A Spectral Algorithm for Latent Dirichlet Allocation*

Animashree Anandkumar[1], Dean P. Foster[3], Daniel Hsu[2],
Sham M. Kakade[2], and Yi-Kai Liu[4]

[1]Department of EECS, University of California, Irvine
[2]Microsoft Research, New England
[3]Department of Statistics, Wharton School, University of Pennsylvania
[4]National Institute of Standards and Technology, Gaithersburg, MD †

Abstract

The problem of topic modeling can be seen as a generalization of the clustering problem, in that it posits that observations are generated due to multiple latent factors (*e.g.*, the words in each document are generated as a mixture of *several* active topics, as opposed to just one). This increased representational power comes at the cost of a more challenging unsupervised learning problem of estimating the topic probability vectors (the distributions over words for each topic), when only the words are observed and the corresponding topics are hidden.

We provide a simple and efficient learning procedure that is guaranteed to recover the parameters for a wide class of mixture models, including the popular latent Dirichlet allocation (LDA) model. For LDA, the procedure correctly recovers both the topic probability vectors and the prior over the topics, using only trigram statistics (*i.e.*, third order moments, which may be estimated with documents containing just three words). The method, termed Excess Correlation Analysis (ECA), is based on a spectral decomposition of low order moments (third and fourth order) via two singular value decompositions (SVDs). Moreover, the algorithm is scalable since the SVD operations are carried out on $k \times k$ matrices, where k is the number of latent factors (e.g. the number of topics), rather than in the d-dimensional observed space (typically $d \gg k$).

1 Introduction

There is general agreement that there are multiple unobserved or latent factors affecting observed data. Mixture models offer a powerful framework to incorporate the effects of these latent variables. A family of mixture models, popularly known as *topic models*, has generated broad interest on both theoretical and practical fronts.

Topic models incorporate latent variables, the topics, to explain the observed co-occurrences of words in documents. They posit that each document has a mixture of active topics (possibly sparse) and that each active topic determines the occurrence of words in the document. Usually, a Dirichlet prior is assigned to the distribution of topics in documents, giving rise to the so-called

*Previous title: "Two SVDs Suffice: Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation".

†Contributions to this work by NIST, an agency of the US government, are not subject to copyright laws.

latent Dirichlet allocation (LDA) (Blei et al., 2003). These models possess a rich representational power since they allow for the words in each document to be generated from more than one topic (*i.e.*, the model permits documents to be about multiple topics). This increased representational power comes at the cost of a more challenging unsupervised estimation problem, when only the words are observed and the corresponding topics are hidden.

In practice, the most common estimation procedures are based on finding maximum likelihood (ML) estimates, through either local search or sampling based methods, *e.g.*, Expectation-Maximization (EM) (Redner and Walker, 1984), Gibbs sampling (Asuncion et al., 2011), and variational approaches (Hoffman et al., 2010). Another body of tools is based on matrix factorization (Hofmann, 1999; Lee and Seung, 1999). For document modeling, typically, the goal is to form a sparse decomposition of a term by document matrix (which represents the word counts in each document) into two parts: one which specifies the active topics in each document and the other which specifies the distributions of words under each topic.

This work provides an alternative approach to parameter recovery based on the method of moments (Lindsay, 1989; Lindsay and Basak, 1993), which attempts to match the observed moments with those posited by the model. Our approach does this efficiently through a spectral decomposition of the observed moments through two singular value decompositions. This method is simple and efficient to implement, based on only low order moments (third or fourth order), and is guaranteed to recover the parameters of a wide class of mixture models, including the LDA model. We exploit exchangeability of the observed variables and, more generally, the availability of multiple views drawn independently from the same hidden component.

1.1 Summary of Contributions

We present an approach known as Excess Correlation Analysis (ECA) based on the knowledge of low order moments between the observed variables, assumed to be exchangeable (or, more generally, drawn from a multi-view mixture model). ECA differs from Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) in that it is based on two singular value decompositions: the first SVD whitens the data (based on the correlation between two variables) and the second SVD utilizes higher order moments (based on third or fourth order) to find directions which exhibit moments that are in *excess* of those suggested by a Gaussian distribution. Both SVDs are performed on matrices of size $k \times k$, where k is the number of latent factors, making the algorithm scalable (typically the dimension of the observed space $d \gg k$).

The method is applicable to a wide class of mixture models including exchangeable and multi-view models. We first consider the class of exchangeable variables with independent latent factors, such as a latent Poisson mixture model (a natural Poisson model for generating the sentences in a document, analogous to LDA's multinomial model for generating the words in a document). We establish that a spectral decomposition, based on third or fourth order central moments, recovers the parameters for this model class. We then consider latent Dirichlet allocation and show that a spectral decomposition of a modified third order moment (exactly) recovers both the probability distributions over words for each topic and the Dirichlet prior. Note that to obtain third order moments, it suffices for documents to contain just 3 words. Finally, we present extensions to multi-view models, where multiple views drawn independently from the same latent factor are available. This includes the case of both pure topic models (where only one active topic is present in each document) and discrete hidden Markov models. For this setting, we establish that ECA correctly recovers the parameters and is simpler than the eigenvector decomposition methods of Anandkumar et al. (2012).

Finally, "plug-in" moment estimates can be used with sampled data. Section 5 provides a sample complexity of the method showing that estimating the third order moments is not as difficult as it might naively seem since we only need a $k \times k$ matrix to be accurate.

Some preliminary experiments that illustrate the efficacy of the proposed algorithm are given in the appendix.

1.2 Related Work

For the case of a single topic per document, the work of Papadimitriou et al. (2000) provides the first guarantees of recovering the topic distributions (*i.e.*, the distributions over words corresponding to each topic), albeit with a rather stringent separation condition (where the words in each topic are essentially non overlapping). Understanding what separation conditions (or lack thereof) permit efficient learning is a natural question; in the clustering literature, a line of work has focussed on understanding the relation between the separation of the mixture components and the complexity of learning. For clustering, the first learnability result (Dasgupta, 1999) was under a somewhat strong separation condition; a subsequent line of results relaxed (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) or removed these conditions (Kalai et al., 2010; Belkin and Sinha, 2010; Moitra and Valiant, 2010); roughly speaking, the less stringent the separation condition assumed, the more difficult the learning problem is, both computationally and statistically. For the topic modeling problem in which only a single topic is present per document, Anandkumar et al. (2012) provides an algorithm for learning topics with no separation (only a certain full rank assumption is utilized).

For the case of latent Dirichlet allocation (where multiple topics are present in each document), the recent work of Arora et al. (2012) provides the first provable result under a certain natural separation condition. The notion of separation utilized is based on the existence of "anchor words" for topics — essentially , each topic contains words that appear (with reasonable probability) only in that topic (this is a milder assumption than that in Papadimitriou et al. (2000)). Under this assumption, Arora et al. (2012) provide the first provably correct algorithm for learning the topic distributions. Their work also justifies the use of non-negative matrix (NMF) as a procedure for this problem (the original motivation for NMF was as a topic modeling algorithm, though, prior to this work, formal guarantees as such were rather limited). Furthermore, Arora et al. (2012) provides results for certain correlated topic models.

Our approach makes further progress on this problem by providing an algorithm which requires no separation condition. The underlying approach we take is a certain diagonalization technique of the observed moments. We know of at least three different settings which utilize this idea for parameter estimation.

Chang (1996) utilizes eigenvector methods for discrete Markov models of evolution, where the models involve multinomial distributions. The idea has been extended to other discrete mixture models such as discrete hidden Markov models (HMMs) and mixture models with single active topics (see Mossel and Roch (2006); Hsu et al. (2009); Anandkumar et al. (2012)). A key idea in Chang (1996) is the ability to handle multinomial distributions, which comes at the cost of being able to handle only certain single latent factor/topic models (where the latent factor is in only one of k states, such as in HMMs). For these single topic models, the work in Anandkumar et al. (2012) shows how this method is quite general in that the noise model is essentially irrelevant, making it applicable to both discrete models like HMMs and certain Gaussian mixture models.

The second setting is the body of algebraic methods used for the problem of blind source separation (Cardoso and Comon, 1996). These approaches rely on tensor decomposition approaches (see Comon and Jutten (2010)) tailored to independent source separation with additive noise (usually Gaussian). Much of literature focuses on understanding the effects of measurement noise (without assuming knowledge of their statistics) on the tensor decomposition, which often requires more sophisticated algebraic tools.

Frieze et al. (1996) also utilize these ideas for learning the columns of a linear transformation (in a noiseless setting). This work provides a different efficient algorithm, based on a certain ascent algorithm (rather than joint diagonalization approach, as in (Cardoso and Comon, 1996)).

The underlying insight that our method exploits is that we have exchangeable (or multi-view) variables, *e.g.*, we have multiple words (or sentences) in a document, which are drawn independently from the same hidden state. This allows us to borrow from both the ideas in Chang (1996) and in Cardoso and Comon (1996). In particular, we show that the "topic" modeling problem exhibits a rather simple algebraic solution, where only two SVDs suffice for parameter estimation. Moreover, this approach also simplifies the algorithms in Mossel and Roch (2006); Hsu et al. (2009); Anandkumar et al. (2012), in that the eigenvector methods are no longer necessary (*e.g.*, the approach leads to methods for parameter estimation in HMMs with only two SVDs rather than using eigenvector approaches, as in previous work).

Furthermore, the exchangeability assumption permits us to have *arbitrary* noise models (rather than additive Gaussian noise, which are not appropriate for multinomial and other discrete distributions). A key technical contribution is that we show how the basic diagonalization approach can be adapted to Dirichlet models, through a rather careful construction. This construction bridges the gap between the single topic models (as in Chang (1996); Anandkumar et al. (2012)) and the independent factor model.

More generally, the multi-view approach has been exploited in previous works for semi-supervised learning and for learning mixtures of well-separated distributions (*e.g.*, as in Ando and Zhang (2007); Kakade and Foster (2007); Chaudhuri and Rao (2008); Chaudhuri et al. (2009)). These previous works essentially use variants of canonical correlation analysis (Hotelling, 1935) between two views. This work shows that having a third view of the data permits rather simple estimation procedures with guaranteed parameter recovery.

2 The Exchangeable and Multi-view Models

We have a random vector $h = (h_1, h_2, \ldots, h_k)^\top \in \mathbb{R}^k$. This vector specifies the latent factors (*i.e.*, the hidden state), where h_i specifies the value taken by i-th factor. Denote the variance of h_i as

$$\sigma_i^2 = \mathbb{E}[(h_i - \mathbb{E}[h_i])^2]$$

which we assume to be strictly positive, for each i, and denote the higher l-th central moments of h_i as:

$$\mu_{i,l} \quad := \quad \mathbb{E}[(h_i - \mathbb{E}[h_i])^l]$$

At most, we only use the first four moments in our analysis.

Suppose we also have a sequence of *exchangeable* random vectors $\{x_1, x_2, x_3, x_4, \dots\} \in \mathbb{R}^d$; these are considered to be the observed variables. Assume throughout that $d \geq k$; that $x_1, x_2, x_3, x_4, \dots \in \mathbb{R}^d$ are conditionally independent given h; and there exists a matrix $O \in \mathbb{R}^{d \times k}$ such that

$$\mathbb{E}[x_v|h] = Oh$$

for each $v \in \{1, 2, 3, 4, \dots\}$. Throughout, we make the following assumption.

