination of this and a non-degeneracy assumption is what allows us to avoid the sample complexity lower bound of Moitra and Valiant (2010) for Gaussian mixture models. The multi-view assumption also naturally arises in many applications, such as in multimedia data with (say) text, audio, and video components (Blaschko and Lampert, 2008; Chaudhuri et al., 2009); as well as in linguistic data, where the different words in a sentence or paragraph are considered noisy predictors of the underlying semantics (Gale et al., 1992). In the vein of this latter example, we consider estimation in a simple bag-of-words document topic model as a warm-up to our general method; even this simpler model illustrates the power of pair-wise and triple-wise (*i.e.*, bigram and trigram) statistics that were not exploited by previous works on multi-view learning.

1.2. Outline

Section 2 first develops the method of moments in the context of a simple discrete mixture model motivated by document topic modeling; an explicit algorithm and convergence analysis are also provided. The general setting is considered in Section 3, where the main algorithm and its accompanying correctness and efficiency guarantee are presented. Applications to learning multi-view mixtures of Gaussians and HMMs are discussed in Section 4. Proofs and additional discussion are provided in the appendices.

1.3. Notations

The standard inner product between vectors $\vec{u}$ and $\vec{v}$ is denoted by $\langle \vec{u}, \vec{v} \rangle = \vec{u}^\top \vec{v}$. We denote the p-norm of a vector $\vec{v}$ by $\|\vec{v}\|_p$. For a matrix $A \in \mathbb{R}^{m \times n}$, we let $\|A\|_2$ denote its spectral norm $\|A\|_2 := \sup_{\vec{v} \neq \vec{0}} \|A\vec{v}\|_2 / \|\vec{v}\|_2$, $\|A\|_{\mathrm{F}}$ denote its Frobenius norm, $\sigma_i(A)$ denote the i-th largest singular value, and $\kappa(A) := \sigma_1(A)/\sigma_{\min(m,n)}(A)$ denote its condition number. Let $\Delta^{n-1} := \{(p_1, p_2, \dots, p_n) \in \mathbb{R}^n : p_i \geq 0 \ \forall i, \ \sum_{i=1}^n p_i = 1\}$ denote the probability simplex in $\mathbb{R}^n$, and let $\mathcal{S}^{n-1} := \{\vec{u} \in \mathbb{R}^n : \|\vec{u}\|_2 = 1\}$ denote the unit sphere in $\mathbb{R}^n$. Let $\vec{e}_i \in \mathbb{R}^d$ denote the i-th coordinate vector whose i-th entry is 1 and the rest are zero. Finally, for a positive integer n, let $[n] := \{1, 2, \dots, n\}$.

2. Warm-up: bag-of-words document topic modeling

We first describe our method of moments in the simpler context of bag-of-words models for documents. Proofs of lemmas and theorems in this section are given in Appendix A.

2.1. Setting

Suppose a document corpus can be partitioned by topic, with each document being assigned a single topic. Further, suppose the words in a document are drawn independently from a multinomial distribution corresponding to the document's topic. Let k be the number of distinct topics in the corpus, d be the number of distinct words in the vocabulary, and $\ell \geq 3$ be the number of words in each document (so the documents may be quite short).

The generative process for a document is given as follows:

1. The document's topic is drawn according to the multinomial distribution specified by the probability vector $\vec{w} = (w_1, w_2, \dots, w_k) \in \Delta^{k-1}$. This is modeled as a discrete random variable h such that

$$\Pr[h = j] = w_j, \quad j \in [k].$$

2. Given the topic h, the document's ℓ words are drawn independently according to the multinomial distribution specified by the probability vector $\vec{\mu}_h \in \Delta^{d-1}$. The random vectors $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_\ell \in \mathbb{R}^d$ represent the ℓ words by setting

$$\vec{x}_v = \vec{e}_i \iff \text{the } v\text{-th word in the document is } i, \quad i \in [d]$$

(the reason for this encoding of words will become clear in the next section). Therefore, for each word $v \in [\ell]$ in the document,

$$\Pr[\vec{x}_v = \vec{e}_i | h = j] = \langle \vec{e}_i, \vec{\mu}_j \rangle = M_{i,j}, \quad i \in [d], j \in [k],$$

where $M \in \mathbb{R}^{d \times k}$ is the matrix of conditional probabilities $M := [\vec{\mu}_1 | \vec{\mu}_2 | \cdots | \vec{\mu}_k]$.

This probabilistic model has the conditional independence structure depicted in Figure 3(a) as a directed graphical model.

We assume the following condition on $\vec{w}$ and M.

Condition 1 (Non-degeneracy: document topic model) *$w_j > 0$ for all $j \in [k]$, and M has rank k.*

This condition requires that each topic has non-zero probability, and also prevents any topic's word distribution from being a mixture of the other topics' word distributions.

2.2. Pair-wise and triple-wise probabilities

Define $\text{Pairs} \in \mathbb{R}^{d \times d}$ to be the matrix of pair-wise probabilities whose (i,j)-th entry is

$$\text{Pairs}_{i,j} := \Pr[\vec{x}_1 = \vec{e}_i, \vec{x}_2 = \vec{e}_j], \quad i, j \in [d].$$

Also define Triples $\in \mathbb{R}^{d \times d \times d}$ to be the third-order tensor of triple-wise probabilities whose (i, j, κ)-th entry is

$$\text{Triples}_{i,j,\kappa} := \Pr[\vec{x}_1 = \vec{e}_i, \vec{x}_2 = \vec{e}_j, \vec{x}_3 = \vec{e}_\kappa], \quad i, j, \kappa \in [d].$$

The identification of words with coordinate vectors allows Pairs and Triples to be viewed as expectations of tensor products of the random vectors $\vec{x}_1$, $\vec{x}_2$, and $\vec{x}_3$:

$$\text{Pairs} = \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2] \quad \text{and} \quad \text{Triples} = \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2 \otimes \vec{x}_3]. \tag{1}$$

We may also view Triples as a linear operator Triples: $\mathbb{R}^d \to \mathbb{R}^{d \times d}$ given by

$$\text{Triples}(\vec{\eta}) := \mathbb{E}[(\vec{x}_1 \otimes \vec{x}_2)\langle \vec{\eta}, \vec{x}_3 \rangle].$$

In other words, the (i, j)-th entry of Triples$(\vec{\eta})$ for $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_d)$ is

$$\text{Triples}(\vec{\eta})_{i,j} = \sum_{x=1}^{d} \eta_x \text{Triples}_{i,j,x} = \sum_{x=1}^{d} \eta_x \text{Triples}(\vec{e}_x)_{i,j}.$$

The following lemma shows that Pairs and Triples$(\vec{\eta})$ can be viewed as certain matrix products involving the model parameters M and $\vec{w}$.

Lemma 1 Pairs $= M \operatorname{diag}(\vec{w}) M^\top$ *and* Triples$(\vec{\eta}) = M \operatorname{diag}(M^\top \vec{\eta}) \operatorname{diag}(\vec{w}) M^\top$ *for all* $\vec{\eta} \in \mathbb{R}^d$.

2.3. Observable operators and their spectral properties

The pair-wise and triple-wise probabilities can be related in a way that essentially reveals the conditional probability matrix M. This is achieved through a matrix called an "observable operator". Similar observable operators were previously used to characterize multiplicity automata (Schützenberger, 1961; Jaeger, 2000) and, more recently, for learning discrete HMMs (via an operator parameterization) (Hsu et al., 2009).

Lemma 2 *Assume Condition 1. Let* $U \in \mathbb{R}^{d \times k}$ *and* $V \in \mathbb{R}^{d \times k}$ *be matrices such that both* $U^\top M$ *and* $V^\top M$ *are invertible. Then* $U^\top \text{Pairs} V$ *is invertible, and for all* $\vec{\eta} \in \mathbb{R}^d$*, the "observable operator"* $B(\vec{\eta}) \in \mathbb{R}^{k \times k}$*, given by*

$$B(\vec{\eta}) := (U^\top \text{Triples}(\vec{\eta}) V)(U^\top \text{Pairs} V)^{-1},$$

satisfies

$$B(\vec{\eta}) = (U^\top M) \operatorname{diag}(M^\top \vec{\eta})(U^\top M)^{-1}.$$

The matrix $B(\vec{\eta})$ is called "observable" because it is only a function of the observable variables' joint probabilities (*e.g.*, $\Pr[\vec{x}_1 = \vec{e}_i, \vec{x}_2 = \vec{e}_j]$). In the case $\vec{\eta} = \vec{e}_x$ for some $x \in [d]$, the matrix $B(\vec{e}_x)$ is *similar* (in the linear algebraic sense) to the diagonal matrix $\operatorname{diag}(M^\top \vec{e}_x)$; the collection of matrices $\{\operatorname{diag}(M^\top \vec{e}_x) : x \in [d]\}$ (together with $\vec{w}$) can be used to compute joint probabilities under the model (see, *e.g.*, Hsu et al. (2009)). Note that the columns of $U^\top M$ are eigenvectors of $B(\vec{e}_x)$, with the j-th column having an associated eigenvalue equal to $\Pr[\vec{x}_v = x | h = j]$. If the word x has distinct probabilities under every topic, then $B(\vec{e}_x)$ has exactly k distinct eigenvalues, each having geometric multiplicity one and corresponding to a column of $U^\top M$.

Algorithm A

1. Obtain empirical frequencies of word pairs and triples from a given sample of documents, and form the tables $\widehat{\text{Pairs}} \in \mathbb{R}^{d \times d}$ and $\widehat{\text{Triples}} \in \mathbb{R}^{d \times d \times d}$ corresponding to the population quantities Pairs and Triples.
2. Let $\hat{U} \in \mathbb{R}^{d \times k}$ and $\hat{V} \in \mathbb{R}^{d \times k}$ be, respectively, matrices of orthonormal left and right singular vectors of $\widehat{\text{Pairs}}$ corresponding to its top k singular values.
3. Pick $\vec{\eta} \in \mathbb{R}^d$ (see remark in the main text), and compute the right eigenvectors $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_k$ (of unit Euclidean norm) of
$$\hat{B}(\vec{\eta}) := (\hat{U}^\top \widehat{\text{Triples}}(\vec{\eta}) \hat{V})(\hat{U}^\top \widehat{\text{Pairs}} \hat{V})^{-1}.$$
(Fail if not possible.)
4. Let $\hat{\mu}_j := \hat{U}\hat{\xi}_j / \langle \vec{1}, \hat{U}\hat{\xi}_j \rangle$ for all $j \in [k]$.
5. Return $\hat{M} := [\hat{\mu}_1 | \hat{\mu}_2 | \cdots | \hat{\mu}_k]$.

Figure 1: Topic-word distribution estimator (Algorithm A).

2.4. Topic-word distribution estimator and convergence guarantee

The spectral properties of the observable operators $B(\vec{\eta})$ implied by Lemma 2 suggest the estimation procedure (Algorithm A) in Figure 1. The procedure is essentially a plug-in approach based on the equations relating the various moments in Lemma 2. We focus on estimating M; the mixing weights $\vec{w}$ can be handled as a secondary step (see Appendix B.5).

On the choice of $\vec{\eta}$. As discussed in the previous section, a suitable choice for $\vec{\eta}$ can be based on prior knowledge about the topic-word distributions, such as $\vec{\eta} = \vec{e}_x$ for some $x \in [d]$ that has different conditional probabilities under each topic. In the absence of such information, one may select $\vec{\eta}$ randomly from the subspace range$(\hat{U})$. Specifically, take $\vec{\eta} := \hat{U}\vec{\theta}$ where $\vec{\theta} \in \mathbb{R}^k$ is a random unit vector distributed uniformly over $\mathcal{S}^{k-1}$.

The following theorem establishes the convergence rate of Algorithm A.

Theorem 3 *There exists a constant $C > 0$ such that the following holds. Pick any $\delta \in (0,1)$. Assume the document topic model from Section 2.1 satisfies Condition 1. Further, assume that in Algorithm A, $\widehat{\text{Pairs}}$ and $\widehat{\text{Triples}}$ are, respectively, the empirical averages of N independent copies of $\vec{x}_1 \otimes \vec{x}_2$ and $\vec{x}_1 \otimes \vec{x}_2 \otimes \vec{x}_3$; and that $\vec{\eta} = \hat{U}\vec{\theta}$ where $\vec{\theta} \in \mathbb{R}^k$ is an independent random unit vector distributed uniformly over $\mathcal{S}^{k-1}$. If*

$$N \geq C \cdot \frac{k^7 \cdot \ln(1/\delta)}{\sigma_k(M)^6 \cdot \sigma_k(\text{Pairs})^4 \cdot \delta^2},$$

then with probability at least $1-\delta$, the parameters returned by Algorithm A have the following guarantee: there exists a permutation τ on $[k]$ and scalars $c_1, c_2, \ldots, c_k \in \mathbb{R}$ such that, for each $j \in [k]$,

$$\|c_j \hat{\mu}_j - \vec{\mu}_{\tau(j)}\|_2 \leq C \cdot \|\vec{\mu}_{\tau(j)}\|_2 \cdot \frac{k^5}{\sigma_k(M)^4 \cdot \sigma_k(\text{Pairs})^2 \cdot \delta} \cdot \sqrt{\frac{\ln(1/\delta)}{N}}.$$

Some illustrative empirical results using Algorithm A are presented in Appendix A.5. A few remarks about the theorem are in order.

On boosting the confidence. Although the convergence depends polynomially on $1/\delta$, where δ is the failure probability, it is possible to boost the confidence by repeating Step 3 of Algorithm A with different random $\vec{\eta}$ until the eigenvalues of $\hat{B}(\vec{\eta})$ are sufficiently separated (as judged by confidence intervals).

On the scaling factors c_j. With a larger sample complexity that depends on d, an error bound can be established for $\|\hat{\mu}_j - \vec{\mu}_{\tau(j)}\|_1$ directly (without the unknown scaling factors c_j), but we do not pursue this as the c_j are removed in Algorithm B anyway.

3. A method of moments for multi-view mixture models

We now consider a much broader class of mixture models and present a general method of moments in this context. Proofs of lemmas and theorems in this section are given in Appendix B.

3.1. General setting

Consider the following multi-view mixture model; k denotes the number of mixture components, and ℓ denotes the number of views. We assume $\ell \geq 3$ throughout. Let $\vec{w} = (w_1, w_2, \ldots, w_k) \in \Delta^{k-1}$ be a vector of mixing weights, and let h be a (hidden) discrete random variable with $\Pr[h = j] = w_j$ for all $j \in [k]$. Let $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \in \mathbb{R}^d$ be ℓ random vectors that are conditionally independent given h; the directed graphical model is depicted in Figure 3(a).

Define the conditional mean vectors as

$$\vec{\mu}_{v,j} := \mathbb{E}[\vec{x}_v | h = j], \quad v \in [\ell], j \in [k],$$

and let $M_v \in \mathbb{R}^{d \times k}$ be the matrix whose j-th column is $\vec{\mu}_{v,j}$. Note that we do not specify anything else about the (conditional) distribution of $\vec{x}_v$—it may be continuous, discrete, or even a hybrid depending on h.

We assume the following conditions on $\vec{w}$ and the M_v.

Condition 2 (Non-degeneracy: general setting) *$w_j > 0$ for all $j \in [k]$, and M_v has rank k for all $v \in [\ell]$.*

We remark that it is easy to generalize to the case where views have different dimensionality (*e.g.*, $\vec{x}_v \in \mathbb{R}^{d_v}$ for possibly different dimensions d_v). For notational simplicity, we stick to the same dimension for each view. Moreover, Condition 2 can be relaxed in some cases; we discuss one such case in Section 4.1 in the context of Gaussian mixture models.

Because the conditional distribution of $\vec{x}_v$ is not specified beyond its conditional means, it is not possible to develop a maximum likelihood approach to parameter estimation. Instead, as in the document topic model, we develop a method of moments based on solving polynomial equations arising from eigenvalue problems.

3.2. Observable moments and operators

We focus on the moments concerning $\{\vec{x}_1, \vec{x}_2, \vec{x}_3\}$, but the same properties hold for other triples of the random vectors $\{\vec{x}_a, \vec{x}_b, \vec{x}_c\} \subseteq \{\vec{x}_v : v \in [\ell]\}$ as well.