Assumption 2.1. *O is full rank.*

This is a mild assumption, which allows for identifiability of the columns of O. The goal is to estimate the matrix O, sometimes referred to as the topic matrix.

Importantly, we make no assumptions on the noise model. In particular, we do not assume that the noise is additive (or that the noise is independent of h).

2.1 Independent Latent Factors

Here, suppose that h has a product distribution, *i.e.*, each component of h_i is independent from the rest. Two important examples of this setting are as follows:

(Multiple) mixtures of Gaussians: Suppose $x_v = Oh + \eta$, where η is Gaussian noise and h is a binary vector (under a product distribution). Here, the i-th column O_i can be considered to be the mean of the i-th Gaussian component. This is somewhat different model than the classic mixture of k-Gaussians, as the model now permits any number of Gaussians to be responsible for generating the hidden state (*i.e.*, h is permitted to be any of the 2^k vectors on the hypercube, while in the classic mixture problem, only one component is responsible. However, this model imposes the independent factor constraint.). We may also allow η to be heteroskedastic (*i.e.*, the noise may depend on h, provided the linearity assumption $\mathbb{E}[x_v|h] = Oh$ holds.)

(Multiple) mixtures of Poissons: Suppose $[Oh]_j$ specifies the Poisson rate of counts for $[x_v]_j$. For example, x_v could be a vector of word counts in the v-th sentence of a document (where $x_1, x_2, \dots$ are words counts of a sequence sentences). Here, O would be a matrix with positive entries, and h_i would scale the rate at which topic i generates words in a sentence (as specified by the i-th column of O). The linearity assumption is satisfied as $\mathbb{E}[x_v|h] = Oh$ (note the noise is not additive in this case). Here, multiple topics may be responsible for generating the words in each sentence. This model provides a natural variant of LDA, where the distribution over h is a product distribution (while in LDA, h is a probability vector).

2.2 The Dirichlet Model

Now suppose the hidden state h is a distribution itself, with a density specified by the Dirichlet distribution with parameter $\alpha \in \mathbb{R}^k_+$ (α is a strictly positive real vector). We often think of h as a distribution over topics. Precisely, the density of $h \in \Delta^{k-1}$ (where the probability simplex Δ^{k-1} denotes the set of possible distributions over k outcomes) is specified by:

$$p_\alpha(h) := \frac{1}{Z(\alpha)} \prod_{i=1}^{k} h_i^{\alpha_i - 1}$$

where

$$Z(\alpha) := \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$$

and

$$\alpha_0 := \alpha_1 + \alpha_2 + \cdots + \alpha_k .$$

Intuitively, α_0 (the sum of the "pseudo-counts") is a crude measure of the uniformity of the distribution. As $\alpha_0 \to 0$, the distribution degenerates to one over pure topics (*i.e.*, the limiting density is one in which, with probability 1, precisely one coordinate of h is 1 and the rest are 0).

Latent Dirichlet Allocation: LDA makes the further assumption that each random variable $x_1, x_2, x_3, \ldots$ takes on discrete values out of d outcomes (*e.g.*, x_v represents what the v-th word in a document is, so d represents the number of words in the language). Each column of O represents a distribution over the outcomes (*e.g.*, these are the topic probabilities). The sampling procedure is specified as follows: First, h is sampled according to the Dirichlet distribution. Then, for each v, independently sample $i \in \{1, 2, \ldots k\}$ according to h, and, finally, sample x_v according to the i-th column of O. Observe this model falls into our setting: represent x_v with a "hot" encoding where $[x_v]_j = 1$ if and only if the v-th outcome is the j-th word in the vocabulary. Hence, $\Pr([x_v]_j = 1|h) = [Oh]_j$ and $\mathbb{E}[x_v|h] = Oh$. (Again, the noise model is not additive).

2.3 The Multi-View Model

The multi-view setting can be considered an extension of the exchangeable model. Here, the random vectors $\{x_1, x_2, x_3, \ldots\}$ are of dimensions $d_1, d_2, d_3, \ldots$. Instead of a single O matrix, suppose for each $v \in \{1, 2, 3, \ldots\}$ there exists an $O_v \in \mathbb{R}^{d_v \times k}$ such that

$$\mathbb{E}[x_v|h] = O_v h$$

Throughout, we make the following assumption.

Assumption 2.2. *O_v is full rank for each v.*

Even though the variables are no longer exchangeable, the setting shares much of the statistical structure as the exchangeable one; furthermore, it allows for significantly richer models. For example, Anandkumar et al. (2012) consider a special case of this multi-view model (where there is only one topic present in h) for the purposes of learning hidden Markov models.

A simple factorial HMM: Here, suppose we have a time series of random hidden vectors $h_1, h_2, h_3, \ldots$ and observations $x_1, x_2, x_3, \ldots$ (we slightly abuse notation as h_1 is a vector). Assume that each factor $[h_t]_i \in \{-1, 1\}$. The model parameters and evolution are specified as follows: We have an initial (product) distribution over the first h_1. The "factorial" assumption we make is that each factor $[h_t]_i$ evolves independently; in particular, for each component i, there are (time independent) transition probabilities $p_{i,1\to-1}$ and $p_{i,1\to-1}$. Also suppose that $\mathbb{E}[x_t|h_t] = Oh_t$ (where, again, O does not depend on the time).

To learn this model, consider the first three observations x_1, x_2, x_3. We can embed this three timestep model into the multiview model using a single hidden state, namely h_2, and, with an appropriate construction (of O_1, O_2, O_3 and means shifts of x_v to make the linearity assumption hold). Furthermore, if we recover O_1, O_2, O_3 we can recover O and the transition model. See Anandkumar et al. (2012) for further discussion of this idea (for the single topic case).

3 Identifiability

The underlying question here is: what may we hope to recover about O with only knowledge of the distribution on $x_1, x_2, x_3, \ldots$. At best, we could only recover the columns of O up to permutation.

At the other extreme, suppose no a priori knowledge of the distribution of h is assumed (*e.g.*, it may not even be a product distribution). Here, at best, we can only recover the range of O. In particular, suppose h is distributed according to a multivariate Gaussian, then clearly the columns of O are not identifiable. To see this, transform O to OM (where M is any $k \times k$ invertible matrix) and transform the distribution on h (by M^{-1}); after this transformation, the distribution over x_v is unaltered and the distribution on h is still a multivariate Gaussian. Hence, O and OM are indistinguishable from any observable statistics. (These issues are well understood in setting of independent source separation, for additive noise models without exchangeable variables. See Comon and Jutten (2010)).

Thus, for the columns of O to be identifiable, the distribution on h must have some non-Gaussian statistical properties. We consider three cases. In the independent factor model, we consider the cases when h is skewed and when h has excess kurtosis. We also consider the case that h is Dirichlet distributed.

4 Excess Correlation Analysis (ECA)

We now present exact and efficient algorithms for recovering O. The algorithm is based on two singular value decompositions: the first SVD whitens the data (based on the correlation between two variables) and the second SVD is carried out on higher order moments (based on third or fourth order). We start with the case of independent factors, as these algorithms make the basic diagonalization approach clear.

As discussed in the Introduction, these approaches can been seen as extensions of the methodologies in Chang (1996); Cardoso and Comon (1996). Furthermore, as we shall see, the Dirichlet distribution bridges between the single topic models (as in Chang (1996); Anandkumar et al. (2012)) and the independent factor model.

Throughout, we use A^+ to denote the pseudo-inverse:

$$A^+ = (A^\top A)^{-1} A^\top \tag{1}$$

for a matrix A with linearly independent columns (this allows us to appropriately invert non-square matrices).

4.1 Independent and Skewed Latent Factors

Denote the pairwise and threeway correlations as:

$$\begin{aligned} \mu &:= \mathbb{E}[x_1] \\ \text{Pairs} &:= \mathbb{E}[(x_1 - \mu)(x_2 - \mu)^\top] \\ \text{Triples} &:= \mathbb{E}[(x_1 - \mu) \otimes (x_2 - \mu) \otimes (x_3 - \mu)] \end{aligned}$$

The dimensions of Pairs and Triples are d^2 and d^3, respectively. It is convenient to project Triples to a matrix as follows:

$$\text{Triples}(\eta) := \mathbb{E}[(x_1 - \mu)(x_2 - \mu)^\top \langle \eta, x_3 - \mu \rangle]$$

Algorithm 1 ECA, with skewed factors

Input: vector $\theta \in \mathbb{R}^k$; the moments Pairs and Triples(η)

1. **Dimensionality Reduction:** Find a matrix $U \in \mathbb{R}^{d \times k}$ such that

$$\text{Range}(U) = \text{Range}(\text{Pairs}).$$

(See Remark 1 for a fast procedure.)

2. **Whiten:** Find $V \in \mathbb{R}^{k \times k}$ so $V^\top (U^\top \,\text{Pairs}\, U) V$ is the $k \times k$ identity matrix. Set:

$$W = UV$$

3. **SVD:** Let Λ be the set of (left) singular vectors, with *unique* singular values, of

$$W^\top \,\text{Triples}(W\theta) W$$

4. **Reconstruct:** Return the set $\widehat{O}$:

$$\widehat{O} = \{\ (W^+)^\top \lambda \ : \lambda \in \Lambda\}$$

where W^+ is the pseudo-inverse (see Eq 1).

Roughly speaking, we can think of Triples(η) as a reweighing of a cross covariance (by $\langle \eta, x_3 - \mu \rangle$).

In addition to O not being identifiable up to permutation, the scale of each column of O is also not identifiable. To see this, observe the model over x_i is unaltered if we both rescale any column O_i and appropriately rescale the variable h_i. Without further assumptions, we can only hope to recover a certain canonical form of O, defined as follows:

Definition 1 (The Canonical O)**.** We say O is in a canonical form if, for each i, $\sigma_i^2 = 1$. In particular, the transformation $O \leftarrow O\,\text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$ (and a rescaling of h) places O in canonical form, and the distribution over $x_1, x_2, x_3, \ldots$ is unaltered. Observe the canonical O is only specified up to the sign of each column (any sign change of a column does not alter the variance of h_i).

Recall $\mu_{i,3}$ is the central third moment. Denote the skewness of h_i as:

$$\gamma_i = \frac{\mu_{i,3}}{\sigma_i^3}$$

The first result considers the case when the skewness is non-zero.

Theorem 4.1 (Independent and skewed factors)**.** *We have that:*

- *(No False Positives) For all $\theta \in \mathbb{R}^k$, Algorithm 1 returns a subset of the columns of O, in a canonical form.*
- *(Exact Recovery) Assume γ_i is nonzero for each i. Suppose $\theta \in \mathbb{R}^k$ is a random vector uniformly sampled over the sphere $\mathcal{S}^{k-1}$. With probability 1, Algorithm 1 returns all columns of O, in a canonical form.*

The proof of this theorem is a consequence of the following lemma:

Lemma 4.1. *We have:*

$$\begin{aligned}\text{Pairs} &= O\,\text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)O^\top \\ \text{Triples}(\eta) &= O\,\text{diag}(O^\top \eta)\,\text{diag}(\mu_{1,3}, \mu_{2,3}, \ldots, \mu_{k,3})O^\top\end{aligned}$$

The proof of this Lemma is provided in the Appendix.

Proof of Theorem 4.1. The analysis is with respect to O it its canonical form. By the full rank assumption, $U^\top \text{Pairs}\, U$, which is a $k \times k$ matrix, is full rank; hence, the whitening step is possible. By construction:

$$\begin{aligned}\text{I} &= W^\top \text{Pairs}\, W \\ &= W^\top O\,\text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)O^\top W \\ &= (W^\top O)(W^\top O)^\top \\ &:= MM^\top\end{aligned}$$

where $M := W^\top O$. Hence, M is a $k \times k$ orthogonal matrix.

Observe:

$$\begin{aligned}W^\top \text{Triples}(W\theta)W &= W^\top O\,\text{diag}(O^\top W\theta)\,\text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_k)O^\top W \\ &= M\,\text{diag}(M^\top \theta)\,\text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_k)M^\top\end{aligned}$$

Since M is an orthogonal matrix, the above is a (not necessarily unique) singular value decomposition of $W^\top \text{Triples}(W\theta)W$. Denote the standard basis as $e_1, e_2, \ldots e_k$. Observe that $Me_1, \ldots Me_k$ are singular vectors. In other words, $W^\top O_1, \ldots W^\top O_k$ are singular vectors, where O_i is the i-th column of O.

An SVD uniquely determines all singular vectors (up to sign) which have unique singular values. The diagonal of the matrix $\text{diag}(M^\top\theta)\,\text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_k)$ is the vector $\text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_k)M^\top\theta$. Also, since M is a rotation matrix, the distribution of $M\theta$ is also uniform on the sphere. Thus, if θ is uniformly sampled over the sphere, then every singular value will be nonzero (and distinct) with probability 1. Finally, for the reconstruction, we have

$$W(W^\top W)^{-1}Me_i = W(W^\top W)^{-1}W^\top O_i = O_i,$$

since $W(W^\top W)^{-1}W^\top$ is a projection operator (and the range of W and O are identical). □

Remark 1 (Finding Range(Pairs) efficiently). Suppose $\Theta \in \mathbb{R}^{d\times k}$ is a random matrix with entries sampled independently from a standard normal. Set $U = \text{Pairs}\,\Theta$. Then, with probability 1, $\text{Range}(U) = \text{Range}(\text{Pairs})$.

Remark 2 (No false positives). Note that if the skewness is 0 for some i then ECA will not recover the corresponding column. However, the algorithm does succeed for those directions in which the skewness is non-zero. This guarantee also provides the practical freedom to run the algorithm with multiple different directions θ, since we need only to find unique singular vectors (which may be easier to determine by running the algorithm with different choices for θ).

Algorithm 2 ECA; with kurtotic factors

Input: vectors $\theta, \theta' \in \mathbb{R}^k$; the moments Pairs and Quadruples(η, η')

1. **Dimensionality Reduction:** Find a matrix $U \in \mathbb{R}^{d\times k}$ such that

$$\text{Range}(U) = \text{Range}(\text{Pairs}).$$

2. **Whiten:** Find $V \in \mathbb{R}^{k\times k}$ so $V^\top(U^\top \,\text{Pairs}\, U)V$ is the $k \times k$ identity matrix. Set:

$$W = UV$$

3. **SVD:** Let Λ be the set of (left) singular vectors, with *unique* singular values, of

$$W^\top \,\text{Quadruples}(W\theta, W\theta')W$$

4. **Reconstruct:** Return the set $\widehat{O}$:

$$\widehat{O} = \{\ (W^+)^\top \lambda\ \ : \lambda \in \Lambda\}$$

where W^+ is the pseudo-inverse (see Eq 1).

Remark 3 (Estimating the skewness). It is straight forward to estimate the skewness corresponding to any column of O. Suppose λ is some unique singular vector (up to sign) found in step 3 of ECA (which was used to construct some column O_i), then:

$$\gamma_i = \lambda^\top W^\top \,\text{Triples}(W\lambda)W\lambda$$

is the corresponding skewness for O_i. This follows from the proof, since λ corresponds to some singular vector Me_i and:

$$(Me_i)^\top M \,\text{diag}(M^\top Me_i)\,\text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_k)M^\top Me_i = \gamma_i$$

using that M is an orthogonal matrix.

4.2 Independent and Kurtotic Latent Factors

Define the following matrix:

$$\begin{aligned}\text{Quadruples}(\eta, \eta') := \mathbb{E}[(x_1 - \mu)(x_2 - \mu)^\top \langle \eta, x_3 - \mu\rangle\langle \eta', x_4 - \mu\rangle] \\ - (\eta^\top \,\text{Pairs}\, \eta')\,\text{Pairs} - (\text{Pairs}\,\eta)(\text{Pairs}\,\eta')^\top - (\text{Pairs}\,\eta')(\text{Pairs}\,\eta)^\top\end{aligned}$$

This is a subspace of the fourth moment tensor.

Recall $\mu_{i,4}$ is the central fourth moment. Denote the excess kurtosis of h_i as:

$$\kappa_i = \frac{\mu_{i,4}}{\sigma_i^4} - 3$$

For Gaussian distributions, recall the kurtosis is 3, and so the excess kurtosis is 0. This function is also common in the source separation approaches (Hyvärinen et al., 2001) [1].

In settings where the latent factors are not skewed, we may hope that they are differentiated from a Gaussian distribution due to their fourth order moments. Here, Algorithm 2 is applicable:

Theorem 4.2 (Independent and kurtotic factors)**.** *We have that:*

- *(No False Positives) For all $\theta, \theta' \in \mathbb{R}^k$, Algorithm 2 returns a subset of the columns of O, in a canonical form.*
- *(Exact Recovery) Assume κ_i is nonzero for each i. Suppose $\theta, \theta' \in \mathbb{R}^k$ are random vectors uniformly and independently sampled over the sphere $\mathcal{S}^{k-1}$. With probability 1, Algorithm 2 returns all the columns of O, in a canonical form.*

Remark 4 (Using both skewed and kurtotic ECA). Note that both algorithms never incorrectly return columns. Hence, if for every i, *either* the skewness or the excess kurtosis is nonzero, then by running both algorithms we will recover O.

The proof of this theorem is a consequence of the following lemma:

Lemma 4.2. *We have:*

$$\text{Quadruples}(\eta, \eta') = O \operatorname{diag}(O^\top \eta) \operatorname{diag}(O^\top \eta') \operatorname{diag}(\mu_{1,4} - 3\sigma_1^4, \mu_{2,4} - 3\sigma_2^4, \ldots, \mu_{k,4} - 3\sigma_k^4) O^\top$$

The proof of this Lemma is provided in the Appendix.

Proof of Theorem 4.2. The distinction from the argument in Theorem 4.1 is that:

$$\begin{aligned} W^\top \text{Quadruples}(W\theta, W\theta'))W &= W^\top O \operatorname{diag}(O^\top W\theta) \operatorname{diag}(O^\top W\theta') \operatorname{diag}(\kappa_1, \kappa_2, \ldots, \kappa_k) O^\top W \\ &= M \operatorname{diag}(M^\top \theta) \operatorname{diag}(M^\top \theta') \operatorname{diag}(\kappa_1, \kappa_2, \ldots, \kappa_k) M^\top \end{aligned}$$

The remainder of the argument follows that of the proof of Theorem 4.1. □

4.3 Latent Dirichlet Allocation

Now let us turn to the case where h has a Dirichlet density, where, each h_i is not sampled independently. Even though the distribution on h is the product of $h_i^{\alpha_1 - 1}, \ldots h_i^{\alpha_k - 1}$, the h_i's are not independent due to the constraint that h lives on the simplex. These dependencies suggest a modification for the moments to be used in ECA, which we now provide.

Suppose α_0 is known. Recall that $\alpha_0 := \alpha_1 + \alpha_2 + \cdots + \alpha_k$ (the sum of the "pseudo-counts"). Knowledge of α_0 is significantly weaker than having full knowledge of the entire parameter vector α. A common practice is to specify the entire parameter vector α in a homogeneous manner, with each component being identical (see Steyvers and Griffiths (2006)). Here, we need only specify the sum, which allows for arbitrary inhomogeneity in the prior.

[1]Their algebraic method require more effort due to the additive noise and the lack of exchangeability. Here, the exchangeability assumption simplifies the approach and allows us to address models with non-additive noise (as in the Poisson count model discussed in the Section 2.

Denote the mean as

$$\mu = \mathbb{E}[x_1]$$

Define a modified second moment as

$$\text{Pairs}_{\alpha_0} := \mathbb{E}[x_1 x_2^\top] - \frac{\alpha_0}{\alpha_0 + 1} \mu\mu^\top$$

and a modified third moment as

$$\text{Triples}_{\alpha_0}(\eta) := \mathbb{E}[x_1 x_2^\top \langle \eta, x_3 \rangle] - \frac{\alpha_0}{\alpha_0 + 2} \Big(\mathbb{E}[x_1 x_2^\top] \eta \mu^\top + \mu \eta^\top \mathbb{E}[x_1 x_2^\top] + \langle \eta, \mu \rangle \mathbb{E}[x_1 x_2^\top] \Big) + \frac{2\alpha_0^2}{(\alpha_0 + 2)(\alpha_0 + 1)} \langle \eta, \mu \rangle \mu\mu^\top$$

Remark 5 (Central vs Non-Central Moments). In the limit as $\alpha_0 \to 0$, the Dirichlet model degenerates so that, with probability 1, only one coordinate of h equals 1 and the rest are 0 (*e.g.*, each document is about 1 topic). Here, we limit to non-central moments:

$$\lim_{\alpha_0 \to 0} \text{Pairs}_{\alpha_0} = \mathbb{E}[x_1 x_2^\top] \quad \lim_{\alpha_0 \to 0} \text{Triples}_{\alpha_0}(\eta) = \mathbb{E}[x_1 x_2^\top \langle \eta, x_3 \rangle]$$

In the other extreme, the behavior limits to the central moments:

$$\lim_{\alpha_0 \to \infty} \text{Pairs}_{\alpha_0} = \mathbb{E}[(x_1 - \mu)(x_2 - \mu)^\top] \quad \lim_{\alpha_0 \to \infty} \text{Triples}_{\alpha_0}(\eta) = \mathbb{E}[(x_1 - \mu)(x_2 - \mu)^\top \langle \eta, (x_3 - \mu) \rangle]$$

(to prove the latter claim, expand the central moment and use that, by exchangeability, $\mathbb{E}[x_1 x_2^\top] = \mathbb{E}[x_2 x_3^\top] = \mathbb{E}[x_1 x_3^\top]$).

Our main result here shows that ECA recovers both the topic matrix O, up to a permutation of the columns (where each column represents a probability distribution over words for a given topic) and the parameter vector α, using only knowledge of α_0 (which, as discussed earlier, is a significantly less restrictive assumption than tuning the entire parameter vector). Also, as discussed in Remark 8, the method applies to cases where x_v is not a multinomial distribution.

Theorem 4.3 (Latent Dirichlet Allocation)**.** *We have that:*

- *(No False Positives) For all $\theta \in \mathbb{R}^k$, Algorithm 3 returns a subset of the columns of O.*
- *(Topic Recovery) Suppose $\theta \in \mathbb{R}^k$ is a random vector uniformly sampled over the sphere $\mathcal{S}^{k-1}$. With probability 1, Algorithm 3 returns all columns of O.*
- *(Parameter Recovery) We have that:*

  $$\alpha = \alpha_0(\alpha_0 + 1) O^+ \,\text{Pairs}_{\alpha_0} (O^+)^\top \vec{1}$$

  where $\vec{1} \in \mathbb{R}^k$ is a vector of all ones.