As in (1), we define the matrix $P_{1,2} \in \mathbb{R}^{d \times d}$ of second-order moments, and the tensor $P_{1,2,3} \in \mathbb{R}^{d \times d \times d}$ of third-order moments, by

$$P_{1,2} := \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2] \quad \text{and} \quad P_{1,2,3} := \mathbb{E}[\vec{x}_1 \otimes \vec{x}_2 \otimes \vec{x}_3].$$

Again, $P_{1,2,3}$ is regarded as the linear operator $P_{1,2,3} \colon \vec{\eta} \mapsto \mathbb{E}[(\vec{x}_1 \otimes \vec{x}_2)\langle \vec{\eta}, \vec{x}_3 \rangle]$.

Lemma 4 and Lemma 5 are straightforward generalizations of Lemma 1 and Lemma 2.

Lemma 4 *$P_{1,2} = M_1 \operatorname{diag}(\vec{w}) M_2^\top$ and $P_{1,2,3}(\vec{\eta}) = M_1 \operatorname{diag}(M_3^\top \vec{\eta}) \operatorname{diag}(\vec{w}) M_2^\top$ for all $\vec{\eta} \in \mathbb{R}^d$.*

Lemma 5 *Assume Condition 2. For $v \in \{1, 2, 3\}$, let $U_v \in \mathbb{R}^{d \times k}$ be a matrix such that $U_v^\top M_v$ is invertible. Then $U_1^\top P_{1,2} U_2$ is invertible, and for all $\vec{\eta} \in \mathbb{R}^d$, the "observable operator" $B_{1,2,3}(\vec{\eta}) \in \mathbb{R}^{k \times k}$, given by $B_{1,2,3}(\vec{\eta}) := (U_1^\top P_{1,2,3}(\vec{\eta}) U_2)(U_1^\top P_{1,2} U_2)^{-1}$, satisfies*

$$B_{1,2,3}(\vec{\eta}) = (U_1^\top M_1) \operatorname{diag}(M_3^\top \vec{\eta})(U_1^\top M_1)^{-1}.$$

In particular, the k roots of the polynomial $\lambda \mapsto \det(B_{1,2,3}(\vec{\eta}) - \lambda I)$ are $\{\langle \vec{\eta}, \vec{\mu}_{3,j} \rangle : j \in [k]\}$.

Recall that Algorithm A relates the eigenvectors of $B(\vec{\eta})$ to the matrix of conditional means M. The eigenvectors are only defined up to a scaling of each vector, so without prior knowledge of the correct scaling, they are not sufficient to recover the parameters M. Nevertheless, the eigenvalues also carry information about the parameters, as shown in Lemma 5, and it is possible to reconstruct the parameters from different the observation operators applied to different vectors $\vec{\eta}$. This idea is captured in the following lemma.

Lemma 6 *Consider the setting and definitions from Lemma 5. Let $\Theta \in \mathbb{R}^{k \times k}$ be an invertible matrix, and let $\vec{\theta}_i^\top \in \mathbb{R}^k$ be its i-th row. Moreover, for all $i \in [k]$, let $\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,k}$ denote the k eigenvalues of $B_{1,2,3}(U_3 \vec{\theta}_i)$ in the order specified by the matrix of right eigenvectors $U_1^\top M_1$. Let $L \in \mathbb{R}^{k \times k}$ be the matrix whose (i, j)-th entry is $\lambda_{i,j}$. Then*

$$\Theta U_3^\top M_3 = L.$$

Observe that the unknown parameters M_3 are expressed as the solution to a linear system in the above equation, where the elements of the right-hand side L are the roots of k-th degree polynomials derived from the second- and third-order observable moments (namely, the characteristic polynomials of the $B_{1,2,3}(U_3 \vec{\theta}_i)$, $\forall i \in [k]$). This template is also found in other moment methods based on decompositions of a Hankel matrix. A crucial distinction, however, is that the k-th degree polynomials in Lemma 6 only involve low-order moments, whereas standard methods may involve up to $\Omega(k)$-th order moments which are difficult to estimate (Lindsay, 1989; Lindsay and Basak, 1993; Gravin et al., 2012).

Algorithm B

1. Compute empirical averages from N independent copies of $\vec{x}_1 \otimes \vec{x}_2$ to form $\hat{P}_{1,2} \in \mathbb{R}^{d \times d}$. Similarly do the same for $\vec{x}_1 \otimes \vec{x}_3$ to form $\hat{P}_{1,3} \in \mathbb{R}^{k \times k}$, and for $\vec{x}_1 \otimes \vec{x}_2 \otimes \vec{x}_3$ to form $\hat{P}_{1,2,3} \in \mathbb{R}^{d \times d \times d}$.
2. Let $\hat{U}_1 \in \mathbb{R}^{d \times k}$ and $\hat{U}_2 \in \mathbb{R}^{d \times k}$ be, respectively, matrices of orthonormal left and right singular vectors of $\hat{P}_{1,2}$ corresponding to its top k singular values. Let $\hat{U}_3 \in \mathbb{R}^{d \times k}$ be the matrix of orthonormal right singular vectors of $\hat{P}_{1,3}$ corresponding to its top k singular values.
3. Pick an invertible matrix $\Theta \in \mathbb{R}^{k \times k}$, with its i-th row denoted as $\vec{\theta}_i^\top \in \mathbb{R}^k$. In the absence of any prior information about M_3, a suitable choice for Θ is a random rotation matrix. Form the matrix $\hat{B}_{1,2,3}(\hat{U}_3\vec{\theta}_1) := (\hat{U}_1^\top \hat{P}_{1,2,3}(\hat{U}_3\vec{\theta}_1)\hat{U}_2)(\hat{U}_1^\top \hat{P}_{1,2}\hat{U}_2)^{-1}$.
   Compute $\hat{R}_1 \in \mathbb{R}^{k \times k}$ (with unit Euclidean norm columns) that diagonalizes $\hat{B}_{1,2,3}(\hat{U}_3\vec{\theta}_1)$, *i.e.*, $\hat{R}_1^{-1}\hat{B}_{1,2,3}(\hat{U}_3\vec{\theta}_1)\hat{R}_1 = \operatorname{diag}(\hat{\lambda}_{1,1}, \hat{\lambda}_{1,2}, \dots, \hat{\lambda}_{1,k})$. (Fail if not possible.)
4. For each $i \in \{2, \dots, k\}$, obtain the diagonal entries $\hat{\lambda}_{i,1}, \hat{\lambda}_{i,2}, \dots, \hat{\lambda}_{i,k}$ of $\hat{R}_1^{-1}\hat{B}_{1,2,3}(\hat{U}_3\vec{\theta}_i)\hat{R}_1$, and form the matrix $\hat{L} \in \mathbb{R}^{k \times k}$ whose (i,j)-th entry is $\hat{\lambda}_{i,j}$.
5. Return $\hat{M}_3 := \hat{U}_3\Theta^{-1}\hat{L}$.

Figure 2: General method of moments estimator (Algorithm B).

3.3. Main result: general estimation procedure and sample complexity bound

The lemmas in the previous section suggest the estimation procedure (Algorithm B) presented in Figure 2.

As stated, the Algorithm B yields an estimator for M_3, but the method can easily be applied to estimate M_v for all other views v. One caveat is that the estimators may not yield the same ordering of the columns, due to the unspecified order of the eigenvectors obtained in the third step of the method, and therefore some care is needed to obtain a consistent ordering. We outline one solution in Appendix B.4.

The sample complexity of Algorithm B depends on the specific concentration properties of $\vec{x}_1, \vec{x}_2, \vec{x}_3$. We abstract away this dependence in the following condition.

Condition 3 *There exist positive scalars N_0, $C_{1,2}$, $C_{1,3}$, $C_{1,2,3}$, and a function $f(N,\delta)$ (decreasing in N and δ) such that for any $N \geq N_0$ and $\delta \in (0,1)$,*

1. $\Pr\Big[\|\hat{P}_{a,b} - P_{a,b}\|_2 \leq C_{a,b} \cdot f(N,\delta)\Big] \geq 1-\delta$ *for* $\{a,b\} \in \{\{1,2\},\{1,3\}\}$,
2. $\forall \vec{v} \in \mathbb{R}^d$, $\Pr\Big[\|\hat{P}_{1,2,3}(\vec{v}) - P_{1,2,3}(\vec{v})\|_2 \leq C_{1,2,3} \cdot \|\vec{v}\|_2 \cdot f(N,\delta)\Big] \geq 1-\delta$.

Moreover (for technical convenience), $\hat{P}_{1,3}$ is independent of $\hat{P}_{1,2,3}$ (which may be achieved, say, by splitting a sample of size $2N$).

For the discrete models such as the document topic model of Section 2.1 and discrete HMMs (Mossel and Roch, 2006; Hsu et al., 2009), Condition 3 holds with $N_0 = C_{1,2} = C_{1,3} = C_{1,2,3} = 1$, and $f(N,\delta) = (1+\sqrt{\ln(1/\delta)})/\sqrt{N}$. Using standard techniques (*e.g.*, Chaudhuri et al. (2009); Vershynin (2012)), the condition can also be shown to hold for mixtures of various continuous distributions such as multivariate Gaussians.

Now we are ready to present the main theorem of this section.

Theorem 7 *There exists a constant $C > 0$ such that the following holds. Assume the three-view mixture model satisfies Condition 2 and Condition 3. Pick any $\epsilon \in (0,1)$ and $\delta \in (0, \delta_0)$. Further, assume $\Theta \in \mathbb{R}^{k \times k}$ is an independent random rotation matrix distributed uniformly over the Stiefel manifold $\{Q \in \mathbb{R}^{k \times k} : Q^\top Q = I\}$. If the number of samples N satisfies $N \geq N_0$ and*

$$f(N, \delta/k) \leq C \cdot \frac{\min_{i \neq j} \|M_3(\vec{e}_i - \vec{e}_j)\|_2 \cdot \sigma_k(P_{1,2})}{C_{1,2,3} \cdot k^5 \cdot \kappa(M_1)^4} \cdot \frac{\delta}{\ln(k/\delta)} \cdot \epsilon,$$

$$f(N, \delta) \leq C \cdot \min \left\{ \frac{\min_{i \neq j} \|M_3(\vec{e}_i - \vec{e}_j)\|_2 \cdot \sigma_k(P_{1,2})^2}{C_{1,2} \cdot \|P_{1,2,3}\|_2 \cdot k^5 \cdot \kappa(M_1)^4} \cdot \frac{\delta}{\ln(k/\delta)}, \ \frac{\sigma_k(P_{1,3})}{C_{1,3}} \right\} \cdot \epsilon$$

where $\|P_{1,2,3}\|_2 := \max_{\vec{v} \neq \vec{0}} \|P_{1,2,3}(\vec{v})\|_2$, then with probability at least $1 - 5\delta$, Algorithm B returns $\hat{M}_3 = [\hat{\mu}_{3,1} | \hat{\mu}_{3,2} | \cdots | \hat{\mu}_{3,k}]$ with the following guarantee: there exists a permutation τ on $[k]$ such that for each $j \in [k]$,

$$\|\hat{\mu}_{3,j} - \vec{\mu}_{3,\tau(j)}\|_2 \leq \max_{j' \in [k]} \|\vec{\mu}_{3,j'}\|_2 \cdot \epsilon.$$

4. Applications

In addition to the document clustering model from Section 2, a number of natural latent variable models fit into this multi-view framework. We describe two such cases in this section: Gaussian mixture models and HMMs, both of which have been (at least partially) studied in the literature. In both cases, the estimation technique of Algorithm B leads to new learnability results that were not achieved by previous works.

4.1. Multi-view and axis-aligned Gaussian mixture models

The standard Gaussian mixture model is parameterized by a mixing weight w_j, mean vector $\vec{\mu}_j \in \mathbb{R}^D$, and covariance matrix $\Sigma_j \in \mathbb{R}^{D \times D}$ for each mixture component $j \in [k]$. The hidden discrete random variable h selects a component j with probability $\Pr[h = j] = w_j$; the conditional distribution of the observed random vector $\vec{x}$ given h is a multivariate Gaussian with mean $\vec{\mu}_h$ and covariance Σ_h.

The multi-view assumption for Gaussian mixture models asserts that for each component j, the covariance Σ_j has a block diagonal structure $\Sigma_j = \operatorname{blkdiag}(\Sigma_{1,j}, \Sigma_{2,j}, \ldots, \Sigma_{\ell,j})$ (a special case is an axis-aligned Gaussian). The various blocks correspond to the ℓ different views of the data $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_\ell \in \mathbb{R}^d$ (for $d = D/\ell$), which are conditionally independent given h. The mean vector for each component j is similarly partitioned into the views as $\vec{\mu}_j = (\vec{\mu}_{1,j}, \vec{\mu}_{2,j}, \ldots, \vec{\mu}_{\ell,j})$. In the case of an axis-aligned Gaussian, each covariance matrix Σ_j is diagonal, and therefore the original coordinates $[D]$ can be partitioned into $\ell = O(D/k)$ views (each of dimension $d = \Omega(k)$) in any way (say, randomly).[2]

2. For product distributions (*e.g.*, axis-aligned Gaussians) satisfying a certain incoherence condition, Condition 2 can be established using a random partitioning of the coordinates; see Appendix D.1 for details.

Condition 2 requires that the conditional mean matrix $M_v = [\vec{\mu}_{v,1}|\vec{\mu}_{v,2}|\cdots|\vec{\mu}_{v,k}]$ for each view v have full column rank (see Appendix D.2 for a possible relaxation). This is similar to the non-degeneracy and spreading conditions used in previous studies of multi-view clustering (Chaudhuri and Rao, 2008; Chaudhuri et al., 2009). In these previous works, the multi-view and non-degeneracy assumptions are shown to reduce the minimum separation required for various efficient algorithms to learn the model parameters. In comparison, Algorithm B does not require a minimum separation condition at all.

Condition 3 can be established for this class of mixture models (in fact, even when the component distributions are simply subgaussian; see Appendix D.3 for details). Therefore, Algorithm B can be used to recover the means of each component distribution (and the covariances can be recovered as well; see Appendix D.4).

4.2. Hidden Markov models

A hidden Markov model is a latent variable model in which a hidden state sequence $h_1, h_2, \ldots, h_\ell$ forms a Markov chain $h_1 \to h_2 \to \cdots \to h_\ell$ over k possible states $[k]$; and given the state h_t at time $t \in [k]$, the observation $\vec{x}_t$ at time t (a random vector taking values in $\mathbb{R}^d$) is conditionally independent of all other observations and states. The directed graphical model is depicted in Figure 3(b).

The vector $\vec{\pi} \in \Delta^{k-1}$ is the initial state distribution: $\Pr[h_1 = i] = \pi_i$ for all $i \in [k]$. For simplicity, we only consider time-homogeneous HMMs, although it is possible to generalize to the time-varying setting. The matrix $T \in \mathbb{R}^{k\times k}$ is a stochastic matrix describing the hidden state Markov chain: $\Pr[h_{t+1} = i|h_t = j] = T_{i,j}$ for all $i, j \in [k], t \in [\ell - 1]$. Finally, the columns of the matrix $O = [\vec{o}_1|\vec{o}_2|\cdots|\vec{o}_k] \in \mathbb{R}^{d\times k}$ are the conditional means of the observation $\vec{x}_t$ at time t given the corresponding hidden state h_t: $\mathbb{E}[\vec{x}_t|h_t = i] = O\vec{e}_i = \vec{o}_i$ for all $i \in [k], t \in [\ell]$. Note that both discrete and continuous observations are readily handled in this framework. For instance, the conditional distribution of $\vec{x}_t$ given $h_t = i$ (for $i \in [k]$) could be a high-dimensional multivariate Gaussian with mean $\vec{o}_i \in \mathbb{R}^d$. Such models were not handled by previous methods (Mossel and Roch, 2006; Hsu et al., 2009).

The restriction of the HMM to three time steps, say $t \in \{1, 2, 3\}$, is an instance of the three-view mixture model.