The proof is a consequence of the following lemma:

Lemma 4.3. *We have:*

$$\text{Pairs}_{\alpha_0} = \frac{1}{(\alpha_0 + 1)\alpha_0} O \operatorname{diag}(\alpha) O^\top$$

and

$$\text{Triples}_{\alpha_0}(\eta) = \frac{2}{(\alpha_0 + 2)(\alpha_0 + 1)\alpha_0} O \operatorname{diag}(O^\top \eta) \operatorname{diag}(\alpha) O^\top$$

Algorithm 3 ECA for latent Dirichlet allocation

Input: a vector $\theta \in \mathbb{R}^k$; the moments Pairs_{α_0} and $\text{Triples}_{\alpha_0}$

1. **Dimensionality Reduction:** Find a matrix $U \in \mathbb{R}^{d \times k}$ such that

$$\text{Range}(U) = \text{Range}(\text{Pairs}_{\alpha_0}).$$

   (See Remark 1 for a fast procedure.)

2. **Whiten:** Find $V \in \mathbb{R}^{k \times k}$ so $V^\top (U^\top \text{Pairs}_{\alpha_0} U) V$ is the $k \times k$ identity matrix. Set:

$$W = UV$$

3. **SVD:** Let Λ be the set of (left) singular vectors, with *unique* singular values, of

$$W^\top \text{Triples}_{\alpha_0}(W\theta) W$$

4. **Reconstruct and Normalize:** Return the set $\widehat{O}$:

$$\widehat{O} = \left\{ \frac{(W^+)^\top \lambda}{\vec{1}^\top (W^+)^\top \lambda} \ : \lambda \in \Lambda \right\}$$

   where $\vec{1} \in \mathbb{R}^d$ is a vector of all ones and W^+ is the pseudo-inverse (see Eq 1).

The proof of this Lemma is provided in the Appendix.

Proof of Theorem 4.3. Note that with the following rescaling of columns:

$$\tilde{O} = \frac{1}{\sqrt{(\alpha_0 + 1)\alpha_0}} O \, \text{diag}(\sqrt{\alpha_1}, \sqrt{\alpha_2}, \ldots, \sqrt{\alpha_k})$$

we have that h is in canonical form (*i.e.*, the variance of each h_i is 1). The remainder of the proof is identical to that of Theorem 4.1. The only modification is that we simply normalize the output of Algorithm 1. Finally, observe that claim for estimating α holds due to the functional form of Pairs_{α_0}. □

Remark 6 (Limiting behaviors). ECA seamlessly blends between the single topic model ($\alpha_0 \to 0$) of Anandkumar et al. (2012) and the skewness based ECA, Algorithm 1 ($\alpha_0 \to \infty$). In the single topic case, Anandkumar et al. (2012) provide eigenvector based algorithms. This work shows that two SVDs suffice for parameter recovery.

Remark 7 (Skewed and Kurtotic ECA for LDA). We conjecture that the fourth moments can be utilized in the Dirichlet case such that the resulting algorithm limits to the kurtotic based ECA, when $\alpha_0 \to \infty$. Furthermore, the mixture of Poissions model discussed in Section 2 provides a natural alternative to the LDA model in this regime.

Remark 8 (The Dirichlet model, more generally). It is not necessary that we have a multinomial distribution on x_v, so long as $\mathbb{E}[x_v|h] = Oh$. In some applications, it might be natural for the

Algorithm 4 ECA; the multi-view case

Input: vector $\theta \in \mathbb{R}^k$; the moments $\mathrm{Pairs}_{v,v'}$ and $\mathrm{Triples}_{132}(\eta)$

1. **Project views** 1 **and** 2**:** Find matrices $A \in \mathbb{R}^{k\times d_1}$ and $B \in \mathbb{R}^{k\times d_2}$ such that $A\,\mathrm{Pairs}_{12}\,B^\top$ is invertible. Set:

$$\begin{aligned}
\widetilde{\mathrm{Pairs}}_{12} &:= A\,\mathrm{Pairs}_{12}\,B^\top \\
\widetilde{\mathrm{Pairs}}_{31} &:= \mathrm{Pairs}_{31}\,A^\top \\
\widetilde{\mathrm{Pairs}}_{32} &:= \mathrm{Pairs}_{32}\,B^\top \\
\widetilde{\mathrm{Triples}}_{132}(\eta) &:= A\,\mathrm{Triples}_{132}(\eta)B^\top
\end{aligned}$$

(See Remark 10 for a fast procedure.)

2. **Symmetrize:** Reduce to a single view:

$$\begin{aligned}
\mathrm{Pairs}_3 &:= \widetilde{\mathrm{Pairs}}_{31}(\widetilde{\mathrm{Pairs}}_{12}^\top)^{-1}\widetilde{\mathrm{Pairs}}_{23} \\
\mathrm{Triples}_3(\eta) &:= \widetilde{\mathrm{Pairs}}_{32}(\widetilde{\mathrm{Pairs}}_{12})^{-1}\widetilde{\mathrm{Triples}}_{132}(\eta)(\widetilde{\mathrm{Pairs}}_{12})^{-1}\widetilde{\mathrm{Pairs}}_{13}
\end{aligned}$$

3. **Estimate** O_3 **with ECA:** Call Algorithm 1, with θ, Pairs_3, and $\mathrm{Triples}_3(\eta)$.

---

observations to come from a different distribution (say x_v may represent pixel intensities in an image or some other real valued quantity). For this case, where h has a Dirichlet prior (and where x_v may not be multinomial), ECA still correctly recovers the columns of O. Furthermore, we need not normalize; the set $\{(W^+)^\top\lambda : \lambda \in \Lambda\}$ recovers O in a canonical form.

4.4 The Multi-View Extension

Rather than O being identical for each x_v, suppose for each $v \in \{1, 2, 3, 4, \dots\}$ there exists an $O_v \in \mathbb{R}^{d_v\times k}$ such that

$$\mathbb{E}[x_v|h] = O_v h$$

For $v \in \{1, 2, 3\}$, define

$$\begin{aligned}
\mathrm{Pairs}_{v,v'} &:= \mathbb{E}[(x_v-\mu)(x'_v-\mu)^\top] \\
\mathrm{Triples}_{132}(\eta) &:= \mathbb{E}[(x_1-\mu)(x_2-\mu)^\top\langle\eta, x_3-\mu\rangle]
\end{aligned}$$

We use the notation 132 to stress that $\mathrm{Triples}_{132}(\eta)$ is a $d_1 \times d_2$ sized matrix.

Lemma 4.4. *For* $v \in \{1, 2, 3\}$*,*

$$\begin{aligned}
\mathrm{Pairs}_{v,v'} &= O_v\,\mathrm{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2)O_v'^\top \\
\mathrm{Triples}_{132}(\eta) &= O_1\,\mathrm{diag}(O_3^\top\eta)\,\mathrm{diag}(\mu_{1,3}, \mu_{2,3}, \dots, \mu_{k,3})O_2^\top
\end{aligned}$$

The proof for Lemma 4.4 is analogous to those in Appendix A.

These functional forms make deriving an SVD based algorithm more subtle. Using the methods in Anandkumar et al. (2012), eigenvector based method are straightforward to derive. However, SVD based algorithms are preferred due to their greater simplicity. The following lemma shows how the symmetrization step in the algorithm makes this possible.

Lemma 4.5. *For* Pairs_3 *and* $\mathrm{Triples}_3(\eta)$ *defined in Algorithm 4, we have:*

$$\begin{aligned} \mathrm{Pairs}_3 &= O_3 \,\mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2) O_3^\top \\ \mathrm{Triples}_3(\eta) &= O_3 \,\mathrm{diag}(O_3^\top \eta) \,\mathrm{diag}(\mu_{1,3}, \mu_{2,3}, \ldots, \mu_{k,3}) O_3^\top \end{aligned}$$

Proof. Without loss of generality, suppose O_v are in canonical form (for each i, $\sigma_i^2 = 1$). Hence, $A \,\mathrm{Pairs}_{12}\, B^\top = AO_1(BO_2)^\top$. Hence, AO_1 and BO_2 are invertible. Note that:

$$\mathrm{Pairs}_{31}\, A^\top (B \,\mathrm{Pairs}_{21}\, A^\top)^{-1} B \,\mathrm{Pairs}_{23} = O_3 O_1^\top A^\top (B O_2 O_1^\top A^\top)^{-1} B O_2 O_3^\top = O_3 O_3^\top$$

which proves the first claim. The proof of the second claim is analogous. □

Again, we say that all O_v are in a canonical form if, for each i, $\sigma_i^2 = 1$.

Theorem 4.4 (The multi-view case)**.** *We have:*

- *(No False Positives) For all* $\theta \in \mathbb{R}^k$*, Algorithm 4 returns a subset of* O_3*, in a canonical form.*
- *(Exact Recovery) Assume that* γ_i *is nonzero for each* i*. Suppose* $\theta \in \mathbb{R}^k$ *is a random vector uniformly sampled over the sphere* $\mathcal{S}^{k-1}$*. With probability* 1*, Algorithm 4 returns all columns of* O_3*, in a canonical form.*

Proof of Theorem 4.4. The proof is identical to that of Theorem 4.1. □

Remark 9 (Simpler algorithms for HMMs)*.* Mossel and Roch (2006); Anandkumar et al. (2012) provide eigenvector based algorithms for HMM parameter estimation. These results show that we can achieve parameter estimation with only two SVDs (see Anandkumar et al. (2012) for the reduction of an HMM to the multi-view setting). The key idea is the symmetrization that reduces the problem to a single view.

Remark 10 (Finding A and B)*.* Suppose $\Theta, \Theta' \in \mathbb{R}^{d \times k}$ are random matrices with entries sampled independently from a standard normal. Set $A = \mathrm{Pairs}_{1,2}\, \Theta$ and $B = \mathrm{Pairs}_{2,1}\, \Theta'$. With probability 1, $\mathrm{Range}(A) = \mathrm{Range}(O_1)$ and $\mathrm{Range}(B) = \mathrm{Range}(O_2)$, and the invertibility condition will be satisfied (provided that O_1 and O_2 are full rank).

5 Sample Complexity

Let us now provide an efficient algorithm utilizing samples from documents, rather than exact statistics. The following theorem shows that the empirical version of ECA returns accurate estimates of the topics. Furthermore, each run of the algorithm succeeds with probability greater than 3/4 so the algorithm may be repeatedly run. Primarily for theoretical analysis, Algorithm 5 uses a rescaling procedure (rather than explicitly normalizing the topics, which would involve some thresholding procedure; see Remark 11).

Algorithm 5 Empirical ECA for LDA

Input: an integer k; an integer N; vector $\theta \in \mathbb{R}^k$; the sum α_0

1. **Find Empirical Averages:** With N independent samples (of documents), compute the empirical first, second, and third moments. Then compute the empirical moments $\widehat{\text{Pairs}}_{\alpha_0}$ and $\widehat{\text{Triples}}_{\alpha_0}(\eta)$.

2. **Whiten:** Let $\hat{W} = A\Sigma^{-1/2} \in \mathbb{R}^{d\times k}$ where $A \in \mathbb{R}^{d\times k}$ is the matrix of the orthonormal left singular vectors of $\widehat{\text{Pairs}}_{\alpha_0}$, corresponding to the largest k singular values, and $\Sigma \in \mathbb{R}^{k\times k}$ is the corresponding diagonal matrix of the k largest singular values.