Proposition 8 *If the hidden variable h (from the three-view mixture model of Section 3.1) is identified with the second hidden state h_2, then $\{\vec{x}_1, \vec{x}_2, \vec{x}_3\}$ are conditionally independent given h, and the parameters of the resulting three-view mixture model on $(h, \vec{x}_1, \vec{x}_2, \vec{x}_3)$ are $\vec{w} := T\vec{\pi}$, $M_1 := O \operatorname{diag}(\vec{\pi})T^\top \operatorname{diag}(T\vec{\pi})^{-1}$, $M_2 := O$, and $M_3 := OT$.*

From Proposition 8, it is easy to verify that $B_{3,1,2}(\vec{\eta}) = (U_3^\top OT) \operatorname{diag}(O^\top\vec{\eta})(U_3^\top OT)^{-1}$. Therefore, after recovering the observation conditional mean matrix O using Algorithm B, the Markov chain transition matrix can be recovered using the matrix of right eigenvectors R of $B_{3,1,2}(\vec{\eta})$ and the equation $(U_3^\top O)^{-1}R = T$ (up to scaling of the columns).

Acknowledgments

We thank Kamalika Chaudhuri and Tong Zhang for many useful discussions, Karl Stratos for comments on an early draft, and David Sontag and an anonymous reviewer for some pointers to related work.

References

D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *COLT*, 2005.

R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

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

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

M. B. Blaschko and C. H. Lampert. Correlational spectral clustering. In *CVPR*, 2008.

D. L. Boley, F. T. Luk, and D. Vandevoorde. Vandermonde factorization of a Hankel matrix. In *Scientific Computing*, 1997.

S. C. Brubaker and S. Vempala. Isotropic PCA and affine-invariant clustering. In *FOCS*, 2008.

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.

K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, 2009.

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.

J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *FOCS*, 2005.

J. Feldman, R. O'Donnell, and R. Servedio. PAC learning mixtures of axis-aligned Gaussians with no separation assumption. In *COLT*, 2006.

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

W. A. Gale, K. W. Church, and D. Yarowsky. One sense per discourse. In *4th DARPA Speech and Natural Language Workshop*, 1992.

N. Gravin, J. Lasserre, D. Pasechnik, and S. Robins. The inverse moment problem for convex polytopes. *Discrete and Computational Geometry*, 2012. To appear.

H. Hotelling. The most predictable criterion. *Journal of Educational Psychology*, 26(2): 139–142, 1935.

D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. In *COLT*, 2009.

D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 2012. To appear.

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

H. Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6), 2000.

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

R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *COLT*, 2005.

B. G. Lindsay. Moment matrices: applications in mixtures. *Annals of Statistics*, 17(2): 722–740, 1989.

B. G. Lindsay. *Mixture models: theory, geometry and applications.* American Statistical Association, 1995.

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

F. McSherry. Spectral partitioning of random graphs. In *FOCS*, 2001.

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.

B. Recht. A simpler approach to matrix completion, 2009. arXiv:0910.0651v2.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

M. P. Schützenberger. On the definition of a family of automata. *Information and Control*, 4:245–270, 1961.

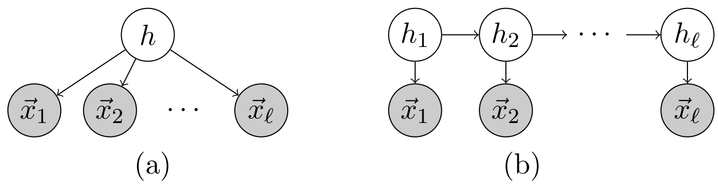

Figure 3: (a) The multi-view mixture model. (b) A hidden Markov model.

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.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

Appendix A. Proofs and details from Section 2

In this section, we provide omitted proofs and discussion from Section 2 (deferring most perturbation arguments to Appendix C), and also present some illustrative empirical results on text data using a modified version of Algorithm A.

A.1. Proof of Lemma 1

Since $\vec{x}_1$, $\vec{x}_2$, and $\vec{x}_3$ are conditionally independent given h,

$$\begin{aligned}\mathrm{Pairs}_{i,j} = \Pr[\vec{x}_1 = \vec{e}_i, \vec{x}_2 = \vec{e}_j] &= \sum_{t=1}^{k} \Pr[\vec{x}_1 = \vec{e}_i, \vec{x}_2 = \vec{e}_j | h = t] \cdot \Pr[h = t] \\ &= \sum_{t=1}^{k} \Pr[\vec{x}_1 = \vec{e}_i | h = t] \cdot \Pr[\vec{x}_2 = \vec{e}_j | h = t] \cdot \Pr[h = t] = \sum_{t=1}^{k} M_{i,t} \cdot M_{j,t} \cdot w_t\end{aligned}$$

so $\mathrm{Pairs} = M \operatorname{diag}(\vec{w}) M^\top$. Moreover, writing $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_d)$,

$$\begin{aligned}\mathrm{Triples}(\vec{\eta})_{i,j} &= \sum_{x=1}^{d} \eta_x \Pr[\vec{x}_1 = \vec{e}_i, \vec{x}_2 = \vec{e}_j, \vec{x}_3 = \vec{e}_x] \\ &= \sum_{x=1}^{d} \sum_{t=1}^{k} \eta_x \cdot M_{i,t} \cdot M_{j,t} \cdot M_{x,t} \cdot w_t = \sum_{t=1}^{k} M_{i,t} \cdot M_{j,t} \cdot w_t \cdot (M^\top \vec{\eta})_t\end{aligned}$$

so $\mathrm{Triples}(\vec{\eta}) = M \operatorname{diag}(M^\top \vec{\eta}) \operatorname{diag}(\vec{w}) M^\top$.

A.2. Proof of Lemma 2

Since $\operatorname{diag}(\vec{w}) \succ 0$ by Condition 1 and $U^\top \mathrm{Pairs} V = (U^\top M) \operatorname{diag}(\vec{w}) M^\top V$ by Lemma 1, it follows that $U^\top \mathrm{Pairs} V$ is invertible by the assumptions on U and V. Moreover, also by Lemma 1,

$$\begin{aligned} B(\vec{\eta}) &= (U^\top \mathrm{Triples}(\vec{\eta}) V)\ (U^\top \mathrm{Pairs} V)^{-1} \\ &= (U^\top M \operatorname{diag}(M^\top \vec{\eta}) \operatorname{diag}(\vec{w}) M^\top V)\ (U^\top \mathrm{Pairs} V)^{-1} \\ &= (U^\top M) \operatorname{diag}(M^\top \vec{\eta}) (U^\top M)^{-1}\ (U^\top M \operatorname{diag}(\vec{w}) M^\top V)\ (U^\top \mathrm{Pairs} V)^{-1} \\ &= (U^\top M) \operatorname{diag}(M^\top \vec{\eta}) (U^\top M)^{-1}. \end{aligned}$$

A.3. Accuracy of moment estimates

Lemma 9 *Fix $\delta \in (0,1)$. Let $\widehat{\mathrm{Pairs}}$ be the empirical average of N independent copies of $\vec{x}_1 \otimes \vec{x}_2$, and let $\widehat{\mathrm{Triples}}$ be the empirical average of N independent copies of $(\vec{x}_1 \otimes \vec{x}_2)\langle \vec{\eta}, \vec{x}_3 \rangle$. Then*

1. $\Pr\left[\|\widehat{\mathrm{Pairs}} - \mathrm{Pairs}\|_{\mathrm{F}} \leq \frac{1 + \sqrt{\ln(1/\delta)}}{\sqrt{N}} \right] \geq 1 - \delta$, *and*

2. $\Pr\left[\forall \vec{\eta} \in \mathbb{R}^d,\ \|\widehat{\mathrm{Triples}}(\vec{\eta}) - \mathrm{Triples}(\vec{\eta})\|_{\mathrm{F}} \leq \frac{\|\vec{\eta}\|_2 (1 + \sqrt{\ln(1/\delta)})}{\sqrt{N}} \right] \geq 1 - \delta.$

Proof The first claim follows from applying Lemma 24 to the vectorizations of $\widehat{\mathrm{Pairs}}$ and Pairs (whereupon the Frobenius norm is the Euclidean norm of the vectorized matrices). For the second claim, we also apply Lemma 24 to $\widehat{\mathrm{Triples}}$ and Triples in the same way to obtain, with probability at least $1 - \delta$,

$$\sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{x=1}^{d} (\widehat{\mathrm{Triples}}_{i,j,x} - \mathrm{Triples}_{i,j,x})^2 \leq \frac{(1 + \sqrt{\ln(1/\delta)})^2}{N}.$$

Now condition on this event. For any $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_d) \in \mathbb{R}^d$,

$$\begin{aligned} \|\widehat{\mathrm{Triples}}(\vec{\eta}) - \mathrm{Triples}(\vec{\eta})\|_{\mathrm{F}}^2 &= \sum_{i=1}^{d} \sum_{j=1}^{d} \left| \sum_{x=1}^{d} \eta_x (\widehat{\mathrm{Triples}}_{i,j,x} - \mathrm{Triples}_{i,j,x}) \right|^2 \\ &\leq \sum_{i=1}^{d} \sum_{j=1}^{d} \|\vec{\eta}\|_2^2 \sum_{x=1}^{d} (\widehat{\mathrm{Triples}}_{i,j,x} - \mathrm{Triples}_{i,j,x})^2 \\ &\leq \frac{\|\vec{\eta}\|_2^2 (1 + \sqrt{\ln(1/\delta)})^2}{N} \end{aligned}$$

where the first inequality follows by Cauchy-Schwarz. ∎

A.4. Proof of Theorem 3

Let E_1 be the event in which

$$\|\widehat{\text{Pairs}} - \text{Pairs}\|_2 \le \frac{1 + \sqrt{\ln(1/\delta)}}{\sqrt{N}} \tag{2}$$

and

$$\|\widehat{\text{Triples}}(\vec{v}) - \text{Triples}(\vec{v})\|_2 \le \frac{\|v\|_2(1 + \sqrt{\ln(1/\delta)})}{\sqrt{N}} \tag{3}$$

for all $\vec{v} \in \mathbb{R}^d$. By Lemma 9, a union bound, and the fact that $\|A\|_2 \le \|A\|_{\mathrm{F}}$, we have $\Pr[E_1] \ge 1 - 2\delta$. Now condition on E_1, and let E_2 be the event in which

$$\gamma := \min_{i \neq j} |\langle \hat{U}\vec{\theta}, M(\vec{e}_i - \vec{e}_j)\rangle| = \min_{i \neq j} |\langle \vec{\theta}, \hat{U}^\top M(\vec{e}_i - \vec{e}_j)\rangle| > \frac{\sqrt{2}\sigma_k(\hat{U}^\top M) \cdot \delta}{\sqrt{e}k\binom{k}{2}}. \tag{4}$$

By Lemma 15 and the fact $\|\hat{U}^\top M(\vec{e}_i - \vec{e}_j)\|_2 \ge \sqrt{2}\sigma_k(\hat{U}^\top M)$, we have $\Pr[E_2|E_1] \ge 1 - \delta$, and thus $\Pr[E_1 \cap E_2] \ge (1-2\delta)(1-\delta) \ge 1 - 3\delta$. So henceforth condition on this joint event $E_1 \cap E_2$.

Let $\varepsilon_0 := \frac{\|\widehat{\text{Pairs}} - \text{Pairs}\|_2}{\sigma_k(\text{Pairs})}$, $\varepsilon_1 := \frac{\varepsilon_0}{1-\varepsilon_0}$, and $\varepsilon_2 := \frac{\varepsilon_0}{(1-\varepsilon_1^2)\cdot(1-\varepsilon_0-\varepsilon_1^2)}$. The conditions on N and the bound in (2) implies that $\varepsilon_0 < \frac{1}{1+\sqrt{2}} \le \frac{1}{2}$, so Lemma 10 implies that $\sigma_k(\hat{U}^\top M) \ge \sqrt{1-\varepsilon_1^2} \cdot \sigma_k(M)$, $\kappa(\hat{U}^\top M) \le \frac{\|M\|_2}{\sqrt{1-\varepsilon_1^2 \cdot \sigma_k(M)}}$, and that $\hat{U}^\top \text{Pairs} \hat{V}$ is invertible. By Lemma 2,

$$\tilde{B}(\vec{\eta}) := (\hat{U}^\top \text{Triples}(\vec{\eta})\hat{V})(\hat{U}^\top \text{Pairs}\hat{V})^{-1} = (\hat{U}^\top M)\operatorname{diag}(M^\top \vec{\eta})(\hat{U}^\top M)^{-1}.$$

Thus, Lemma 11 implies

$$\|\hat{B}(\vec{\eta}) - \tilde{B}(\vec{\eta})\|_2 \le \frac{\|\widehat{\text{Triples}}(\vec{\eta}) - \text{Triples}(\vec{\eta})\|_2}{(1-\varepsilon_0)\cdot\sigma_k(\text{Pairs})} + \frac{\varepsilon_2}{\sigma_k(\text{Pairs})}. \tag{5}$$

Let $R := \hat{U}^\top M \operatorname{diag}(\|\hat{U}^\top M\vec{e}_1\|_2, \|\hat{U}^\top M\vec{e}_2\|_2, \ldots, \|\hat{U}^\top M\vec{e}_k\|_2)^{-1}$ and $\varepsilon_3 := \frac{\|\hat{B}(\vec{\eta}) - \tilde{B}(\vec{\eta})\|_2 \cdot \kappa(R)}{\gamma}$. Note that R has unit norm columns, and that $R^{-1}\tilde{B}(\vec{\eta})R = \operatorname{diag}(M^\top\vec{\eta})$. By Lemma 14 and the fact $\|M\|_2 \le \sqrt{k}\|M\|_1 = \sqrt{k}$,

$$\|R^{-1}\|_2 \le \kappa(\hat{U}^\top M) \le \frac{\|M\|_2}{\sqrt{1-\varepsilon_1^2}\cdot\sigma_k(M)} \le \frac{\sqrt{k}}{\sqrt{1-\varepsilon_1^2}\cdot\sigma_k(M)} \tag{6}$$

and

$$\kappa(R) \le \kappa(\hat{U}^\top M)^2 \le \frac{k}{(1-\varepsilon_1^2)\cdot\sigma_k(M)^2}. \tag{7}$$

The conditions on N and the bounds in (2), (3), (4), (5), and (7) imply that $\varepsilon_3 < \frac{1}{2}$. By Lemma 12, there exists a permutation τ on $[k]$ such that, for all $j \in [k]$,

$$\|s_j\hat{\xi}_j - \hat{U}^\top\vec{\mu}_{\tau(j)}/c'_j\|_2 = \|s_j\hat{\xi}_j - R\vec{e}_{\tau(j)}\|_2 \le 4k \cdot \|R^{-1}\|_2 \cdot \varepsilon_3 \tag{8}$$

where $s_j := \mathrm{sign}(\langle \hat{\xi}_j, \hat{U}^\top \vec{\mu}_{\tau(j)} \rangle)$ and $c'_j := \|\hat{U}^\top \vec{\mu}_{\tau(j)}\|_2 \leq \|\vec{\mu}_{\tau(j)}\|_2$ (the eigenvectors $\hat{\xi}_j$ are unique up to sign s_j because each eigenvalue has geometric multiplicity 1). Since $\vec{\mu}_{\tau(j)} \in \mathrm{range}(U)$, Lemma 10 and the bounds in (8) and (6) imply

$$\begin{aligned}\|s_j \hat{U} \hat{\xi}_j - \vec{\mu}_{\tau(j)}/c'_j\|_2 &\leq \sqrt{\|s_j \hat{\xi}_j - \hat{U}^\top \vec{\mu}_{\tau(j)}/c'_j\|_2^2 + \|\vec{\mu}_{\tau(j)}/c'_j\|_2^2 \cdot \varepsilon_1^2} \\ &\leq \|s_j \hat{\xi}_j - \hat{U}^\top \vec{\mu}_{\tau(j)}/c'_j\|_2 + \|\vec{\mu}_{\tau(j)}/c'_j\|_2 \cdot \varepsilon_1 \\ &\leq 4k \cdot \|R^{-1}\|_2 \cdot \varepsilon_3 + \varepsilon_1 \\ &\leq 4k \cdot \frac{\sqrt{k}}{\sqrt{1-\varepsilon_1^2} \cdot \sigma_k(M)} \cdot \varepsilon_3 + \varepsilon_1.\end{aligned}$$