3. **SVD:** Let $\{\hat{v}_1, \hat{v}_2, \ldots \hat{v}_k\}$ be the set of (left) singular vectors of

$$\hat{W}^\top \widehat{\text{Triples}}_{\alpha_0}(\hat{W}\theta)\hat{W}$$

4. **Reconstruct and Scale:** Return the set $\{\hat{O}_1, \hat{O}_2, \ldots \hat{O}_k\}$ where

$$\hat{Z}_i = \frac{2}{(\alpha_0+2)(\hat{W}\hat{v}_i)^\top \widehat{\text{Triples}}_{\alpha_0}(\hat{W}\hat{v}_i)\hat{W}\hat{v}_i}$$

$$\hat{O}_i = \frac{1}{\hat{Z}_i}(\hat{W}^+)^\top \hat{v}_i$$

(See Remark 11 for a procedure which explicitly normalizes $\hat{O}_i$.)

Theorem 5.1 (Sample Complexity for LDA)**.** *Fix $\delta \in (0,1)$. Let $p_{\min} = \min_i \frac{\alpha_i}{\alpha_0}$ and let $\sigma_k(O)$ denote the smallest (non-zero) singular value of O. Suppose that we obtain $N \geq \left(\frac{(\alpha_0+1)(6+6\sqrt{\ln(3/\delta)})}{p_{\min}\sigma_k(O)^2}\right)^2$ independent samples of x_1, x_2, x_3 in the LDA model. With probability greater than $1-\delta$, the following holds: for $\theta \in \mathbb{R}^k$ sampled uniformly sampled over the sphere $\mathcal{S}^{k-1}$, with probability greater than 3/4, Algorithm 5 returns a set $\{\hat{O}_1, \hat{O}_2, \ldots \hat{O}_k\}$ such that there exists a permutation σ of $\{1, 2, \ldots k\}$ (a permutation of the columns) so that for all $i \in \{1, 2, \ldots k\}$*

$$\|O_i - \hat{O}_{\sigma(i)}\|_2 \leq c\, \frac{(\alpha_0+1)^2 k^3}{p_{\min}^2 \sigma_k(O)^3} \left(\frac{1+\sqrt{\ln(1/\delta)}}{\sqrt{N}}\right)$$

where c is a universal constant.

Remark 11 (Normalizing and ℓ_1 accuracy). An alternative procedure would be to just explicitly normalize $\hat{O}_i$. If d large, to do this robustly, one should first set to 0 the smallest elements and then normalize. The reason for clipping the smallest elements is related to obtaining low ℓ_1 error.

Our theorem currently guarantees ℓ_2 norm accuracy of each column. Another natural error measure for probability distributions is the ℓ_1 error (the total variation error). Ideally, we would like the ℓ_1 error to be small with a number of samples does not depend on the dimension d (*e.g.*, the size of the vocabulary). Unfortunately, in general, this is not possible. For example, in the simplest case where $k = 1$ (*i.e.*, every document is about the same topic), then this amounts to estimating the distributions over words for this topic; in other words, we must estimate a distribution over d, which may require $\Omega(d)$ samples to obtain some fixed target ℓ_1-error. However, this situation occurs only when the target distribution is near to uniform. If instead, for each topic, say most of the probability mass is contained within the most frequent $d_{\text{effective}}$ words (for that topic), then it is possible to translate our ℓ_2 error guarantee into an ℓ_1 guarantee (in terms of $d_{\text{effective}}$).

6 Discussion: Sparsity

Note that sparsity considerations have not entered into our analysis. Often, in high dimensional statistics, notions of sparsity are desired as this generally decreases the sample size requirements (often at an increased computational burden).

Here, while these results have no explicit dependence on the sparsity level, sparsity is helpful in that it does implicitly affect the skewness (and the whitening) , which determines the sample complexity. As the model becomes less sparse, the skewness tends to 0. In particular, for the case of LDA, as $\alpha_0 \to \infty$ note that error increases (see Theorem 5.1).

Perhaps surprisingly, the sparsity level has no direct impact on the computational requirements of a "plug-in" empirical algorithm (beyond the linear time requirement of reading the data in order to construct the empirical statistics).

Acknowledgements

We thank Kamalika Chaudhuri, Adam Kalai, Percy Liang, Chris Meek, David Sontag, and Tong Zhang for many invaluable insights. We also give warm thanks to Rong Ge for sharing early insights and their preliminary results (in Arora et al. (2012)) into this problem with us.

References

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

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

R. Ando and T. Zhang. Two-view feature generation model for semi-supervised learning. In *ICML*, 2007.

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

S. Arora, R. Ge, and A. Moitra. Learning topic models — going beyond svd. *arXiv:1204.1956*, Apr 2012.

A. Asuncion, P. Smyth, M. Welling, D. Newman, I. Porteous, and S. Triglia. Distributed gibbs sampling for latent variable models. In *Scaling Up Machine Learning: Parallel and Distributed Approaches*. Cambridge Univ Pr, 2011.

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

David M. Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

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

Jean-Franois Cardoso and Pierre Comon. Independent component analysis, a survey of some algebraic methods. In *IEEE International Symposium on Circuits and Systems*, pages 93–96, 1996.

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

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

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

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

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

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

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

Alan M. Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In *FOCS*, 1996.

M.D. Hoffman, D.M. Blei, and F. Bach. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2010.

Thomas Hofmann. Probilistic latent semantic analysis. In *UAI*, 1999.

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.

Aapo Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

Sham M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysi s. In Nader H. Bshouty and Claudio Gentile, editors, *COLT*, volume 4539 of *Lecture Notes in Computer Science*, pages 82–96. Springer, 2007.

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.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401, 1999.

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

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

A. 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.

Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *J. Comput. Syst. Sci.*, 61(2), 2000.

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

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

Mark Steyvers and Tom Griffiths. Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum, 2006. URL `http://cocosci.berkeley.edu/tom/papers/SteyversGriffiths.pdf`.

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

A Analysis with Independent Factors

Lemma A.1 (Hidden state moments)**.** *Let $z = h - \mathbb{E}[h]$. For any vectors $u, v \in \mathbb{R}^k$,*

$$\begin{aligned} \mathbb{E}[zz^\top] &= \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2) \\ \mathbb{E}[zz^\top \langle u, z\rangle] &= \operatorname{diag}(u) \operatorname{diag}(\mu_{i,3}, \mu_{2,3}, \ldots, \mu_{k,3}) \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[zz^\top \langle u, z\rangle \langle v, z\rangle] = \operatorname{diag}(u) \operatorname{diag}(v) \operatorname{diag}(\mu_{1,4} - 3\sigma_1^4, \mu_{2,4} - 3\sigma_2^4, \ldots, \mu_{k,4} - 3\sigma_k^4) \\ + (u^\top \mathbb{E}[zz^\top] v)\mathbb{E}[zz^\top] + (\mathbb{E}[zz^\top]u)(\mathbb{E}[zz^\top]v)^\top + (\mathbb{E}[zz^\top]v)(\mathbb{E}[zz^\top]u)^\top \end{aligned}$$

Proof. Let a, b, u and v be vectors. Since the $\{z_t\}$ are independent and have mean zero, we have:

$$\mathbb{E}[\langle a, z\rangle \langle b, z\rangle] = \mathbb{E}\left[\left(\sum_{i=1}^k a_i z_i\right)\left(\sum_{i=1}^k b_i z_i\right)\right] = \sum_{i=1}^k a_i b_i \mathbb{E}[z_i^2] = \sum_{i=1}^k a_i b_i \sigma_i^2$$

and

$$\mathbb{E}[\langle a, z\rangle \langle b, z\rangle \langle u, z\rangle] = \mathbb{E}\left[\left(\sum_{i=1}^k a_i z_i\right)\left(\sum_{i=1}^k b_i z_i\right)\left(\sum_{i=1}^k u_i z_i\right)\right] = \sum_{i=1}^k a_i b_i u_i \mathbb{E}[z_i^3] = \sum_{i=1}^k a_i b_i u_i \mu_{i,3}.$$

For the final claim, let us compute the diagonal and non-diagonal entries separately. First,

$$\begin{aligned} \mathbb{E}[z_i z_i \langle u, z\rangle \langle v, z\rangle] &= \mathbb{E}[\sum_{j,k} u_j v_k z_i z_i z_j z_k] \\ &= u_i v_i \mathbb{E}[z_i^4] + \sum_{j \neq i} u_j v_j \mathbb{E}[z_i^2]\mathbb{E}[z_j^2] \\ &= u_i v_i \mu_{i,4} + \sigma_i^2 \sum_{j \neq i} u_j v_j \sigma_j^2 \\ &= u_i v_i \mu_{i,4} - u_i v_i (\sigma_i^2)^2 + \sigma_i^2 \sum_j u_j v_j \sigma_j^2 \\ &= u_i v_i \mu_{i,4} - u_i v_i (\sigma_i^2)^2 + (u^\top \mathbb{E}[zz^\top] v)\sigma_i^2 \end{aligned}$$

For $j \neq i$

$$\begin{aligned} \mathbb{E}[z_i z_j \langle u, z\rangle \langle v, z\rangle] &= \mathbb{E}[\sum_{k,l} u_k v_l z_i z_j z_k z_l] \\ &= u_i v_j \mathbb{E}[z_i^2 z_j^2] + u_j v_i \mathbb{E}[z_i^2 z_j^2] \\ &= u_i v_j \sigma_i^2 \sigma_j^2 + u_j v_i \sigma_i^2 \sigma_j^2 \\ &= [\mathbb{E}[zz^\top]u]_i [\mathbb{E}[zz^\top]v]_j + [\mathbb{E}[zz^\top]u]_j [\mathbb{E}[zz^\top]v]_i \end{aligned}$$

The proof is completed by noting the (i, j)-th components of $\mathbb{E}[zz^\top \langle u, z\rangle \langle v, z\rangle]$ agree with the above moment expressions. ∎

The proofs of Lemmas 4.1 and 4.2 follow.

Proof of Lemmas 4.1 and 4.2. By the conditional independence of $\{x_1, x_2, x_3\}$ given h,

$$\mathbb{E}[x_1] = O\mathbb{E}[h]$$

and

$$\begin{aligned}\mathbb{E}[(x_1-\mu)(x_2-\mu)^\top] &= \mathbb{E}[\mathbb{E}[(x_1-\mu)(x_2-\mu)^\top|h]] \\ &= \mathbb{E}[\mathbb{E}[(x_1-\mu)|h]\mathbb{E}[(x_2-\mu)^\top|h]] \\ &= O\mathbb{E}[(h-\mathbb{E}[h])(h-\mathbb{E}[h])^\top]O^\top \\ &= O\operatorname{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2)O^\top\end{aligned}$$

by Lemma A.1.