Therefore, for $c_j := s_j c'_j \langle \vec{1}, \hat{U}\hat{\xi}_j \rangle$, we have

$$\|c_j \hat{\mu}_j - \vec{\mu}_{\tau(j)}\|_2 = \|c'_j s_j \hat{U} \hat{\xi}_j - \vec{\mu}_{\tau(j)}\|_2 \leq \|\vec{\mu}_{\tau(j)}\|_2 \cdot \left(4k \cdot \frac{\sqrt{k}}{\sqrt{1-\varepsilon_1^2} \cdot \sigma_k(M)} \cdot \varepsilon_3 + \varepsilon_1\right).$$

Making all of the substitutions into the above bound gives

$$\begin{aligned}\frac{\|c_j \hat{\mu}_j - \vec{\mu}_{\tau(j)}\|_2}{\|\vec{\mu}_{\tau(j)}\|_2} &\leq \frac{4k^{1.5}}{\sqrt{1-\varepsilon_1^2} \cdot \sigma_k(M)} \cdot \frac{k}{(1-\varepsilon_1^2) \cdot \sigma_k(M)^2} \cdot \frac{\sqrt{ek} \cdot \binom{k}{2}}{\sqrt{2(1-\varepsilon_1^2)} \cdot \sigma_k(M) \cdot \delta} \\ &\quad \cdot \left(\frac{\|\widehat{\mathrm{Triples}}(\vec{\eta}) - \mathrm{Triples}(\vec{\eta})\|_2}{(1-\varepsilon_0) \cdot \sigma_k(\mathrm{Pairs})} + \frac{\|\widehat{\mathrm{Pairs}} - \mathrm{Pairs}\|_2}{(1-\varepsilon_1^2) \cdot (1-\varepsilon_0-\varepsilon_1^2) \cdot \sigma_k(\mathrm{Pairs})^2} \right) \\ &\quad + \frac{\|\widehat{\mathrm{Pairs}} - \mathrm{Pairs}\|_2}{(1-\varepsilon_0) \cdot \sigma_k(\mathrm{Pairs})} \\ &\leq C \cdot \frac{k^5}{\sigma_k(M)^4 \cdot \sigma_k(\mathrm{Pairs})^2 \cdot \delta} \cdot \sqrt{\frac{\ln(1/\delta)}{N}}.\end{aligned}$$

A.5. Some illustrative empirical results

As a demonstration of feasibility, we applied a modified version of Algorithm A to a subset of articles from the "20 Newsgroups" dataset, specifically those in `comp.graphics`, `rec.sport.baseball`, `sci.crypt`, and `soc.religion.christian`, where $\vec{x}_1, \vec{x}_2, \vec{x}_3$ represent three words from the beginning (first third), middle (middle third), and end (last third) of an article. We used $k = 25$ (although results were similar for $k \in \{10, 15, 20, 25, 30\}$) and $d = 5441$ (after removing a standard set of 524 stop-words and applying Porter stemming). Instead of using a single $\vec{\eta}$ and extracting all eigenvectors of $\hat{B}(\vec{\eta})$, we extracted a single eigenvector $\vec{\xi}_x$ from $\hat{B}(\vec{e}_x)$ for several words $x \in [d]$ (these x's were chosen using an automatic heuristic based on their statistical leverage scores in $\widehat{\mathrm{Pairs}}$). Below, for each such $(\hat{B}(\vec{e}_x), \vec{\xi}_x)$, we report the top 15 words y ordered by $\vec{e}_y^\top \hat{U} \vec{\xi}_x$ value.

$\hat{B}(\vec{e}_{\text{format}})$	$\hat{B}(\vec{e}_{\text{god}})$	$\hat{B}(\vec{e}_{\text{key}})$	$\hat{B}(\vec{e}_{\text{polygon}})$	$\hat{B}(\vec{e}_{\text{team}})$	$\hat{B}(\vec{e}_{\text{today}})$
source	god	key	polygon	win	game
find	write	bit	time	game	tiger
post	jesus	chip	save	run	bit
image	christian	system	refer	team	run
feal	christ	encrypt	book	year	pitch
intersect	people	car	source	don	day
email	time	repository	man	watch	team
rpi	apr	ve	routine	good	true
time	sin	public	netcom	score	lot
problem	bible	escrow	gif	yankees	book
file	day	secure	record	pitch	lost
program	church	make	subscribe	start	colorado
gif	person	clipper	change	bit	fan
bit	book	write	algorithm	time	apr
jpeg	life	nsa	scott	wonder	watch

The first and fourth topics appear to be about computer graphics (`comp.graphics`), the fifth and sixth about baseball (`rec.sports.baseball`), the third about encryption (`sci.crypt`), and the second about Christianity (`soc.religion.christian`).

We also remark that Algorithm A can be implemented so that it makes just two passes over the training data, and that simple hashing or random projection tricks can reduce the memory requirement to $O(k^2 + kd)$ (*i.e.*, $\widehat{\text{Pairs}}$ and $\widehat{\text{Triples}}$ never need to be explicitly formed).

Appendix B. Proofs and details from Section 3

In this section, we provide omitted proofs and discussion from Section 3.

B.1. Proof of Lemma 4

By conditional independence,

$$\begin{aligned} P_{1,2} &= \mathbb{E}[\mathbb{E}[\vec{x}_1 \otimes \vec{x}_2 | h]] = \mathbb{E}[\mathbb{E}[\vec{x}_1 | h] \otimes \mathbb{E}[\vec{x}_2 | h]] \\ &= \mathbb{E}[(M_1 \vec{e}_h) \otimes (M_2 \vec{e}_h)] = M_1 \left(\sum_{t=1}^{k} w_t \vec{e}_t \otimes \vec{e}_t \right) M_2^\top = M_1 \operatorname{diag}(\vec{w}) M_2^\top . \end{aligned}$$

Similarly,

$$\begin{aligned} P_{1,2,3}(\vec{\eta}) &= \mathbb{E}[\mathbb{E}[(\vec{x}_1 \otimes \vec{x}_2) \langle \vec{\eta}, \vec{x}_3 \rangle | h]] = \mathbb{E}[\mathbb{E}[\vec{x}_1 | h] \otimes \mathbb{E}[\vec{x}_2 | h] \langle \vec{\eta}, \mathbb{E}[\vec{x}_3 | h] \rangle] \\ &= \mathbb{E}[(M_1 \vec{e}_h) \otimes (M_2 \vec{e}_h) \langle \vec{\eta}, M_3 \vec{e}_h \rangle] = M_1 \left(\sum_{t=1}^{k} w_t \vec{e}_h \otimes \vec{e}_h \langle \vec{\eta}, M_3 \vec{e}_h \rangle \right) M_2^\top \\ &= M_1 \operatorname{diag}(M_3^\top \vec{\eta}) \operatorname{diag}(\vec{w}) M_2^\top . \end{aligned}$$

B.2. Proof of Lemma 5

We have $U_1^\top P_{1,2} U_2 = (U_1^\top M_1) \operatorname{diag}(\vec{w})(M_2^\top U_2)$ by Lemma 4, which is invertible by the assumptions on U_v and Condition 2. Moreover, also by Lemma 4,

$$\begin{aligned}
B_{1,2,3}(\vec{\eta}) &= (U_1^\top P_{1,2,3}(\vec{\eta}) U_2)\ (U_1^\top P_{1,2} U_2)^{-1} \\
&= (U_1^\top M_1 \operatorname{diag}(M_3^\top \vec{\eta}) \operatorname{diag}(\vec{w}) M_2^\top U_2)\ (U_1^\top P_{1,2} U_2)^{-1} \\
&= (U_1^\top M_1) \operatorname{diag}(M_3^\top \vec{\eta})(U_1^\top M_1)^{-1}\ (U_1^\top M_1 \operatorname{diag}(\vec{w}) M_2^\top U_2)\ (U_1^\top P_{1,2} U_2)^{-1} \\
&= (U_1^\top M_1) \operatorname{diag}(M_3^\top \vec{\eta})(U_1^\top M_1)^{-1}\ (U_1^\top P_{1,2} U_2)\ (U_1^\top P_{1,2} U_2)^{-1} \\
&= (U_1^\top M_1) \operatorname{diag}(M_3^\top \vec{\eta})(U_1^\top M_1)^{-1}.
\end{aligned}$$

B.3. Proof of Lemma 6

By Lemma 5,

$$\begin{aligned}
(U_1^\top M_1)^{-1} B_{1,2,3}(U_3 \vec{\theta}_i)(U_1^\top M_1) &= \operatorname{diag}(M_3^\top U_3 \vec{\theta}_i) \\
&= \operatorname{diag}(\langle \vec{\theta}_i, U_3^\top M_3 \vec{e}_1 \rangle, \langle \vec{\theta}_i, U_3^\top M_3 \vec{e}_2 \rangle, \ldots \langle \vec{\theta}_i, U_3^\top M_3 \vec{e}_k \rangle) \\
&= \operatorname{diag}(\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,k})
\end{aligned}$$

for all $i \in [k]$, and therefore

$$L = \begin{bmatrix}
\langle \vec{\theta}_1, U_3^\top M_3 \vec{e}_1 \rangle & \langle \vec{\theta}_1, U_3^\top M_3 \vec{e}_2 \rangle & \cdots & \langle \vec{\theta}_1, U_3^\top M_3 \vec{e}_k \rangle \\
\langle \vec{\theta}_2, U_3^\top M_3 \vec{e}_1 \rangle & \langle \vec{\theta}_2, U_3^\top M_3 \vec{e}_2 \rangle & \cdots & \langle \vec{\theta}_2, U_3^\top M_3 \vec{e}_3 \rangle \\
\vdots & \vdots & \ddots & \vdots \\
\langle \vec{\theta}_k, U_3^\top M_3 \vec{e}_1 \rangle & \langle \vec{\theta}_k, U_3^\top M_3 \vec{e}_2 \rangle & \cdots & \langle \vec{\theta}_k, U_3^\top M_3 \vec{e}_k \rangle
\end{bmatrix} = \Theta U_3^\top M_3.$$

B.4. Ordering issues

Although Algorithm B only explicitly yields estimates for M_3, it can easily be applied to estimate M_v for all other views v. The main caveat is that the estimators may not yield the same ordering of the columns, due to the unspecified order of the eigenvectors obtained in the third step of the method, and therefore some care is needed to obtain a consistent ordering. However, this ordering issue can be handled by exploiting consistency across the multiple views.

The first step is to perform the estimation of M_3 using Algorithm B as is. Then, to estimate M_2, one may re-use the eigenvectors in $\hat{R}_1$ to diagonalize $\hat{B}_{1,3,2}(\vec{\eta})$, as $B_{1,2,3}(\vec{\eta})$ and $B_{1,3,2}(\vec{\eta})$ share the same eigenvectors. The same goes for estimating M_v for other all other views v except $v = 1$.

It remains to provide a way to estimate M_1. Observe that M_2 can be estimated in at least two ways: via the operators $\hat{B}_{1,3,2}(\vec{\eta})$, or via the operators $\hat{B}_{3,1,2}(\vec{\eta})$. This is because the eigenvalues of $B_{3,1,2}(\vec{\eta})$ and $B_{1,3,2}(\vec{\eta})$ are the identical. Because the eigenvalues are also sufficiently separated from each other, the eigenvectors $\hat{R}_3$ of $\hat{B}_{3,1,2}(\vec{\eta})$ can be put in the same order as the eigenvectors $\hat{R}_1$ of $\hat{B}_{1,3,2}$ by (approximately) matching up their respective corresponding eigenvalues. Finally, the appropriately re-ordered eigenvectors $\hat{R}_3$ can then be used to diagonalize $\hat{B}_{3,2,1}(\vec{\eta})$ to estimate M_1.

B.5. Estimating the mixing weights

Given the estimate of $\hat{M}_3$, one can obtain an estimate of $\vec{w}$ using

$$\hat{w} := \hat{M}_3^{\dagger}\hat{\mathbb{E}}[\vec{x}_3]$$

where $A^{\dagger}$ denotes the Moore-Penrose pseudoinverse of A (though other generalized inverses may work as well), and $\hat{\mathbb{E}}[\vec{x}_3]$ is the empirical average of $\vec{x}_3$. This estimator is based on the following observation:

$$\mathbb{E}[\vec{x}_3] = \mathbb{E}[\mathbb{E}[\vec{x}_3|h]] = M_3\mathbb{E}[\vec{e}_h] = M_3\vec{w}$$

and therefore

$$M_3^{\dagger}\mathbb{E}[\vec{x}_3] = M_3^{\dagger}M_3\vec{w} = \vec{w}$$

since M_3 has full column rank.

B.6. Proof of Theorem 7

The proof is similar to that of Theorem 3, so we just describe the essential differences. As before, most perturbation arguments are deferred to Appendix C.

First, let E_1 be the event in which

$$\|\hat{P}_{1,2} - P_{1,2}\|_2 \leq C_{1,2} \cdot f(N,\delta),$$
$$\|\hat{P}_{1,3} - P_{1,3}\|_2 \leq C_{1,3} \cdot f(N,\delta)$$

and

$$\|\hat{P}_{1,2,3}(\hat{U}_3\vec{\theta}_i) - P_{1,2,3}(\hat{U}_3\vec{\theta}_i)\|_2 \leq C_{1,2,3} \cdot f(N,\delta/k)$$

for all $i \in [k]$. Therefore by Condition 3 and a union bound, we have $\Pr[E_1] \geq 1 - 3\delta$. Second, let E_2 be the event in which

$$\gamma := \min_{i\in[k]}\min_{j\neq j'} |\langle\vec{\theta}_i, \hat{U}_3^{\top}M_3(\vec{e}_j - \vec{e}_{j'})\rangle| > \frac{\min_{j\neq j'}\|\hat{U}_3^{\top}M_3(\vec{e}_j - \vec{e}_{j'})\|_2 \cdot \delta}{\sqrt{ek}\binom{k}{2}k}$$

and

$$\lambda_{\max} := \max_{i,j\in[k]} |\langle\vec{\theta}_i, \hat{U}_3^{\top}M_3\vec{e}_j\rangle| \leq \frac{\max_{j\in[k]}\|M_3\vec{e}_j\|_2}{\sqrt{k}}\Big(1 + \sqrt{2\ln(k^2/\delta)}\Big).$$

Since each $\vec{\theta}_i$ is distributed uniformly over $\mathcal{S}^{k-1}$, it follows from Lemma 15 and a union bound that $\Pr[E_2|E_1] \geq 1 - 2\delta$. Therefore $\Pr[E_1 \cap E_2] \geq (1-3\delta)(1-2\delta) \geq 1 - 5\delta$.

Let $U_3 \in \mathbb{R}^{d\times k}$ be the matrix of top k orthonormal left singular vectors of M_3. By Lemma 10 and the conditions on N, we have $\sigma_k(\hat{U}_3^{\top}U_3) \geq 1/2$, and therefore

$$\gamma \geq \frac{\min_{i\neq i'}\|M_3(\vec{e}_i - \vec{e}_{i'})\|_2 \cdot \delta}{2\sqrt{ek}\binom{k}{2}k} \quad \text{and} \quad \frac{\lambda_{\max}}{\gamma} \leq \frac{\sqrt{e}k^3(1 + \sqrt{2\ln(k^2/\delta)})}{\delta} \cdot \kappa'(M_3)$$

where

$$\kappa'(M_3) := \frac{\max_{i\in[m]}\|M_3\vec{e}_i\|_2}{\min_{i\neq i'}\|M_3(\vec{e}_i - \vec{e}_{i'})\|_2}.$$

Let $\vec{\eta}_i := \hat{U}_3 \vec{\theta}_i$ for $i \in [k]$. By Lemma 10, $\hat{U}_1^\top P_{1,2} \hat{U}_2$ is invertible, so we may define $\tilde{B}_{1,2,3}(\vec{\eta}_i) := (\hat{U}_1^\top P_{1,2,3}(\vec{\eta}_i)\hat{U}_2)(\hat{U}_1^\top P_{1,2}\hat{U}_2)^{-1}$. By Lemma 5,

$$\tilde{B}_{1,2,3}(\vec{\eta}_i) = (\hat{U}_1^\top M_1) \operatorname{diag}(M_3^\top \vec{\eta}_i)(\hat{U}_1^\top M_1)^{-1}.$$

Also define $R := \hat{U}_1^\top M_1 \operatorname{diag}(\|\hat{U}_1^\top M_1 \vec{e}_1\|_2, \|\hat{U}_1^\top M_1 \vec{e}_2\|_2, \ldots, \|\hat{U}_1^\top M_1 \vec{e}_k\|_2)^{-1}$. Using most of the same arguments in the proof of Theorem 3, we have

$$\|R^{-1}\|_2 \le 2\kappa(M_1), \tag{9}$$

$$\kappa(R) \le 4\kappa(M_1)^2, \tag{10}$$

$$\|\hat{B}_{1,2,3}(\vec{\eta}_i) - \tilde{B}_{1,2,3}(\vec{\eta}_i)\|_2 \le \frac{2\|\hat{P}_{1,2,3}(\vec{\eta}_i) - P_{1,2,3}(\vec{\eta}_i)\|_2}{\sigma_k(P_{1,2})} + \frac{2\|P_{1,2,3}\|_2 \cdot \|\hat{P}_{1,2} - P_{1,2}\|_2}{\sigma_k(P_{1,2})^2}.$$

By Lemma 12, the operator $\hat{B}_{1,2,3}(\vec{\eta}_1)$ has k distinct eigenvalues, and hence its matrix of right eigenvectors $\hat{R}_1$ is unique up to column scaling and ordering. This in turn implies that $\hat{R}_1^{-1}$ is unique up to row scaling and ordering. Therefore, for each $i \in [k]$, the $\hat{\lambda}_{i,j} = \vec{e}_j^\top \hat{R}_1^{-1} \hat{B}_{1,2,3}(\vec{\eta}_i)\hat{R}_1 \vec{e}_j$ for $j \in [k]$ are uniquely defined up to ordering. Moreover, by Lemma 13 and the above bounds on $\|\hat{B}_{1,2,3}(\vec{\eta}_i) - \tilde{B}_{1,2,3}(\vec{\eta}_i)\|_2$ and γ, there exists a permutation τ on $[k]$ such that, for all $i, j \in [k]$,

$$\begin{aligned} |\hat{\lambda}_{i,j} - \lambda_{i,\tau(j)}| &\le \Big(3\kappa(R) + 16k^{1.5} \cdot \kappa(R) \cdot \|R^{-1}\|_2^2 \cdot \lambda_{\max}/\gamma\Big) \cdot \|\hat{B}_{1,2,3}(\vec{\eta}_i) - \tilde{B}_{1,2,3}(\vec{\eta}_i)\|_2 \\ &\le \Big(12\kappa(M_1)^2 + 256k^{1.5} \cdot \kappa(M_1)^4 \cdot \lambda_{\max}/\gamma\Big) \cdot \|\hat{B}_{1,2,3}(\vec{\eta}_i) - \tilde{B}_{1,2,3}(\vec{\eta}_i)\|_2 \quad (11) \end{aligned}$$

where the second inequality uses (9) and (10). Let $\hat{\nu}_j := (\hat{\lambda}_{1,j}, \hat{\lambda}_{2,j}, \ldots, \hat{\lambda}_{k,j}) \in \mathbb{R}^k$ and $\vec{\nu}_j := (\lambda_{1,j}, \lambda_{2,j}, \ldots, \lambda_{k,j}) \in \mathbb{R}^k$. Observe that $\vec{\nu}_j = \Theta \hat{U}_3^\top M_3 \vec{e}_j = \Theta \hat{U}_3^\top \vec{\mu}_{3,j}$ by Lemma 6. By the orthogonality of Θ, the fact $\|\vec{v}\|_2 \le \sqrt{k}\|\vec{v}\|_\infty$ for $\vec{v} \in \mathbb{R}^k$, and (11)

$$\begin{aligned} \|\Theta^{-1}\hat{\nu}_j - \hat{U}_3^\top \vec{\mu}_{3,\tau(j)}\|_2 &= \|\Theta^{-1}(\hat{\nu}_j - \vec{\nu}_{\tau(j)})\|_2 \\ &= \|\hat{\nu}_j - \vec{\nu}_{\tau(j)}\|_2 \\ &\le \sqrt{k} \cdot \|\hat{\nu}_j - \vec{\nu}_{\tau(j)}\|_\infty \\ &= \sqrt{k} \cdot \max_i |\hat{\lambda}_{i,j} - \lambda_{i,\tau(j)}| \\ &\le \Big(12\sqrt{k} \cdot \kappa(M_1)^2 + 256k^2 \cdot \kappa(M_1)^4 \cdot \lambda_{\max}/\gamma\Big) \cdot \|\hat{B}_{1,2,3}(\vec{\eta}_i) - \tilde{B}_{1,2,3}(\vec{\eta}_i)\|_2. \end{aligned}$$

Finally, by Lemma 10 (as applied to $\hat{P}_{1,3}$ and $P_{1,3}$),

$$\|\hat{\mu}_{3,j} - \vec{\mu}_{3,\tau(j)}\|_2 \le \|\Theta^{-1}\hat{\nu}_j - \hat{U}_3^\top \vec{\mu}_{3,\tau(j)}\|_2 + 2\|\vec{\mu}_{3,\tau(j)}\|_2 \cdot \frac{\|\hat{P}_{1,3} - P_{1,3}\|_2}{\sigma_k(P_{1,3})}.$$

Making all of the substitutions into the above bound gives

$$\begin{aligned}
\|\hat{\mu}_{3,j} - \vec{\mu}_{3,\tau(j)}\|_2 &\le \frac{C}{6} \cdot k^5 \cdot \kappa(M_1)^4 \cdot \kappa'(M_3) \cdot \frac{\ln(k/\delta)}{\delta} \cdot \biggl(\frac{C_{1,2,3} \cdot f(N, \delta/k)}{\sigma_k(P_{1,2})} \\
&\qquad + \frac{\|P_{1,2,3}\|_2 \cdot C_{1,2} \cdot f(N/\delta)}{\sigma_k(P_{1,2})^2} \biggr) + \frac{C}{6} \cdot \|\vec{\mu}_{3,\tau(j)}\|_2 \cdot \frac{C_{1,3} \cdot f(N,\delta)}{\sigma_k(P_{1,3})} \\
&\le \frac{1}{2} \Bigl(\max_{j' \in [k]} \|\vec{\mu}_{3,j'}\|_2 + \|\vec{\mu}_{3,\tau(j)}\|_2 \Bigr) \cdot \epsilon \\
&\le \max_{j' \in [k]} \|\vec{\mu}_{3,j'}\|_2 \cdot \epsilon.
\end{aligned}$$

Appendix C. Perturbation analysis for observable operators

The following lemma establishes the accuracy of approximating the fundamental subspaces (*i.e.*, the row and column spaces) of a matrix X by computing the singular value decomposition of a perturbation $\hat{X}$ of X.

Lemma 10 *Let $X \in \mathbb{R}^{m \times n}$ be a matrix of rank k. Let $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$ be matrices with orthonormal columns such that* $\operatorname{range}(U)$ *and* $\operatorname{range}(V)$ *are spanned by, respectively, the left and right singular vectors of X corresponding to its k largest singular values. Similarly define $\hat{U} \in \mathbb{R}^{m \times k}$ and $\hat{V} \in \mathbb{R}^{n \times k}$ relative to a matrix $\hat{X} \in \mathbb{R}^{m \times n}$. Define $\epsilon_X := \|\hat{X} - X\|_2$, $\varepsilon_0 := \frac{\epsilon_X}{\sigma_k(X)}$, and $\varepsilon_1 := \frac{\varepsilon_0}{1-\varepsilon_0}$. Assume $\varepsilon_0 < \frac{1}{2}$. Then*

1. $\varepsilon_1 < 1$;
2. $\sigma_k(\hat{X}) = \sigma_k(\hat{U}^\top \hat{X} \hat{V}) \ge (1 - \varepsilon_0) \cdot \sigma_k(X) > 0$;
3. $\sigma_k(\hat{U}^\top U) \ge \sqrt{1 - \varepsilon_1^2}$;
4. $\sigma_k(\hat{V}^\top V) \ge \sqrt{1 - \varepsilon_1^2}$;
5. $\sigma_k(\hat{U}^\top X \hat{V}) \ge (1 - \varepsilon_1^2) \cdot \sigma_k(X)$;
6. *for any* $\hat{\alpha} \in \mathbb{R}^k$ *and* $\vec{v} \in \operatorname{range}(U)$, $\|\hat{U}\hat{\alpha} - \vec{v}\|_2^2 \le \|\hat{\alpha} - \hat{U}^\top \vec{v}\|_2^2 + \|\vec{v}\|_2^2 \cdot \varepsilon_1^2$.

Proof The first claim follows from the assumption on ε_0. The second claim follows from the assumptions and Weyl's theorem (Lemma 20). Let the columns of $\hat{U}_\perp \in \mathbb{R}^{m \times (m-k)}$ be an orthonormal basis for the orthogonal complement of $\operatorname{range}(\hat{U})$, so that $\|\hat{U}_\perp^\top U\|_2 \le \epsilon_X / \sigma_k(\hat{X}) \le \varepsilon_1$ by Wedin's theorem (Lemma 21). The third claim then follows because $\|\hat{U}^\top U\|_2^2 = 1 - \|\hat{U}_\perp^\top U\|_2^2 \ge 1 - \varepsilon_1^2$. The fourth claim is analogous to the third claim, and the fifth claim follows from the third and fourth. The sixth claim follows writing $\vec{v} = U\vec{\alpha}$ for some $\vec{\alpha} \in \mathbb{R}^k$, and using the decomposition $\|\hat{U}\hat{\alpha} - \vec{v}\|_2^2 = \|\hat{U}\hat{\alpha} - \hat{U}\hat{U}^\top \vec{v}\|_2^2 + \|\hat{U}_\perp \hat{U}_\perp^\top \vec{v}\|_2^2 = \|\hat{\alpha} - \hat{U}^\top \vec{v}\|_2^2 + \|\hat{U}_\perp^\top (U\vec{\alpha})\|_2^2 \le \|\hat{\alpha} - \hat{U}^\top \vec{v}\|_2^2 + \|\hat{U}_\perp^\top U\|_2^2 \|\vec{\alpha}\|_2^2 \le \|\hat{\alpha} - \hat{U}^\top \vec{v}\|_2^2 + \|\vec{\alpha}\|_2^2 \cdot \varepsilon_1^2 = \|\hat{\alpha} - \hat{U}^\top U\vec{\alpha}\|_2^2 + \|\vec{v}\|_2^2 \cdot \varepsilon_1^2$ where the last inequality follows from the argument for the third claim, and the last equality uses the orthonormality of the columns of U. ■

The next lemma bounds the error of the observation operator in terms of the errors in estimating the second-order and third-order moments.

Lemma 11 *Consider the setting and definitions from Lemma 10, and let $Y \in \mathbb{R}^{m\times n}$ and $\hat{Y} \in \mathbb{R}^{m\times n}$ be given. Define $\varepsilon_2 := \frac{\varepsilon_0}{(1-\varepsilon_1^2)\cdot(1-\varepsilon_0-\varepsilon_1^2)}$ and $\epsilon_Y := \|\hat{Y} - Y\|_2$. Assume $\varepsilon_0 < \frac{1}{1+\sqrt{2}}$. Then*

1. *$\hat{U}^\top \hat{X}\hat{V}$ and $\hat{U}^\top X\hat{V}$ are both invertible, and $\|(\hat{U}^\top \hat{X}\hat{V})^{-1} - (\hat{U}^\top X\hat{V})^{-1}\|_2 \le \frac{\varepsilon_2}{\sigma_k(X)}$;*
2. *$\|(\hat{U}^\top \hat{Y}\hat{V})(\hat{U}^\top \hat{X}\hat{V})^{-1} - (\hat{U}^\top Y\hat{V})(\hat{U}^\top X\hat{V})^{-1}\|_2 \le \frac{\epsilon_Y}{(1-\varepsilon_0)\cdot\sigma_k(X)} + \frac{\|Y\|_2\cdot\varepsilon_2}{\sigma_k(X)}$.*

Proof Let $\hat{S} := \hat{U}^\top \hat{X}\hat{V}$ and $\tilde{S} := \hat{U}^\top X\hat{V}$. By Lemma 10, $\hat{U}^\top \hat{X}\hat{V}$ is invertible, $\sigma_k(\tilde{S}) \ge \sigma_k(\hat{U}^\top U)\cdot\sigma_k(X)\cdot\sigma_k(\hat{V}^\top V) \ge (1-\varepsilon_1^2)\cdot\sigma_k(X)$ (so $\tilde{S}$ is also invertible), and $\|\hat{S} - \tilde{S}\|_2 \le \varepsilon_0\cdot\sigma_k(X) \le \frac{\varepsilon_0}{1-\varepsilon_1^2}\cdot\sigma_k(\tilde{S})$. The assumption on ε_0 implies $\frac{\varepsilon_0}{1-\varepsilon_1^2} < 1$; therefore the Lemma 23 implies $\|\hat{S}^{-1} - \tilde{S}^{-1}\|_2 \le \frac{\|\hat{S}-\tilde{S}\|_2/\sigma_k(\tilde{S})}{1-\|\hat{S}-\tilde{S}\|_2/\sigma_k(\tilde{S})}\cdot\frac{1}{\sigma_k(\tilde{S})} \le \frac{\varepsilon_2}{\sigma_k(X)}$, which proves the first claim. For the second claim, observe that

$$
\begin{aligned}
&\|(\hat{U}^\top \hat{Y}\hat{V})(\hat{U}^\top \hat{X}\hat{V})^{-1} - (\hat{U}^\top Y\hat{V})(\hat{U}^\top X\hat{V})^{-1}\|_2 \\
&\le \|(\hat{U}^\top \hat{Y}\hat{V})(\hat{U}^\top \hat{X}\hat{V})^{-1} - (\hat{U}^\top Y\hat{V})(\hat{U}^\top \hat{X}\hat{V})^{-1}\|_2 \\
&\quad + \|(\hat{U}^\top Y\hat{V})(\hat{U}^\top \hat{X}\hat{V})^{-1} - (\hat{U}^\top Y\hat{V})(\hat{U}^\top X\hat{V})^{-1}\|_2 \\
&\le \|\hat{U}^\top \hat{Y}\hat{V} - \hat{U}^\top Y\hat{V}\|_2\cdot\|(\hat{U}^\top \hat{X}\hat{V})^{-1}\|_2 + \|\hat{U}^\top Y\hat{V}\|_2\cdot\|(\hat{U}^\top \hat{X}\hat{V})^{-1} - (\hat{U}^\top X\hat{V})^{-1}\|_2 \\
&\le \frac{\epsilon_Y}{(1-\varepsilon_0)\cdot\sigma_k(X)} + \frac{\|Y\|_2\cdot\varepsilon_2}{\sigma_k(X)}
\end{aligned}
$$

where the first inequality follows from the triangle inequality, the second follows from the sub-multiplicative property of the spectral norm, and the last follows from Lemma 10 and the first claim. ■

The following lemma establishes standard eigenvalue and eigenvector perturbation bounds.