Similarly, the (i,j)-th entry of Triples(η) is

$$\begin{aligned}\mathbb{E}\big[\langle e_i, x_1-\mu\rangle\langle e_j, x_2-\mu\rangle\langle \eta, x_3-\mu\rangle\big] &= \mathbb{E}\big[\mathbb{E}[\langle e_i, x_1-\mu\rangle\langle e_j, x_2-\mu\rangle\langle \eta, x_3-\mu\rangle|h]\big] \\ &= \mathbb{E}\big[\mathbb{E}[\langle e_i, x_1-\mu\rangle|h]\cdot\mathbb{E}[\langle e_j, x_2-\mu\rangle|h]\cdot\mathbb{E}[\langle \eta, x_3-\mu\rangle|h]\big] \\ &= \mathbb{E}\big[\langle e_i, O(h-\mathbb{E}[h])\rangle\langle e_j, O(h-\mathbb{E}[h])\rangle\langle \eta, O(h-\mathbb{E}[h])\rangle\big] \\ &= \mathbb{E}\big[\langle O^\top e_i, h-\mathbb{E}[h]\rangle\langle O^\top e_j, h-\mathbb{E}[h]\rangle\langle O^\top \eta, h-\mathbb{E}[h]\rangle\big] \\ &= e_i^\top O\operatorname{diag}(O^\top\eta)\operatorname{diag}(\mu_{i,3}, \mu_{2,3}, \dots, \mu_{k,3})O^\top e_j.\end{aligned}$$

The proof for Quadruples(η, η') is analogous. □

The proof for Lemma 4.4 is analogous to the above proofs.

B Analysis with Dirichlet Factors

We first provide the functional forms of the first, second, and third moments. With these, we prove Lemma 4.3.

B.1 Dirichlet moments

Lemma B.1 (Dirichlet moments)**.** *We have:*

$$\mathbb{E}[h\otimes h] = \frac{1}{(\alpha_0+1)\alpha_0}\big(\operatorname{diag}(\alpha) + \alpha\alpha^\top\big)$$

and

$$\begin{aligned}\mathbb{E}[h\otimes h\otimes h] = \frac{1}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\Big(&\alpha\otimes\alpha\otimes\alpha + \sum_{i=1}^k \alpha_i\big(e_i\otimes e_i\otimes\alpha\big) + \sum_{i=1}^k \alpha_i\big(\alpha\otimes e_i\otimes e_i\big) \\ &+ \sum_{i=1}^k \alpha_i\big(e_i\otimes\alpha\otimes e_i\big) + 2\sum_{i=1}^k \alpha_i\big(e_i\otimes e_i\otimes e_i\big)\Big).\end{aligned}$$

Hence, for $v \in \mathbb{R}^k$,

$$\begin{aligned}\mathbb{E}[(h\otimes h)\langle v, h\rangle] = \frac{1}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\Big(&\langle v, \alpha\rangle\alpha\alpha^\top + \operatorname{diag}(\alpha)v\alpha^\top + \alpha v^\top\operatorname{diag}(\alpha) \\ &+ \langle v, \alpha\rangle\operatorname{diag}(\alpha) + 2\operatorname{diag}(v)\operatorname{diag}(\alpha)\Big)\end{aligned}$$

Proof. First, let us specify the following scalar moments.

Univariate moments: Fix some $i \in [k]$, and let $\alpha' := \alpha + p \cdot e_i$ for some positive integer p. Then

$$\begin{aligned}\mathbb{E}[h_i^p] &= \frac{Z(\alpha')}{Z(\alpha)} \\ &= \frac{\Gamma(\alpha_i + p)}{\Gamma(\alpha_i)} \cdot \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + p)} \\ &= \frac{(\alpha_i + p - 1)(\alpha_i + p - 2) \cdots \alpha_i}{(\alpha_0 + p - 1)(\alpha_0 + p - 2) \cdots \alpha_0}.\end{aligned}$$

In particular,

$$\begin{aligned}\mathbb{E}[h_i] &= \frac{\alpha_i}{\alpha_0} \\ \mathbb{E}[h_i^2] &= \frac{(\alpha_i + 1)\alpha_i}{(\alpha_0 + 1)\alpha_0} \\ \mathbb{E}[h_i^3] &= \frac{(\alpha_i + 2)(\alpha_i + 1)\alpha_i}{(\alpha_0 + 2)(\alpha_0 + 1)\alpha_0}.\end{aligned}$$

Bivariate moments: Fix $i, j \in [k]$ with $i \neq j$, and let $\alpha' := \alpha + p \cdot e_i + q \cdot e_j$ for some positive integers p and q. Then

$$\begin{aligned}\mathbb{E}[h_i^p h_j^q] &= \frac{Z(\alpha')}{Z(\alpha)} \\ &= \frac{\Gamma(\alpha_i + p) \cdot \Gamma(\alpha_j + q)}{\Gamma(\alpha_i) \cdot \Gamma(\alpha_j)} \cdot \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + p + q)} \\ &= \frac{\big((\alpha_i + p - 1)(\alpha_i + p - 2) \cdots \alpha_i\big) \cdot \big((\alpha_j + q - 1)(\alpha_j + q - 2) \cdots \alpha_j\big)}{(\alpha_0 + p + q - 1)(\alpha_0 + p + q - 2) \cdots \alpha_0}.\end{aligned}$$

In particular,

$$\begin{aligned}\mathbb{E}[h_i h_j] &= \frac{\alpha_i \alpha_j}{(\alpha_0 + 1)\alpha_0} \\ \mathbb{E}[h_i^2 h_j] &= \frac{(\alpha_i + 1)\alpha_i \alpha_j}{(\alpha_0 + 2)(\alpha_0 + 1)\alpha_0}.\end{aligned}$$

Trivariate moments: Fix $i, j, \kappa \in [k]$ all distinct, and let $\alpha' := \alpha + e_i + e_j + e_\kappa$. Then

$$\begin{aligned}\mathbb{E}[h_i h_j h_\kappa] &= \frac{Z(\alpha')}{Z(\alpha)} \\ &= \frac{\Gamma(\alpha_i + 1) \cdot \Gamma(\alpha_j + 1) \cdot \Gamma(\alpha_\kappa + 1)}{\Gamma(\alpha_i) \cdot \Gamma(\alpha_j) \cdot \Gamma(\alpha_\kappa)} \cdot \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + 3)} \\ &= \frac{\alpha_i \alpha_j \alpha_\kappa}{(\alpha_0 + 2)(\alpha_0 + 1)\alpha_0}.\end{aligned}$$

Completing the proof: The proof for the second moment matrix and the third moment tensor follows by observing that each component agrees with the above expressions. For the final claim,

$$\begin{aligned}\mathbb{E}[(h \otimes h)\langle v, h\rangle] &= \frac{1}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\Big(\langle v, \alpha\rangle(\alpha \otimes \alpha) + \sum_{i=1}^{k} \alpha_i v_i \big(e_i \otimes \alpha\big) + \sum_{i=1}^{k} \alpha_i v_i \big(\alpha \otimes e_i\big) \\ &\quad + \sum_{i=1}^{k} \alpha_i \langle v, \alpha\rangle (e_i \otimes e_i) + 2\sum_{i=1}^{k} \alpha_i v_i (e_i \otimes e_i)\Big) \\ &= \frac{1}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\Big(\langle v, \alpha\rangle \alpha\alpha^\top + \operatorname{diag}(\alpha) v \alpha^\top + \alpha v^\top \operatorname{diag}(\alpha) \\ &\qquad + \langle v, \alpha\rangle \operatorname{diag}(\alpha) + 2 \operatorname{diag}(v) \operatorname{diag}(\alpha)\Big)\end{aligned}$$

which completes the proof. □

B.2 The proof of Lemma 4.3

Proof. Observe:

$$\mathbb{E}[x_1] = O\mathbb{E}[h]$$

and

$$\mathbb{E}[x_1 x_2^\top] = \mathbb{E}[\mathbb{E}[x_1 x_2^\top | h]] = O\mathbb{E}[hh^\top]O^\top$$

Define the analogous quantity:

$$\mathrm{Pairs}_h = \mathbb{E}[hh^\top] - \frac{\alpha_0}{\alpha_0+1}\mathbb{E}[h]\mathbb{E}[h]^\top$$

and so:

$$\mathrm{Pairs}_{\alpha_0} = O \, \mathrm{Pairs}_h \, O^\top$$

Observe:

$$\begin{aligned}\mathrm{Pairs}_h &= \mathbb{E}[hh^\top] - \frac{1}{(\alpha_0+1)\alpha_0}\alpha\alpha^\top \\ &= \frac{1}{(\alpha_0+1)\alpha_0}\operatorname{diag}(\alpha)\end{aligned}$$

Hence,

$$\mathrm{Pairs}_{\alpha_0} = O\,\mathrm{Pairs}_h\,O^\top = \frac{1}{(\alpha_0+1)\alpha_0} O \operatorname{diag}(\alpha) O^\top$$

which proves the first claim.

Also, define:

$$\begin{aligned}\mathrm{Triples}_h(v) := \mathbb{E}[(h\otimes h)\langle v, h\rangle] - \frac{\alpha_0}{\alpha_0+2}\Big(\mathbb{E}[hh^\top]v\mathbb{E}[h]^\top + \mathbb{E}[h]v^\top\mathbb{E}[hh^\top] + \langle v, \mathbb{E}[h]\rangle \mathbb{E}[hh^\top]\Big) \\ + \frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}\langle v, \mathbb{E}[h]\rangle \mathbb{E}[h]\mathbb{E}[h]^\top\end{aligned}$$

Since

$$\begin{aligned}\mathbb{E}[x_1 x_2^\top \langle \eta, x_3\rangle] &= \mathbb{E}[\mathbb{E}[x_1 x_2^\top \langle \eta, x_3\rangle | h]] \\ &= O\mathbb{E}[hh^\top \langle \eta, Oh\rangle]O^\top \\ &= O\mathbb{E}[hh^\top \langle O^\top \eta, h\rangle]O^\top\end{aligned}$$

we have

$$\mathrm{Triples}_{\alpha_0}(\eta) = O\,\mathrm{Triples}_h(O^\top \eta)O^\top$$

Let us complete the proof by showing:

$$\mathrm{Triples}_h(v) := \frac{2}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\,\mathrm{diag}(v)\,\mathrm{diag}(\alpha)$$

Observe:

$$\begin{aligned}\frac{2}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\,\mathrm{diag}(v)\,\mathrm{diag}(\alpha) = \mathbb{E}[(h\otimes h)\langle v, h\rangle] - \frac{1}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\Big(&\langle v,\alpha\rangle\alpha\alpha^\top + \mathrm{diag}(\alpha)v\alpha^\top \\ &+ \alpha v^\top\,\mathrm{diag}(\alpha) + \langle v,\alpha\rangle\,\mathrm{diag}(\alpha)\Big)\end{aligned}$$

Let us handle each term separately. First,

$$\frac{1}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\langle v,\alpha\rangle\alpha\alpha^\top = \frac{\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}\langle v, \mathbb{E}[h]\rangle \mathbb{E}[h]\mathbb{E}[h]^\top$$

Also, since:

$$\frac{1}{(\alpha_0+1)\alpha_0}\,\mathrm{diag}(\alpha) = \mathbb{E}[hh^\top] - \frac{1}{(\alpha_0+1)\alpha_0}\alpha\alpha^\top$$

we have:

$$\begin{aligned}&\frac{1}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\Big(\mathrm{diag}(\alpha)v\alpha^\top + \alpha v^\top\,\mathrm{diag}(\alpha) + \langle v,\alpha\rangle\,\mathrm{diag}(\alpha)\Big) \\ =&\frac{1}{\alpha_0+2}\Big(\mathbb{E}[hh^\top]v\alpha^\top + \alpha v^\top \mathbb{E}[hh^\top] + \langle v,\alpha\rangle\mathbb{E}[hh^\top]\Big) - \frac{3}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\langle v,\alpha\rangle\alpha\alpha^\top \\ =&\frac{\alpha_0}{\alpha_0+2}\Big(\mathbb{E}[hh^\top]v\mathbb{E}[h]^\top + \mathbb{E}[h]v^\top\mathbb{E}[hh^\top] + \langle v,\mathbb{E}[h]\rangle\mathbb{E}[hh^\top]\Big) - \frac{3\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}\langle v,\mathbb{E}[h]\rangle\mathbb{E}[h]\mathbb{E}[h]^\top\end{aligned}$$