Lemma 12 *Let $A \in \mathbb{R}^{k\times k}$ be a diagonalizable matrix with k distinct real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{R}$ corresponding to the (right) eigenvectors $\vec{\xi}_1, \vec{\xi}_2, \ldots, \vec{\xi}_k \in \mathbb{R}^k$ all normalized to have $\|\vec{\xi}_i\|_2 = 1$. Let $R \in \mathbb{R}^{k\times k}$ be the matrix whose i-th column is $\vec{\xi}_i$. Let $\hat{A} \in \mathbb{R}^{k\times k}$ be a matrix. Define $\epsilon_A := \|\hat{A} - A\|_2$, $\gamma_A := \min_{i\ne j}|\lambda_i - \lambda_j|$, and $\varepsilon_3 := \frac{\kappa(R)\cdot\epsilon_A}{\gamma_A}$. Assume $\varepsilon_3 < \frac{1}{2}$. Then there exists a permutation τ on $[k]$ such that the following holds:*

1. *$\hat{A}$ has k distinct real eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k \in \mathbb{R}$, and $|\hat{\lambda}_{\tau(i)} - \lambda_i| \le \varepsilon_3\cdot\gamma_A$ for all $i \in [k]$;*
2. *$\hat{A}$ has corresponding (right) eigenvectors $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_k \in \mathbb{R}^k$, normalized to have $\|\hat{\xi}_i\|_2 = 1$, which satisfy $\|\hat{\xi}_{\tau(i)} - \vec{\xi}_i\|_2 \le 4(k-1)\cdot\|R^{-1}\|_2\cdot\varepsilon_3$ for all $i \in [k]$;*
3. *the matrix $\hat{R} \in \mathbb{R}^{k\times k}$ whose i-th column is $\hat{\xi}_{\tau(i)}$ satisfies $\|\hat{R} - R\|_2 \le \|\hat{R} - R\|_{\mathrm{F}} \le 4k^{1/2}(k-1)\cdot\|R^{-1}\|_2\cdot\varepsilon_3$.*

Proof The Bauer-Fike theorem (Lemma 22) implies that for every eigenvalue $\hat{\lambda}_i$ of $\hat{A}$, there exists an eigenvalue λ_j of A such that $|\hat{\lambda}_i - \lambda_j| \le \|R^{-1}(\hat{A} - A)R\|_2 \le \varepsilon_3\cdot\gamma_A$. Therefore, the assumption on ε_3 implies that there exists a permutation τ such that $|\hat{\lambda}_{\tau(i)} - \lambda_i| \leq \varepsilon_3 \cdot \gamma_A < \frac{\gamma_A}{2}$. In particular,

$$\left| \left[\lambda_i - \frac{\gamma_A}{2}, \lambda_i + \frac{\gamma_A}{2} \right] \cap \{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k\} \right| = 1, \quad \forall i \in [k]. \tag{12}$$

Since $\hat{A}$ is real, all non-real eigenvalues of $\hat{A}$ must come in conjugate pairs; so the existence of a non-real eigenvalue of $\hat{A}$ would contradict (12). This proves the first claim.

For the second claim, assume for notational simplicity that the permutation τ is the identity permutation. Let $\hat{R} \in \mathbb{R}^{k \times k}$ be the matrix whose i-th column is $\hat{\xi}_i$. Define $\vec{\zeta}_i^\top \in \mathbb{R}^k$ to be the i-th row of R^{-1} (*i.e.*, the i-th left eigenvector of A), and similarly define $\hat{\zeta}_i^\top \in \mathbb{R}^k$ to be the i-th row of $\hat{R}^{-1}$. Fix a particular $i \in [k]$. Since $\{\vec{\xi}_1, \vec{\xi}_2, \ldots, \vec{\xi}_k\}$ forms a basis for $\mathbb{R}^k$, we can write $\hat{\xi}_i = \sum_{j=1}^k c_{i,j} \vec{\xi}_j$ for some coefficients $c_{i,1}, c_{i,2}, \ldots, c_{i,k} \in \mathbb{R}$. We may assume $c_{i,i} \geq 0$ (or else we replace $\hat{\xi}_i$ with $-\hat{\xi}_i$). The fact that $\|\hat{\xi}_i\|_2 = \|\vec{\xi}_j\|_2 = 1$ for all $j \in [k]$ and the triangle inequality imply $1 = \|\hat{\xi}_i\|_2 \leq c_{i,i}\|\vec{\xi}_i\|_2 + \sum_{j \neq i} |c_{i,j}| \|\vec{\xi}_j\|_2 = c_{i,i} + \sum_{j \neq i} |c_{i,j}|$, and therefore

$$\|\hat{\xi}_i - \vec{\xi}_i\|_2 \leq |1 - c_{i,i}| \|\vec{\xi}_i\|_2 + \sum_{j \neq i} |c_{i,j}| \|\vec{\xi}_j\|_2 \leq 2 \sum_{j \neq i} |c_{i,j}|$$

again by the triangle inequality. Therefore, it suffices to show $|c_{i,j}| \leq 2\|R^{-1}\|_2 \cdot \varepsilon_3$ for $j \neq i$ to prove the second claim.

Observe that $A\hat{\xi}_i = A(\sum_{i'=1}^k c_{i,i'} \vec{\xi}_{i'}) = \sum_{i'=1}^k c_{i,i'} \lambda_{i'} \vec{\xi}_{i'}$, and therefore

$$\sum_{i'=1}^k c_{i,i'} \lambda_{i'} \vec{\xi}_{i'} + (\hat{A} - A)\hat{\xi}_i = \hat{A}\hat{\xi}_i = \hat{\lambda}_i \hat{\xi}_i = \lambda_i \sum_{i'=1}^k c_{i,i'} \vec{\xi}_{i'} + (\hat{\lambda}_i - \lambda_i)\hat{\xi}_i.$$

Multiplying through the above equation by $\vec{\zeta}_j^\top$, and using the fact that $\vec{\zeta}_j^\top \vec{\xi}_{i'} = \mathbb{1}\{j = i'\}$ gives

$$c_{i,j}\lambda_j + \vec{\zeta}_i^\top (\hat{A} - A)\hat{\xi}_i = \lambda_i c_{i,j} + (\hat{\lambda}_i - \lambda_i)\vec{\zeta}_j^\top \hat{\xi}_i.$$

The above equation rearranges to $(\lambda_j - \lambda_i)c_{i,j} = (\hat{\lambda}_i - \lambda_i)\vec{\zeta}_j^\top \hat{\xi}_i + \vec{\zeta}_j^\top (A - \hat{A})\hat{\xi}_i$ and therefore

$$|c_{i,j}| \leq \frac{\|\vec{\zeta}_j\|_2 \cdot (|\hat{\lambda}_i - \lambda_i| + \|(\hat{A} - A)\hat{\xi}_i\|_2)}{|\lambda_j - \lambda_i|} \leq \frac{\|R^{-1}\|_2 \cdot (|\hat{\lambda}_i - \lambda_i| + \|\hat{A} - A\|_2)}{|\lambda_j - \lambda_i|}$$

by the Cauchy-Schwarz and triangle inequalities and the sub-multiplicative property of the spectral norm. The bound $|c_{i,j}| \leq 2\|R^{-1}\|_2 \cdot \varepsilon_3$ then follows from the first claim.

The third claim follows from standard comparisons of matrix norms. ∎

The next lemma gives perturbation bounds for estimating the eigenvalues of simultaneously diagonalizable matrices $A_1, A_2, \ldots, A_k$. The eigenvectors $\hat{R}$ are taken from a perturbation of the first matrix A_1, and are then subsequently used to approximately diagonalize the perturbations of the remaining matrices $A_2, \ldots, A_k$. In practice, one may use Jacobi-like procedures to approximately solve the joint eigenvalue problem.

Lemma 13 *Let $A_1, A_2, \ldots, A_k \in \mathbb{R}^{k \times k}$ be diagonalizable matrices that are diagonalized by the same matrix invertible $R \in \mathbb{R}^{k \times k}$ with unit length columns $\|R\vec{e}_j\|_2 = 1$, such that each A_i has k distinct real eigenvalues:*

$$R^{-1} A_i R = \operatorname{diag}(\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,k}).$$

Let $\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_k \in \mathbb{R}^{k \times k}$ be given. Define $\epsilon_A := \max_i \|\hat{A}_i - A_i\|_2$, $\gamma_A := \min_i \min_{j \neq j'} |\lambda_{i,j} - \lambda_{i,j'}|$, $\lambda_{\max} := \max_{i,j} |\lambda_{i,j}|$, $\varepsilon_3 := \frac{\kappa(R) \cdot \epsilon_A}{\gamma_A}$, and $\varepsilon_4 := 4k^{1.5} \cdot \|R^{-1}\|_2^2 \cdot \varepsilon_3$. Assume $\varepsilon_3 < \frac{1}{2}$ and $\varepsilon_4 < 1$. Then there exists a permutation τ on $[k]$ such that the following holds.

1. *The matrix $\hat{A}_1$ has k distinct real eigenvalues $\hat{\lambda}_{1,1}, \hat{\lambda}_{1,2}, \ldots, \hat{\lambda}_{1,k} \in \mathbb{R}$, and $|\hat{\lambda}_{1,j} - \lambda_{1,\tau(j)}| \leq \varepsilon_3 \cdot \gamma_A$ for all $j \in [k]$.*

2. *There exists a matrix $\hat{R} \in \mathbb{R}^{k \times k}$ whose j-th column is a right eigenvector corresponding to $\hat{\lambda}_{1,j}$, scaled so $\|\hat{R}\vec{e}_j\|_2 = 1$ for all $j \in [k]$, such that $\|\hat{R} - R_\tau\|_2 \leq \frac{\varepsilon_4}{\|R^{-1}\|_2}$, where R_τ is the matrix obtained by permuting the columns of R with τ.*

3. *The matrix $\hat{R}$ is invertible and its inverse satisfies $\|\hat{R}^{-1} - R_\tau^{-1}\|_2 \leq \|R^{-1}\|_2 \cdot \frac{\varepsilon_4}{1 - \varepsilon_4}$;*

4. *For all $i \in \{2, 3, \ldots, k\}$ and all $j \in [k]$, the (j,j)-th element of $\hat{R}^{-1}\hat{A}_i\hat{R}$, denoted by $\hat{\lambda}_{i,j} := \vec{e}_j^\top \hat{R}^{-1} \hat{A}_i \hat{R} \vec{e}_j$, satisfies*

$$\begin{aligned} |\hat{\lambda}_{i,j} - \lambda_{i,\tau(j)}| &\leq \left(1 + \frac{\varepsilon_4}{1 - \varepsilon_4}\right) \cdot \left(1 + \frac{\varepsilon_4}{\sqrt{k} \cdot \kappa(R)}\right) \cdot \varepsilon_3 \cdot \gamma_A \\ &\quad + \kappa(R) \cdot \left(\frac{1}{1 - \varepsilon_4} + \frac{1}{\sqrt{k} \cdot \kappa(R)} + \frac{1}{\sqrt{k}} \cdot \frac{\varepsilon_4}{1 - \varepsilon_4}\right) \cdot \varepsilon_4 \cdot \lambda_{\max}. \end{aligned}$$

   If $\varepsilon_4 \leq \frac{1}{2}$, then $|\hat{\lambda}_{i,j} - \lambda_{i,\tau(j)}| \leq 3\varepsilon_3 \cdot \gamma_A + 4\kappa(R) \cdot \varepsilon_4 \cdot \lambda_{\max}$.

Proof The first and second claims follow from applying Lemma 12 to A_1 and $\hat{A}_1$. The third claim follows from applying Lemma 23 to $\hat{R}$ and R_τ. To prove the last claim, first define $\vec{\zeta}_j^\top \in \mathbb{R}^k$ ($\hat{\zeta}_j^\top$) to be the j-th row of R_τ^{-1} ($\hat{R}^{-1}$), and $\vec{\xi}_j \in \mathbb{R}^k$ ($\hat{\xi}_j$) to be the j-th column of R_τ ($\hat{R}$), so $\vec{\zeta}_j^\top A_i \vec{\xi}_j = \lambda_{i,\tau(j)}$ and $\hat{\zeta}_j^\top \hat{A}_i \hat{\xi}_j = \vec{e}_j^\top \hat{R}^{-1} \hat{A}_i \hat{R} \vec{e}_j = \hat{\lambda}_{i,j}$. By the triangle and Cauchy-Schwarz inequalities and the sub-multiplicative property of the spectral norm,

$$\begin{aligned} &|\hat{\lambda}_{i,j} - \lambda_{i,\tau(j)}| \\ &= |\hat{\zeta}_j^\top \hat{A}_i \hat{\xi}_j - \vec{\zeta}_j^\top A_i \vec{\xi}_j| \\ &= |\vec{\zeta}_j^\top (\hat{A}_i - A_i)\vec{\xi}_j + \vec{\zeta}_j^\top (\hat{A}_i - A_i)(\hat{\xi}_j - \vec{\xi}_j) + (\hat{\zeta}_j - \vec{\zeta}_j)^\top (\hat{A}_i - A_i)\vec{\xi}_j \\ &\quad + (\hat{\zeta}_j - \vec{\zeta}_j)^\top (\hat{A}_i - A_i)(\hat{\xi}_j - \vec{\xi}_j) + (\hat{\zeta}_j - \vec{\zeta}_j)^\top A_i \vec{\xi}_j + \vec{\zeta}_j^\top A_i (\hat{\xi}_j - \vec{\xi}_j) + (\hat{\zeta}_j - \vec{\zeta}_j)^\top A_i (\hat{\xi}_j - \vec{\xi}_j)| \\ &\leq |\vec{\zeta}_j^\top (\hat{A}_i - A_i)\vec{\xi}_j| + |\vec{\zeta}_j^\top (\hat{A}_i - A_i)(\hat{\xi}_j - \vec{\xi}_j)| + |(\hat{\zeta}_j - \vec{\zeta}_j)^\top (\hat{A}_i - A_i)\vec{\xi}_j| \\ &\quad + |(\hat{\zeta}_j - \vec{\zeta}_j)^\top (\hat{A}_i - A_i)(\hat{\xi}_j - \vec{\xi}_j)| + |(\hat{\zeta}_j - \vec{\zeta}_j)^\top A_i \vec{\xi}_j| + |\vec{\zeta}_j^\top A_i (\hat{\xi}_j - \vec{\xi}_j)| + |(\hat{\zeta}_j - \vec{\zeta}_j)^\top A_i (\hat{\xi}_j - \vec{\xi}_j)| \\ &\leq \|\vec{\zeta}_j\|_2 \cdot \|\hat{A}_i - A_i\|_2 \cdot \|\vec{\xi}_j\|_2 + \|\vec{\zeta}_j\|_2 \cdot \|\hat{A}_i - A_i\|_2 \cdot \|\hat{\xi}_j - \vec{\xi}_j\|_2 + \|\hat{\zeta}_j - \vec{\zeta}_j\|_2 \cdot \|\hat{A}_i - A_i\|_2 \|\vec{\xi}_j\|_2 \\ &\quad + \|\hat{\zeta}_j - \vec{\zeta}_j\|_2 \cdot \|\hat{A}_i - A_i\|_2 \cdot \|\hat{\xi}_j - \vec{\xi}_j\|_2 \\ &\quad + \|\hat{\zeta}_j - \vec{\zeta}_j\|_2 \cdot \|\lambda_{i,\tau(j)}\vec{\xi}_j\|_2 + \|\lambda_{i,\tau(j)}\vec{\zeta}_j\|_2 \cdot \|\hat{\xi}_j - \vec{\xi}_j\|_2 + \|\hat{\zeta}_j - \vec{\zeta}_j\|_2 \cdot \|A_i\|_2 \cdot \|\hat{\xi}_j - \vec{\xi}_j\|_2. \end{aligned} \tag{13}$$

Observe that $\|\vec{\zeta}_j\|_2 \leq \|R^{-1}\|_2$, $\|\vec{\xi}_j\|_2 \leq \|R\|_2$, $\|\hat{\zeta}_j - \vec{\zeta}_j\|_2 \leq \|\hat{R}^{-1} - R_\tau^{-1}\|_2 \leq \|R^{-1}\|_2 \cdot \frac{\varepsilon_4}{1-\varepsilon_4}$, $\|\hat{\xi}_j - \vec{\xi}_j\|_2 \leq 4k \cdot \|R^{-1}\|_2 \cdot \varepsilon_3$ (by Lemma 12), and $\|A_i\|_2 \leq \|R\|_2 \cdot (\max_j |\lambda_{i,j}|) \cdot \|R^{-1}\|_2$. Therefore, continuing from (13), $|\hat{\lambda}_{i,j} - \lambda_{i,\tau(j)}|$ is bounded as

$$\begin{aligned}
|\hat{\lambda}_{i,j} - \lambda_{i,\tau(j)}| \leq{} & \|R^{-1}\|_2 \cdot \|R\|_2 \cdot \epsilon_A + \|R^{-1}\|_2 \cdot \epsilon_A \cdot 4k \cdot \|R^{-1}\|_2 \cdot \varepsilon_3 + \|R^{-1}\|_2 \cdot \frac{\varepsilon_4}{1-\varepsilon_4} \cdot \epsilon_A \cdot \|R\|_2 \\
& + \|R^{-1}\|_2 \cdot \frac{\varepsilon_4}{1-\varepsilon_4} \cdot \epsilon_A \cdot 4k \cdot \|R^{-1}\|_2 \cdot \varepsilon_3 \\
& + \lambda_{\max} \cdot \|R^{-1}\|_2 \cdot \frac{\varepsilon_4}{1-\varepsilon_4} \cdot \|R\|_2 + \lambda_{\max} \cdot \|R^{-1}\|_2 \cdot 4k \cdot \|R^{-1}\|_2 \cdot \varepsilon_3 \\
& + \|R^{-1}\|_2 \cdot \frac{\varepsilon_4}{1-\varepsilon_4} \cdot \|R\|_2 \cdot \lambda_{\max} \cdot \|R^{-1}\|_2 \cdot 4k \cdot \|R^{-1}\|_2 \cdot \varepsilon_3 \\
={} & \varepsilon_3 \cdot \gamma_A + \frac{\varepsilon_4}{\sqrt{k} \cdot \kappa(R)} \cdot \varepsilon_3 \cdot \gamma_A + \frac{\varepsilon_4}{1-\varepsilon_4} \cdot \varepsilon_3 \cdot \gamma_A \\
& + \frac{\varepsilon_4}{\sqrt{k} \cdot \kappa(R)} \cdot \frac{\varepsilon_4}{1-\varepsilon_4} \cdot \varepsilon_3 \cdot \gamma_A \\
& + \kappa(R) \cdot \frac{1}{1-\varepsilon_4} \cdot \varepsilon_4 \cdot \lambda_{\max} + \frac{1}{\sqrt{k}} \cdot \varepsilon_4 \cdot \lambda_{\max} + \frac{\kappa(R)}{\sqrt{k}} \cdot \frac{\varepsilon_4}{1-\varepsilon_4} \cdot \varepsilon_4 \cdot \lambda_{\max}.
\end{aligned}$$

Rearranging gives the claimed inequality. ■

Lemma 14 *Let $V \in \mathbb{R}^{k \times k}$ be an invertible matrix, and let $R \in \mathbb{R}^{k \times k}$ be the matrix whose j-th column is $V\vec{e}_j / \|V\vec{e}_j\|_2$. Then $\|R\|_2 \leq \kappa(V)$, $\|R^{-1}\|_2 \leq \kappa(V)$, and $\kappa(R) \leq \kappa(V)^2$.*

Proof We have $R = V \operatorname{diag}(\|V\vec{e}_1\|_2, \|V\vec{e}_2\|_2, \ldots, \|V\vec{e}_k\|_2)^{-1}$, so by the sub-multiplicative property of the spectral norm, $\|R\|_2 \leq \|V\|_2 / \min_j \|V\vec{e}_j\|_2 \leq \|V\|_2 / \sigma_k(V) = \kappa(V)$. Similarly, $\|R^{-1}\|_2 \leq \|V^{-1}\|_2 \cdot \max_j \|V\vec{e}_j\|_2 \leq \|V^{-1}\|_2 \cdot \|V\|_2 = \kappa(V)$. ■

The next lemma shows that randomly projecting a collection of vectors to $\mathbb{R}$ does not collapse any two too close together, nor does it send any of them too far away from zero.

Lemma 15 *Fix any $\delta \in (0,1)$ and matrix $A \in \mathbb{R}^{m \times n}$ (with $m \leq n$). Let $\vec{\theta} \in \mathbb{R}^m$ be a random vector distributed uniformly over $\mathcal{S}^{m-1}$.*

1. $\Pr\left[\min_{i \neq j} |\langle \vec{\theta}, A(\vec{e}_i - \vec{e}_j) \rangle| > \frac{\min_{i \neq j} \|A(\vec{e}_i - \vec{e}_j)\|_2 \cdot \delta}{\sqrt{em}\binom{n}{2}}\right] \geq 1 - \delta.$
2. $\Pr\left[\forall i \in [m],\ |\langle \vec{\theta}, A\vec{e}_i \rangle| \leq \frac{\|A\vec{e}_i\|_2}{\sqrt{m}}\left(1 + \sqrt{2\ln(m/\delta)}\right)\right] \geq 1 - \delta.$

Proof For the first claim, let $\delta_0 := \delta / \binom{n}{2}$. By Lemma 25, for any fixed pair $\{i,j\} \subseteq [n]$ and $\beta := \delta_0/\sqrt{e}$,

$$\Pr\left[|\langle \vec{\theta}, A(\vec{e}_i - \vec{e}_j) \rangle| \leq \|A(\vec{e}_i - \vec{e}_j)\|_2 \cdot \frac{1}{\sqrt{m}} \cdot \frac{\delta_0}{\sqrt{e}}\right] \leq \exp\left(\frac{1}{2}(1 - (\delta_0^2/e) + \ln(\delta_0^2/e))\right) \leq \delta_0.$$

Therefore the first claim follows by a union bound over all $\binom{n}{2}$ pairs $\{i, j\}$.

For the second claim, apply Lemma 25 with $\beta := 1 + t$ and $t := \sqrt{2\ln(m/\delta)}$ to obtain

$$\begin{aligned} \Pr\left[|\langle \vec{\theta}, A\vec{e}_i \rangle| \geq \frac{\|A\vec{e}_i\|_2}{\sqrt{m}} \cdot (1 + t) \right] &\leq \exp\left(\frac{1}{2}\Big(1 - (1+t)^2 + 2\ln(1+t)\Big) \right) \\ &\leq \exp\left(\frac{1}{2}\Big(1 - (1+t)^2 + 2t\Big) \right) \\ &= e^{-t^2/2} = \delta/m. \end{aligned}$$

Therefore the second claim follows by taking a union bound over all $i \in [m]$. ■

Appendix D. Proofs and details from Section 4

In this section, we provide omitted proofs and details from Section 4.

D.1. Learning mixtures of product distributions

In this section, we show how to use Algorithm B with mixtures of product distributions in $\mathbb{R}^n$ that satisfy an *incoherence condition* on the means $\vec{\mu}_1, \vec{\mu}_2, \dots, \vec{\mu}_k \in \mathbb{R}^n$ of k component distributions. Note that product distributions are just a special case of the more general class of multi-view distributions, which are directly handled by Algorithm B.

The basic idea is to randomly partition the coordinates into $\ell \geq 3$ "views", each of roughly the same dimension. Under the assumption that the component distributions are product distributions, the multi-view assumption is satisfied. What remains to be checked is that the non-degeneracy condition (Condition 2) is satisfied. Theorem 16 (below) shows that it suffices that the original matrix of component means have rank k and satisfy the following incoherence condition.

Condition 4 (Incoherence condition) *Let $\delta \in (0,1)$, $\ell \in [n]$, and $M = [\vec{\mu}_1|\vec{\mu}_2|\cdots|\vec{\mu}_k] \in \mathbb{R}^{n\times k}$ be given; let $M = USV^\top$ be the thin singular value decomposition of M, where $U \in \mathbb{R}^{n\times k}$ is a matrix of orthonormal columns, $S = \operatorname{diag}(\sigma_1(M), \sigma_2(M), \dots, \sigma_k(M)) \in \mathbb{R}^{k\times k}$, and $V \in \mathbb{R}^{k\times k}$ is orthogonal; and let*

$$c_M := \max_{j\in[n]} \left\{ \frac{n}{k} \cdot \|U^\top \vec{e}_j\|_2^2 \right\}.$$

The following inequality holds:

$$c_M \leq \frac{9}{32} \cdot \frac{\lfloor n/\ell \rfloor}{k \cdot \ln \frac{\ell\cdot k}{\delta}}.$$

Note that c_M is always in the interval $[1, n/k]$; it is smallest when the left singular vectors in U have $\pm 1/\sqrt{n}$ entries (as in a Hadamard basis), and largest when the singular vectors are the coordinate axes. Roughly speaking, the incoherence condition requires that the non-degeneracy of a matrix M be witnessed by many vertical blocks of M. When the condition is satisfied, then with high probability, a random partitioning of the coordinates into ℓ groups induces a block partitioning of M into ℓ matrices $M_1, M_2, \dots, M_\ell$ (with roughly equal number of rows) such that the k-th largest singular value of M_v is not much smaller than that of M (for each $v \in [\ell]$).

Chaudhuri and Rao (2008) show that under a similar condition (which they call a *spreading condition*), a random partitioning of the coordinates into two "views" preserves the separation between the means of k component distributions. They then follow this preprocessing with a projection based on the correlations across the two views (similar to CCA). However, their overall algorithm requires a minimum separation condition on the means of the component distributions. In contrast, Algorithm B does not require a minimum separation condition at all in this setting.

Theorem 16 *Assume Condition 4 holds. With probability at least $1-\delta$, a uniformly chosen random partitioning of $[n]$ into ℓ disjoint sets $[n] = \mathcal{I}_1 \cup \mathcal{I}_2 \cup \cdots \cup \mathcal{I}_\ell$, each of size at least*

$$|\mathcal{I}_v| \geq \left\lceil \frac{32}{9} \cdot c_M \cdot k \cdot \ln \frac{\ell \cdot k}{\delta} \right\rceil,$$

has the following property: for each $v \in [\ell]$, the matrix $M_v \in \mathbb{R}^{|\mathcal{I}_v| \times k}$ formed by selecting the rows of M indexed by $\mathcal{I}_v$ and scaling by $\sqrt{n/|\mathcal{I}_v|}$, satisfies

$$\sigma_k(M_v) \geq \sigma_k(M)/2.$$

Proof Follows from Lemma 17 (below) together with a union bound. ∎

Lemma 17 *Assume Condition 4 holds. Consider a random subset $\{J_1, J_2, \ldots, J_d\} \subseteq [n]$ of size d chosen uniformly at random without replacement, and let $\widetilde{M}$ be the random $d \times k$ matrix given by*

$$\widetilde{M} := \sqrt{\frac{n}{d}} \cdot \begin{bmatrix} \vec{e}_{J_1}^{\top} M \\ \vec{e}_{J_2}^{\top} M \\ \vdots \\ \vec{e}_{J_d}^{\top} M \end{bmatrix}.$$

If

$$d \geq \frac{32}{9} \cdot c_M \cdot k \cdot \ln \frac{k}{\delta},$$

then

$$\Pr\left[\sigma_k(\widetilde{M}) \geq \sigma_k(M)/2\right] \geq 1 - \delta.$$

Proof Let $\{I_1, I_2, \ldots, I_d\} \subseteq [n]$ be a random subset of size d chosen uniformly at random with replacement, and let $\widehat{M}$ be the random $d \times k$ matrix given by

$$\widehat{M} := \sqrt{\frac{n}{d}} \cdot \begin{bmatrix} \vec{e}_{I_1}^{\top} M \\ \vec{e}_{I_2}^{\top} M \\ \vdots \\ \vec{e}_{I_d}^{\top} M \end{bmatrix}.$$

By Proposition 18, for any $\tau > 0$,

$$\Pr[\sigma_k(\widetilde{M}) < \tau] \leq \Pr[\sigma_k(\widehat{M}) < \tau].$$

Therefore, henceforth, we just work with $\widehat{M}$ (*i.e.*, sampling with replacement).

Note that

$$\sigma_k(\widehat{M}) = \sqrt{\lambda_{\min}(\widehat{M}^\top \widehat{M})}.$$

For each $j \in [d]$, let $X_j := n \cdot (U^\top \vec{e}_{I_j}) \otimes (U^\top \vec{e}_{I_j})$, so

$$\widehat{M}^\top \widehat{M} = \frac{n}{d} \sum_{j=1}^{d} (M^\top \vec{e}_{I_j}) \otimes (M^\top \vec{e}_{I_j}) = VS\left(\frac{1}{d}\sum_{j=1}^{d} X_j\right) SV^\top$$

and

$$\lambda_{\min}(\widehat{M}^\top \widehat{M}) \geq \lambda_{\min}(S)^2 \cdot \lambda_{\min}\left(\frac{1}{d}\sum_{j=1}^{d} X_j\right) = \sigma_k(M)^2 \cdot \lambda_{\min}\left(\frac{1}{d}\sum_{j=1}^{d} X_j\right).$$

Observe that

$$\mathbb{E}[X_j] = \sum_{i=1}^{n} \Pr[I_j = i] \cdot n \cdot (U^\top \vec{e}_i) \otimes (U^\top \vec{e}_i) = I$$

and that

$$\lambda_{\max}(X_j) \leq n \cdot \max_{i \in [n]} \{\|U^\top \vec{e}_i\|_2^2\} = c_M \cdot k, \quad \text{almost surely.}$$

By Lemma 26 (a Chernoff bound on extremal eigenvalues of random symmetric matrices),

$$\Pr\left[\lambda_{\min}\left(\frac{1}{d}\sum_{j=1}^{d} X_j\right) \leq \frac{1}{4}\right] \leq k \cdot e^{-d(3/4)^2/(2c_M k)} \leq \delta.$$

The claim follows. ■

Proposition 18 (Reduction to sampling with replacement) *Consider any $m \times n$ matrix A. For any $t \in [m]$, let $\widetilde{A}_t$ be a random $t \times n$ submatrix of A formed by choosing a random subset of t rows of A uniformly at random* without replacement*; and let $\widehat{A}_t$ be a random $t \times n$ submatrix of A formed by choosing a random subset of t rows of A uniformly at random* with replacement*. Fix any $t \in [m]$ and $\tau > 0$. Then*

$$\Pr[\sigma_n(\widetilde{A}_t) < \tau] \;\leq\; \Pr[\sigma_n(\widehat{A}_t) < \tau].$$

Proof This argument is similar to one given by Recht (2009). We first prove that $\Pr[\sigma_n(\widetilde{A}_t) < \tau]$ is non-increasing in t. For any $t' \leq t$, consider the following coupling between $\widetilde{A}_t$ and $\widetilde{A}_{t'}$:

1. First, sample t row indices in $[m]$ uniformly at random without replacement, and select those rows in A to form $\widetilde{A}_t$.

2. Then, given these t row indices, choose $t - t'$ of them uniformly at random without replacement, and remove them to form $\widetilde{A}_{t'}$.

Since $\sigma_n(\widetilde{A}_{t'}) \leq \sigma_n(\widetilde{A}_t)$,

$$\sigma_n(\widetilde{A}_t) < \tau \implies \sigma_n(\widetilde{A}_{t'}) < \tau,$$

and consequently

$$\Pr[\sigma_n(\widetilde{A}_t) < \tau] \;\leq\; \Pr[\sigma_n(\widetilde{A}_{t'}) < \tau]. \tag{14}$$

Now we prove the proposition. Let $\mathrm{Unique}_t \in [t]$ be the number of distinct row indices selected to form $\widehat{A}_t$. Then

$$\begin{aligned}
\Pr[\widehat{A}_t < \tau] &= \sum_{i=1}^{t} \Pr[\widehat{A}_t < \tau \,|\, \mathrm{Unique}_t = i] \Pr[\mathrm{Unique}_t = i] \\
&= \sum_{i=1}^{t} \Pr[\widetilde{A}_i < \tau] \Pr[\mathrm{Unique}_t = i] \\
&\geq \sum_{i=1}^{t} \Pr[\widetilde{A}_t < \tau] \Pr[\mathrm{Unique}_t = i] \quad \text{(by (14), as } i \leq t) \\
&= \Pr[\widetilde{A}_t < \tau] \sum_{i=1}^{t} \Pr[\mathrm{Unique}_t = i] \\
&= \Pr[\widetilde{A}_t < \tau].
\end{aligned}$$

■

D.2. Relaxation of Condition 2 using higher-order moments

Even if Condition 2 does not hold (*e.g.*, if $\vec{\mu}_{v,j} \equiv \vec{m} \in \mathbb{R}^d$ (say) for all $v \in [\ell], j \in [k]$ so all of the component distributions have the same mean), one may still apply Algorithm B to the model $(h, \vec{y}_1, \vec{y}_2, \ldots, \vec{y}_\ell)$ where $\vec{y}_v \in \mathbb{R}^{d+d(d+1)/2}$ is the random vector that include both first- and second-order terms of $\vec{x}_v$, *i.e.*, $\vec{y}_v$ is the concatenation of x_v and the upper triangular part of $\vec{x}_v \otimes \vec{x}_v$. In this case, Condition 2 is replaced by a requirement that the matrices

$$M'_v := \left[\ \mathbb{E}[\vec{y}_v|h=1] \;\middle|\; \mathbb{E}[\vec{y}_v|h=2] \;\middle|\; \cdots \;\middle|\; \mathbb{E}[\vec{y}_v|h=k]\ \right] \in \mathbb{R}^{(d+d(d+1)/2)\times k}$$

of conditional means and covariances have full rank. This requirement can be met even if the means $\vec{\mu}_{v,j}$ of the mixture components are all the same. Extending this to higher-order terms is immediate.