Hence,

$$\begin{aligned}\frac{2}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\,\mathrm{diag}(v)\,\mathrm{diag}(\alpha) &= \mathbb{E}[(h\otimes h)\langle v,h\rangle] \\ &\quad - \frac{\alpha_0}{\alpha_0+2}\Big(\mathbb{E}[hh^\top]v\mathbb{E}[h]^\top + \mathbb{E}[h]v^\top\mathbb{E}[hh^\top] + \langle v,\mathbb{E}[h]\rangle\mathbb{E}[hh^\top]\Big) \\ &\quad + \frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}\langle v,\mathbb{E}[h]\rangle\mathbb{E}[h]\mathbb{E}[h]^\top\end{aligned}$$

which proves the claim. □

C Sample Complexity Analysis

Throughout, we work in a canonical form. Define:

$$\tilde{O} := \frac{1}{\sqrt{(\alpha_0+1)\alpha_0}} O \operatorname{diag}(\sqrt{\alpha_1}, \sqrt{\alpha_2}, \ldots, \sqrt{\alpha_k})$$

Under this transformation, we have:

$$\mathrm{Pairs}_{\alpha_0} = \frac{1}{(\alpha_0+1)\alpha_0} O \operatorname{diag}(\alpha) O^\top = \tilde{O}\tilde{O}^\top$$

Using the definition:

$$\gamma_i := 2\sqrt{\frac{\alpha_0(\alpha_0+1)}{(\alpha_0+2)^2}\ \frac{1}{\alpha_i}}$$

we also have that :

$$\mathrm{Triples}_{\alpha_0}(\eta) = \tilde{O} \operatorname{diag}(\tilde{O}^\top \eta) \operatorname{diag}(\gamma) \tilde{O}^\top$$

Hence, we can consider γ_i to be the effective skewness. Let us also define:

$$p_{\min} := \min_i \frac{\alpha_i}{\alpha_0}$$

Since $\alpha_i \le \alpha_0$, we have that:

$$\frac{1}{\sqrt{\alpha_0+2}} \le \gamma_i \le 2\frac{1}{\sqrt{p_{\min}(\alpha_0+2)}}$$

Note that:

$$\sigma_k(O)\sqrt{\frac{p_{\min}}{\alpha_0+1}} \le \sigma_k(\tilde{O}) \le 1$$

and

$$\sigma_1(\tilde{O}) \le \sigma_1(O)\frac{1}{\sqrt{\alpha_0+1}} \le \frac{1}{\sqrt{\alpha_0+1}}$$

where $\sigma_j(\cdot)$ denotes the j-th largest singular value. These lower bounds are relevant for lower bounding certain singular values in our analysis.

We use $\|M\|$ to denote the spectral norm of a matrix M. Let us suppose that for all η,

$$\begin{aligned} \|\widehat{\mathrm{Pairs}_{\alpha_0}} - \mathrm{Pairs}_{\alpha_0}\| &= E_P \\ \|\mathrm{Triples}_{\alpha_0}(\eta) - \widehat{\mathrm{Triples}_{\alpha_0}}(\eta)\| &\le \|\eta\| E_T \end{aligned}$$

for some E_P and E_T (which we set later).

C.1 Perturbation Lemmas

Let $\widehat{\text{Pairs}}_{\alpha_0,k}$ be the best rank k approximation to Pairs_{α_0}. We have that $\hat{W}$, as defined in Algorithm 5, whitens $\widehat{\text{Pairs}}_{\alpha_0,k}$, *i.e.*,

$$\hat{W}^\top \ \widehat{\text{Pairs}}_{\alpha_0,k}\hat{W} = \mathrm{I}\,.$$

Let

$$\hat{W}^\top \text{Pairs}_{\alpha_0} \hat{W} = ADA^\top$$

be an SVD of $\hat{W}^\top \text{Pairs}_{\alpha_0} \hat{W}$, where $A \in \mathbb{R}^{k\times k}$. Define:

$$W := \hat{W} A D^{-1/2} A^\top$$

and observe that W also whitens Pairs_{α_0}, *i.e.*,

$$W^\top \text{Pairs}_{\alpha_0} W = (AD^{-1/2}A^\top)^\top \hat{W}^\top \text{Pairs}_{\alpha_0} \hat{W}(AD^{-1/2}A^\top) = \mathrm{I}$$

Due to sampling error, the range(W) may not equal the range$(\text{Pairs}_{\alpha_0})$.

Define:

$$M := W^\top \tilde{O}, \qquad \hat{M} = \hat{W}^\top \tilde{O}$$

Lemma C.1. *Let Π_W be the orthogonal projection onto the range of W and Π be the orthogonal projection onto the range of O. Suppose $E_P \leq \sigma_k(\text{Pairs}_{\alpha_0})/2$. We have that:*

$$\begin{aligned}
\|M\| &= 1 \\
\|\hat{M}\| &\leq 2 \\
\|\hat{W}\| &\leq \frac{2}{\sigma_k(\tilde{O})} \\
\|\hat{W}^+\| &\leq 2\sigma_1(\tilde{O}) \\
\|W^+\| &\leq 3\sigma_1(\tilde{O}) \\
\|M - \hat{M}\| &\leq \frac{4}{\sigma_k(\tilde{O})^2} E_P \\
\|\hat{W}^+ - W^+\| &\leq \frac{6\sigma_1(\tilde{O})}{\sigma_k(\tilde{O})^2} E_P \\
\|\Pi - \Pi_W\| &\leq \frac{4}{\sigma_k(\tilde{O})^2} E_P
\end{aligned}$$

Proof. Since W whitens Pairs_{α_0}, we have $MM^\top = W^\top \tilde{O}\tilde{O}^\top W = \mathrm{I}$ and

$$\|M\| = 1$$

By Weyl's theorem (see Lemma E.1),

$$\|\hat{W}\|^2 = \frac{1}{\sigma_k(\widehat{\text{Pairs}}_{\alpha_0})} \leq \frac{1}{\sigma_k(\text{Pairs}_{\alpha_0}) - \|\widehat{\text{Pairs}}_{\alpha_0} - \text{Pairs}_{\alpha_0}\|} \leq \frac{2}{\sigma_k(\text{Pairs}_{\alpha_0})} = \frac{2}{\sigma_k(\tilde{O})^2}$$

Also, $\hat{W} = WAD^{1/2}A^\top$ so that $\hat{M} = AD^{1/2}A^\top M$ and

$$\begin{aligned}
\|M - \hat{M}\| &= \|M - AD^{1/2}A^\top M\| \\
&\leq \|M\|\|\mathrm{I} - AD^{1/2}A^\top\| \\
&= \|\mathrm{I} - D^{1/2}\| \\
&\leq \|\mathrm{I} - D^{1/2}\|\|\mathrm{I} + D^{1/2}\| \\
&= \|\mathrm{I} - D\|
\end{aligned}$$

where we have used that $D \succeq 0$ and D is diagonal.

We can bound this as follows:

$$\begin{aligned}
\|\mathrm{I} - D\| = &= \|\mathrm{I} - ADA^\top\| \\
&= \|\mathrm{I} - \hat{W}^\top \mathrm{Pairs}_{\alpha_0} \hat{W}\| \\
&= \|\hat{W}^\top(\widehat{\mathrm{Pairs}_{\alpha_0,k}} - \mathrm{Pairs}_{\alpha_0})\hat{W}\| \\
&\leq \|\hat{W}\|^2\|\widehat{\mathrm{Pairs}_{\alpha_0,k}} - \mathrm{Pairs}_{\alpha_0}\| \\
&\leq \|\hat{W}\|^2(\|\widehat{\mathrm{Pairs}_{\alpha_0,k}} - \widehat{\mathrm{Pairs}_{\alpha_0}}\| + \|\widehat{\mathrm{Pairs}_{\alpha_0}} - \mathrm{Pairs}_{\alpha_0}\|) \\
&= \|\hat{W}\|^2(\sigma_{k+1}(\widehat{\mathrm{Pairs}_{\alpha_0}}) + \|\widehat{\mathrm{Pairs}_{\alpha_0}} - \mathrm{Pairs}_{\alpha_0}\|) \\
&\leq 2\|\hat{W}\|^2\|\widehat{\mathrm{Pairs}_{\alpha_0}} - \mathrm{Pairs}_{\alpha_0}\| \\
&\leq 4\frac{1}{\sigma_k(\tilde{O})^2}E_P
\end{aligned}$$

using Weyl's theorem in the second to last step.

This implies $\|\mathrm{I} - D\| \leq 4\frac{1}{\sigma_k(\tilde{O})^2}E_P \leq 2$ and so $\|D\| \leq 3$. Since $\hat{M} = AD^{1/2}A^\top M$,

$$\|\hat{M}\|^2 \leq \|M\|^2\|D\| \leq 3\,.$$

Again, by Weyl's theorem,

$$\|\hat{W}^+\|^2 = \sigma_1(\widehat{\mathrm{Pairs}_{\alpha_0}}) \leq \sigma_1(\widehat{\mathrm{Pairs}_{\alpha_0}}) + E_P \leq 1.5\sigma_1(\mathrm{Pairs}_{\alpha_0}) = 1.5\sigma_1(\tilde{O})^2$$

Using that $W = \hat{W}AD^{-1/2}A^\top$, we have:

$$\|W^+\|^2 \leq \|\hat{W}^+\|^2\|D\| \leq 4.5\sigma_1(\tilde{O})^2$$

and

$$\|\hat{W}^+ - W^+\| \leq \|\hat{W}^+\|\|\mathrm{I} - D^{1/2}\| \leq \|\hat{W}^+\|\|\mathrm{I} - D\| \leq \frac{6\sigma_1(\tilde{O})}{\sigma_k(\tilde{O})^2}E_P$$

which completes the argument for the first set of claims.

We now prove the final claim. Let Θ be the matrix of canonical angles between range($\mathrm{Pairs}_{\alpha_0}$) and range($\widehat{\mathrm{Pairs}_{\alpha_0,k}}$). By Wedin's theorem (see Lemma E.3) (and noting that the k-th singular value of $\widehat{\mathrm{Pairs}_{\alpha_0,k}}$ is greater than $\sigma_k(\mathrm{Pairs}_{\alpha_0})/2$), we have

$$\|\sin\Theta\| \leq 2\frac{\|\mathrm{Pairs}_{\alpha_0} - \widehat{\mathrm{Pairs}_{\alpha_0,k}}\|}{\sigma_k(\mathrm{Pairs}_{\alpha_0})} \leq 2\frac{\|\widehat{\mathrm{Pairs}_{\alpha_0,k}} - \widehat{\mathrm{Pairs}_{\alpha_0}}\| + \|\widehat{\mathrm{Pairs}_{\alpha_0}} - \mathrm{Pairs}_{\alpha_0}\|}{\sigma_k(\mathrm{Pairs}_{\alpha_0})} \leq 4\frac{E_P}{\sigma_k(\mathrm{Pairs}_{\alpha_0})}$$

Using Lemma E.4, $\|\Pi - \Pi_W\| = \|\sin\Theta\|$, which completes the proof. □

Lemma C.2. *Suppose $E_P \leq \sigma_k(\text{Pairs}_{\alpha_0})/2$. For $\|\theta\| = 1$, we have:*

$$\|W^\top \text{Triples}_{\alpha_0}(W\theta)W - \hat{W}^\top \widehat{\text{Triples}}_{\alpha_0}(\hat{W}\theta)\hat{W}\| \ \leq \ c \left(\frac{E_P}{\sqrt{p_{\min}(\alpha_0+2)}\ \sigma_k(\tilde{O})^2} + \frac{E_T}{\sigma_k(\tilde{O})^3} \right)$$

where c is a universal constant.