D.3. Empirical moments for multi-view mixtures of subgaussian distributions

The required concentration behavior of the empirical moments used by Algorithm B can be easily established for multi-view Gaussian mixture models using known techniques (Chaudhuri et al., 2009). This is clear for the second-order statistics $\hat{P}_{a,b}$ for $\{a,b\} \in \{\{1,2\},\{1,3\}\}$, and remains true for the third-order statistics $\hat{P}_{1,2,3}$ because $\vec{x}_3$ is conditionally independent of $\vec{x}_1$ and $\vec{x}_2$ given h. The magnitude of $\langle \hat{U}_3\vec{\theta}_i, \vec{x}_3\rangle$ can be bounded for all samples (with a union bound; recall that we make the simplifying assumption that $\hat{P}_{1,3}$ is independent of $\hat{P}_{1,2,3}$, and therefore so are $\hat{U}_3$ and $\hat{P}_{1,2,3}$). Therefore, one effectively only needs spectral norm error bounds for second-order statistics, as provided by existing techniques.

Indeed, it is possible to establish Condition 3 in the case where the conditional distribution of $\vec{x}_v$ given h (for each view v) is subgaussian. Specifically, we assume that there exists some $\alpha > 0$ such that for each view v and each component $j \in [k]$,

$$\mathbb{E}\left[\exp\Big(\lambda\langle\vec{u}, \operatorname{cov}(\vec{x}_v|h=j)^{-1/2}(\vec{x}_v - \mathbb{E}[\vec{x}_v|h=j])\rangle\Big)\right] \le \exp(\alpha\lambda^2/2), \quad \forall\lambda\in\mathbb{R}, \vec{u}\in\mathcal{S}^{d-1}$$

where $\operatorname{cov}(\vec{x}|h=j) := \mathbb{E}[(\vec{x}_v - \mathbb{E}[\vec{x}_v|h=j])\otimes(\vec{x}_v - \mathbb{E}[\vec{x}_v|h=j])|h=j]$ is assumed to be positive definite. Using standard techniques (*e.g.*, Vershynin (2012)), Condition 3 can be shown to hold under the above conditions with the following parameters (for some universal constant $c > 0$):

$$\begin{aligned}
w_{\min} &:= \min_{j\in[k]} w_j \\
N_0 &:= c\cdot\frac{\alpha^{3/2}(d+\log(1/\delta))}{w_{\min}}\log\frac{\alpha^{3/2}(d+\log(1/\delta))}{w_{\min}} \\
C_{a,b} &:= c\cdot\Big(\max\Big\{\|\operatorname{cov}(\vec{x}_v|h=j)\|_2^{1/2}, \|\mathbb{E}[\vec{x}_v|h=j]\|_2 : v\in\{a,b\}, j\in[k]\Big\}\Big)^2 \\
C_{1,2,3} &:= c\cdot\Big(\max\Big\{\|\operatorname{cov}(\vec{x}_v|h=j)\|_2^{1/2}, \|\mathbb{E}[\vec{x}_v|h=j]\|_2 : v\in[3], j\in[k]\Big\}\Big)^3 \\
f(N,\delta) &:= \sqrt{\frac{k^2\log(1/\delta)}{N}} + \sqrt{\frac{\alpha^{3/2}\sqrt{\log(N/\delta)}(d+\log(1/\delta))}{w_{\min}N}}.
\end{aligned}$$

D.4. Recovering the component covariances

While Algorithm B recovers just the means of the mixture components, we remark that a slight variation can be used to recover the covariances as well. Note that

$$\mathbb{E}[\vec{x}_v\otimes\vec{x}_v|h] = (M_v\vec{e}_h)\otimes(M_v\vec{e}_h) + \Sigma_{v,h} = \vec{\mu}_{v,h}\otimes\vec{\mu}_{v,h} + \Sigma_{v,h}$$

for all $v\in[\ell]$. For a pair of vectors $\vec{\phi}\in\mathbb{R}^d$ and $\vec{\psi}\in\mathbb{R}^d$, define the matrix $Q_{1,2,3}(\vec{\phi},\vec{\psi})\in\mathbb{R}^{d\times d}$ of fourth-order moments by $Q_{1,2,3}(\vec{\phi},\vec{\psi}) := \mathbb{E}[(\vec{x}_1\otimes\vec{x}_2)\langle\vec{\phi},\vec{x}_3\rangle\langle\vec{\psi},\vec{x}_3\rangle]$.

Proposition 19 *Under the setting of Lemma 5, the matrix given by*

$$F_{1,2,3}(\vec{\phi},\vec{\psi}) := (U_1^\top Q_{1,2,3}(\vec{\phi},\vec{\psi})U_2)(U_1^\top P_{1,2}U_2)^{-1}$$

satisfies $F_{1,2,3}(\vec{\phi},\vec{\psi}) = (U_1^\top M_1)\operatorname{diag}(\langle\vec{\phi},\vec{\mu}_{3,t}\rangle\langle\vec{\psi},\vec{\mu}_{3,t}\rangle + \langle\vec{\phi},\Sigma_{3,t}\vec{\psi}\rangle : t\in[k])(U_1^\top M_1)^{-1}$ *and hence is diagonalizable (in fact, by the same matrices as* $B_{1,2,3}(\vec{\eta})$*).*

Proof As in the proof of Lemma 4, it is easy to show that

$$\begin{aligned}
Q_{1,2,3}(\vec{\phi},\vec{\psi}) &= \mathbb{E}[\mathbb{E}[\vec{x}_1|h]\otimes\mathbb{E}[\vec{x}_2|h]\langle\vec{\phi},\mathbb{E}[\vec{x}_3\otimes\vec{x}_3|h]\vec{\psi}\rangle] \\
&= M_1\mathbb{E}[\vec{e}_h\otimes\vec{e}_h\langle\vec{\phi},(\vec{\mu}_{3,h}\otimes\vec{\mu}_{3,h}+\Sigma_{3,h})\vec{\psi}\rangle]M_2^\top \\
&= M_1\operatorname{diag}(\langle\vec{\phi},\vec{\mu}_{3,t}\rangle\langle\vec{\psi},\vec{\mu}_{3,t}\rangle + \langle\vec{\phi},\Sigma_{3,t}\vec{\psi}\rangle : t\in[k])\operatorname{diag}(\vec{w})M_2^\top.
\end{aligned}$$

The claim then follows from the same arguments used in the proof of Lemma 5. ∎

D.5. Proof of Proposition 8

The conditional independence properties follow from the HMM conditional independence assumptions. To check the parameters, observe first that

$$\Pr[h_1 = i | h_2 = j] = \frac{\Pr[h_2 = j | h_1 = i] \cdot \Pr[h_1 = i]}{\Pr[h_2 = j]} = \frac{T_{j,i}\pi_i}{(T\vec{\pi})_j} = \vec{e}_i \operatorname{diag}(\vec{\pi}) T^\top \operatorname{diag}(T\vec{\pi})^{-1} \vec{e}_j$$

by Bayes' rule. Therefore

$$M_1 \vec{e}_j = \mathbb{E}[\vec{x}_1 | h_2 = j] = O\mathbb{E}[\vec{e}_{h_1} | h_2 = j] = O \operatorname{diag}(\vec{\pi}) T^\top \operatorname{diag}(T\vec{\pi})^{-1} \vec{e}_j.$$

The rest of the parameters are similar to verify.

Appendix E. General results from matrix perturbation theory

The lemmas in this section are standard results from matrix perturbation theory, taken from Stewart and Sun (1990).

Lemma 20 (Weyl's theorem) *Let $A, E \in \mathbb{R}^{m \times n}$ with $m \geq n$ be given. Then*

$$\max_{i \in [n]} |\sigma_i(A + E) - \sigma_i(A)| \leq \|E\|_2.$$

Proof See Theorem 4.11, p. 204 in Stewart and Sun (1990). ∎

Lemma 21 (Wedin's theorem) *Let $A, E \in \mathbb{R}^{m \times n}$ with $m \geq n$ be given. Let A have the singular value decomposition*

$$\begin{bmatrix} U_1^\top \\ U_2^\top \\ U_3^\top \end{bmatrix} A \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{bmatrix}.$$

Let $\tilde{A} := A + E$, with analogous singular value decomposition $(\tilde{U}_1, \tilde{U}_2, \tilde{U}_3, \tilde{\Sigma}_1, \tilde{\Sigma}_2, \tilde{V}_1 \tilde{V}_2)$. Let Φ be the matrix of canonical angles between $\operatorname{range}(U_1)$ and $\operatorname{range}(\tilde{U}_1)$, and Θ be the matrix of canonical angles between $\operatorname{range}(V_1)$ and $\operatorname{range}(\tilde{V}_1)$. If there exists $\delta, \alpha > 0$ such that $\min_i \sigma_i(\tilde{\Sigma}_1) \geq \alpha + \delta$ and $\max_i \sigma_i(\Sigma_2) \leq \alpha$, then

$$\max\{\|\sin\Phi\|_2, \|\sin\Theta\|_2\} \leq \frac{\|E\|_2}{\delta}.$$

Proof See Theorem 4.4, p. 262 in Stewart and Sun (1990). ∎

Lemma 22 (Bauer-Fike theorem) *Let $A, E \in \mathbb{R}^{k \times k}$ be given. If $A = V \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k) V^{-1}$ for some invertible $V \in \mathbb{R}^{k \times k}$, and $\tilde{A} := A + E$ has eigenvalues $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_k$, then*

$$\max_{i \in [k]} \min_{j \in [k]} |\tilde{\lambda}_i - \lambda_j| \leq \|V^{-1} E V\|_2.$$

Proof See Theorem 3.3, p. 192 in Stewart and Sun (1990). ∎

Lemma 23 (Perturbation of inverses) *Let $A, E \in \mathbb{R}^{k \times k}$ be given. If A is invertible, and $\|A^{-1}E\|_2 < 1$, then $\tilde{A} := A + E$ is invertible, and*

$$\|\tilde{A}^{-1} - A^{-1}\|_2 \leq \frac{\|E\|_2 \|A^{-1}\|_2^2}{1 - \|A^{-1}E\|_2}.$$

Proof See Theorem 2.5, p. 118 in Stewart and Sun (1990). ∎

Appendix F. Probability inequalities

Lemma 24 (Accuracy of empirical probabilities) *Fix $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_n) \in \Delta^{m-1}$. Let $\vec{x}$ be a random vector for which $\Pr[\vec{x} = \vec{e}_i] = \mu_i$ for all $i \in [m]$, and let $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ be n independent copies of $\vec{x}$. Set $\hat{\mu} := (1/n) \sum_{i=1}^n \vec{x}_i$. For all $t > 0$,*

$$\Pr\left[\|\hat{\mu} - \vec{\mu}\|_2 > \frac{1 + \sqrt{t}}{\sqrt{n}}\right] \leq e^{-t}.$$

Proof This is a standard application of McDiarmid's inequality (using the fact that $\|\hat{\mu} - \vec{\mu}\|_2$ has $\sqrt{2}/n$ bounded differences when a single $\vec{x}_i$ is changed), together with the bound $\mathbb{E}[\|\hat{\mu} - \vec{\mu}\|_2] \leq 1/\sqrt{n}$. See Proposition 19 in Hsu et al. (2012). ∎

Lemma 25 (Random projection) *Let $\vec{\theta} \in \mathbb{R}^n$ be a random vector distributed uniformly over $\mathcal{S}^{n-1}$, and fix a vector $\vec{v} \in \mathbb{R}^n$.*

1. *If $\beta \in (0, 1)$, then*

$$\Pr\left[|\langle \vec{\theta}, \vec{v} \rangle| \leq \|\vec{v}\|_2 \cdot \frac{1}{\sqrt{n}} \cdot \beta\right] \leq \exp\left(\frac{1}{2}(1 - \beta^2 + \ln \beta^2)\right).$$

2. *If $\beta > 1$, then*

$$\Pr\left[|\langle \vec{\theta}, \vec{v} \rangle| \geq \|\vec{v}\|_2 \cdot \frac{1}{\sqrt{n}} \cdot \beta\right] \leq \exp\left(\frac{1}{2}(1 - \beta^2 + \ln \beta^2)\right).$$

Proof This is a special case of Lemma 2.2 from Dasgupta and Gupta (2003). ∎

Lemma 26 (Matrix Chernoff bound) *Let X be a symmetric random $m \times m$ matrix such that $0 \preceq X \preceq rI$ almost surely, and set $l := \lambda_{\min}(\mathbb{E}[X])$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. copies of X. For any $\epsilon \in [0, 1]$,*

$$\Pr\left[\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \leq (1-\epsilon)\cdot l\right] \leq m \cdot e^{-n\epsilon^2 l/(2r)}.$$

Proof This is a direct corollary of Theorem 19 from Ahlswede and Winter (2002). ∎

Appendix G. Insufficiency of second-order moments

Chang (1996) shows that a simple class of Markov models used in mathematical phylogenetics cannot be identified from pair-wise probabilities alone. Below, we restate (a specialization of) this result in terms of the document topic model from Section 2.1.

Proposition 27 (Chang, 1996) *Consider the model from Section 2.1 on $(h, x_1, x_2, \ldots, x_\ell)$ with parameters M and $\vec{w}$. Let $Q \in \mathbb{R}^{k\times k}$ be an invertible matrix such that the following hold:*

1. *$\vec{1}^\top Q = \vec{1}^\top$;*
2. *MQ^{-1}, $Q\operatorname{diag}(\vec{w})M^\top \operatorname{diag}(M\vec{w})^{-1}$, and $Q\vec{w}$ have non-negative entries;*
3. *$Q\operatorname{diag}(\vec{w})Q^\top$ is a diagonal matrix.*

Then the marginal distribution over (x_1, x_2) is identical to that in the case where the model has parameters $\tilde{M} := MQ^{-1}$ and $\tilde{w} := Q\vec{w}$.

A simple example for $d = k = 2$ can be obtained from

$$M := \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}, \qquad \vec{w} := \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}, \qquad Q := \begin{bmatrix} p & \frac{1+\sqrt{1+4p(1-p)}}{2} \\ 1-p & \frac{1-\sqrt{1+4p(1-p)}}{2} \end{bmatrix}$$

for some $p \in (0, 1)$. We take $p = 0.25$, in which case Q satisfies the conditions of Proposition 27, and

$$M = \begin{bmatrix} 0.25 & 0.75 \\ 0.75 & 0.25 \end{bmatrix}, \qquad \vec{w} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix},$$

$$\tilde{M} = MQ^{-1} \approx \begin{bmatrix} 0.6614 & 0.1129 \\ 0.3386 & 0.8871 \end{bmatrix}, \qquad \tilde{w} = Q\vec{w} \approx \begin{bmatrix} 0.7057 \\ 0.2943 \end{bmatrix}.$$

In this case, both $(M, \vec{w})$ and $(\tilde{M}, \tilde{w})$ give rise to the same pair-wise probabilities

$$M\operatorname{diag}(\vec{w})M^\top = \tilde{M}\operatorname{diag}(\tilde{w})\tilde{M}^\top \approx \begin{bmatrix} 0.3125 & 0.1875 \\ 0.1875 & 0.3125 \end{bmatrix}.$$

However, the triple-wise probabilities, for $\eta = (1, 0)$, differ: for $(M, \vec{w})$, we have

$$M \operatorname{diag}(M^\top \eta) \operatorname{diag}(\vec{w}) M^\top \approx \begin{bmatrix} 0.2188 & 0.0938 \\ 0.0938 & 0.0938 \end{bmatrix};$$

while for $(\tilde{M}, \tilde{w})$, we have

$$\tilde{M} \operatorname{diag}(\tilde{M}^\top \eta) \operatorname{diag}(\tilde{w}) \tilde{M}^\top \approx \begin{bmatrix} 0.2046 & 0.1079 \\ 0.1079 & 0.0796 \end{bmatrix}.$$