Proof. We have:

$$\begin{aligned}\|W^\top \text{Triples}_{\alpha_0}(W\theta)W - \hat{W}^\top \widehat{\text{Triples}}_{\alpha_0}(\hat{W}\theta)\hat{W}\| &\leq \|W^\top \text{Triples}_{\alpha_0}(W\theta)W - \hat{W}^\top \text{Triples}_{\alpha_0}(\hat{W}\theta)\hat{W}\| \\ &\quad + \|\hat{W}^\top \text{Triples}_{\alpha_0}(\hat{W}\theta)\hat{W} - \hat{W}^\top \widehat{\text{Triples}}_{\alpha_0}(\hat{W}\theta)\hat{W}\|\end{aligned}$$

For the second term:

$$\begin{aligned}\|\hat{W}^\top \text{Triples}_{\alpha_0}(\hat{W}\theta)\hat{W} - \hat{W}^\top \widehat{\text{Triples}}_{\alpha_0}(\hat{W}\theta)\hat{W}\| &\leq \|\hat{W}\|^2 \|\text{Triples}_{\alpha_0}(\hat{W}\theta) - \widehat{\text{Triples}}_{\alpha_0}(\hat{W}\theta)\| \\ &\leq \|\hat{W}\|^3 E_T \\ &\leq \frac{8}{\sigma_k(\tilde{O})^3} E_T\end{aligned}$$

For the first term, by expanding out the terms and using the bounds in Lemma C.1, we have:

$$\begin{aligned}&\|W^\top \text{Triples}_{\alpha_0}(W\theta)W - \hat{W}^\top \text{Triples}_{\alpha_0}(\hat{W}\theta)\hat{W}\| \\ = \ &\|M \operatorname{diag}(M^\top\theta)\operatorname{diag}(\gamma)M^\top - \hat{M}\operatorname{diag}(\hat{M}^\top\theta)\operatorname{diag}(\gamma)\hat{M}^\top\| \\ \leq \ &\|M \operatorname{diag}(M^\top\theta)\operatorname{diag}(\gamma)M^\top - \hat{M}\operatorname{diag}(M^\top\theta)\operatorname{diag}(\gamma)\hat{M}^\top\| \\ &+\|\hat{M}\operatorname{diag}((M-\hat{M})^\top\theta)\operatorname{diag}(\gamma)\hat{M}^\top\| \\ \leq \ &\|M \operatorname{diag}(M^\top\theta)\operatorname{diag}(\gamma)M^\top - \hat{M}\operatorname{diag}(M^\top\theta)\operatorname{diag}(\gamma)\hat{M}^\top\| + \max_i \gamma_i \|\hat{M}\|^2\|M-\hat{M}\| \\ \leq \ &c \max_i \gamma_i \|M - \hat{M}\| \qquad (2)\end{aligned}$$

for some constant c (where the last step follows from expanding out terms). □

C.2 SVD Accuracy

Let σ_i and v_i denote the corresponding i-th singular value (in increasing order) and vector of $W^\top \text{Triples}_{\alpha_0}(W\theta)W$. Similarly, let $\hat{v}_i$ and $\hat{\sigma}_i$ denote the corresponding i-th singular value (in increasing order) and vector of $\hat{W}^\top \widehat{\text{Triples}}_{\alpha_0}(\hat{W}\theta)\hat{W}$. For convenience, choose the sign of $\hat{v}_i$ so that $\langle v_i, \hat{v}_i\rangle \geq 0$.

The following lemma characterizes the accuracy of the SVD:

Lemma C.3 (SVD Accuracy)**.** *Suppose $E_P \leq \sigma_k(\text{Pairs}_{\alpha_0})/2$. With probability greater than $1-\delta'$, we have for all i:*

$$\|v_i - \hat{v}_i\| \leq c\frac{k^3\sqrt{\alpha_0+2}}{\delta'} \left(\frac{E_P}{\sqrt{p_{\min}(\alpha_0+2)}\ \sigma_k(\tilde{O})^2} + \frac{E_T}{\sigma_k(\tilde{O})^3} \right)$$

for some universal constant c.

First, let us provide a few lemmas. Let

$$\|W^\top \operatorname{Triples}_{\alpha_0}(W\theta)W - \hat{W}^\top \widehat{\operatorname{Triples}}_{\alpha_0}(\hat{W}\theta)\hat{W}\| \le E$$

where the bound on E is provided in Lemma C.2.

Lemma C.4. *Suppose for all i*

$$\sigma_i \ge \Delta$$
$$|\sigma_i - \sigma_{i+1}| \ge \Delta$$

For all i, v_i and $\hat{v}_i$, we have:

$$\|v_i - \hat{v}_i\| \le 2\frac{\sqrt{k}E}{\Delta - E}$$

where the sign of $\hat{v}_i$ chosen so $\langle \hat{v}_i, \hat{v}_i\rangle \ge 0$.

Proof. Let $\cos(\theta) = \langle v_i, \hat{v}_i\rangle$ (which is positive since we assume $\langle v_i, \hat{v}_i\rangle \ge 0$). We have:

$$\|v_i - \hat{v}_i\|^2 = 2(1-\cos(\theta)) =\le 2(1-\cos^2(\theta)) = 2\sin^2(\theta)$$

By Weyl's theorem (see Lemma E.1) and by assumption,

$$\min_i |\hat{\sigma}_i - \sigma_j| \ge \Delta - E$$

and

$$\min_{j\ne i} |\hat{\sigma}_i - \sigma_j| \ge \min_{j\ne i} |\sigma_i - \sigma_j| - E \ge \Delta - E$$

By Wedin's theorem (see Lemma E.2 applied to the split where v_i and v_i correspond to the subspaces U_1 and U_1),

$$|\sin(\theta)| \le \sqrt{2}\frac{E_F}{\Delta - E} \le \sqrt{2}\frac{\sqrt{k}E}{\Delta - E}$$

□

Lemma C.5. *Fix any $\delta \in (0,1)$ and matrix $A \in \mathbb{R}^{k\times k}$. Let $\theta \in \mathbb{R}^k$ be a random vector distributed uniformly over $\mathcal{S}^{k-1}$. With probability greater than $1-\delta$, we have*

$$\min_{i\ne j} |\langle \theta, A(e_i - e_j)\rangle| > \frac{\min_{i\ne j} \|A(e_i - e_j)\| \cdot \delta}{\sqrt{e}k^{2.5}}$$

and

$$\min_i |\langle \theta, Ae_i\rangle| > \frac{\min_i \|Ae_i\| \cdot \delta}{\sqrt{e}k^{2.5}}$$

Proof. By Lemma D.2, for any fixed pair $\{i,j\} \subseteq [k]$ and $\beta := \delta_0/\sqrt{e}$,

$$\Pr\left[|\langle \theta, A(e_i - e_j)\rangle| \le \|A(e_i - e_j)\| \cdot \frac{1}{\sqrt{k}} \cdot \frac{\delta_0}{\sqrt{e}}\right] \le exp\left(\frac{1}{2}(1 - (\delta_0^2/e) + \ln(\delta_0^2/e))\right) \le \delta_0.$$

Similarly, for each i

$$\Pr\left[|\langle \theta, Ae_i\rangle)| \le \|Ae_i\| \cdot \frac{1}{\sqrt{k}} \cdot \frac{\delta_0}{\sqrt{e}}\right] \le exp\left(\frac{1}{2}(1 - (\delta_0^2/e) + \ln(\delta_0^2/e))\right) \le \delta_0.$$

Let $\delta_0 := \delta/k^2$. The claim follows by a union bound over all $\binom{k}{2} + k \le k^2$ possibilities. □

We now complete the argument.

Proof of Lemma C.3. Choose $A = \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_k) M^\top$, where $M = W^\top \tilde{O}$. The proof of Theorem 4.1 shows $\langle e_i, A\theta \rangle$ are the singular values. Also the minimal singular value of A is greater than $\min_i \gamma_i \geq \frac{1}{\sqrt{\alpha_0+2}} \leq$ (since $MM^\top = \mathrm{I}$). Hence, we have:

$$\begin{aligned} \sigma_i &\geq \tfrac{\delta}{2k^{2.5}\sqrt{\alpha_0+2}} \quad := \Delta \\ |\sigma_i - \sigma_{i+1}| &\geq \tfrac{\delta}{2k^{2.5}\sqrt{\alpha_0+2}} \end{aligned}$$

Suppose $E \leq \Delta/2$. Here,

$$\|v_i - \hat{v}_i\| \leq 2\frac{\sqrt{k}E}{\Delta - E} \leq 4\frac{\sqrt{k}E}{\Delta} = 8\frac{k^3\sqrt{\alpha_0+2}}{\delta} D$$

Also, since $\|v_i - \hat{v}_i\| \leq 2$ the above also holds for $E > \Delta/2$, which proves the first claim. □

C.3 Reconstruction Accuracy

Lemma C.6. *Suppose $E_P \leq \sigma_k(\mathrm{Pairs}_{\alpha_0})/2$. With probability greater than $1-\delta'$, we have that for all i:*

$$\|O_i - \frac{1}{\hat{Z}_i} (\hat{W}^+)^\top \hat{v}_i\| \leq c\frac{(\alpha_0+2)^2 k^3}{p_{\min}^2 \sigma_k(O)^3 \delta'} \max\{E_P, E_T\}$$

where $\{O_1, O_2, \ldots O_k\}$ is some permutation of the columns of O.

Proof. First, observe that W whitens $\Pi_W \mathrm{Pairs}_{\alpha_0} \Pi_W^\top$. To see, observe that $\Pi_W^\top = \Pi_W$ (since Π_W is an orthogonal projection) and $\Pi_W^\top W = \Pi_W W = W$; so

$$W^\top (\Pi_W \mathrm{Pairs}_{\alpha_0} \Pi_W^\top) W = (\Pi_W^\top W)^\top \mathrm{Pairs}_{\alpha_0} (\Pi_W^\top W) = W^\top \mathrm{Pairs}_{\alpha_0} W = \mathrm{I}$$

Using the definition $M = W^\top \tilde{O} = W^\top \Pi_W \tilde{O}$, we have:

$$W^\top \mathrm{Triples}_{\alpha_0}(W\theta) W = M \,\mathrm{diag}(M^\top \theta)\, \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_k) M^\top$$

Since $\mathrm{range}(W) = \mathrm{range}(\Pi_W \tilde{O})$ the proof of Theorem 4.1 shows that:

$$\Pi_W \tilde{O}_i = (W^+)^\top v_i$$

Define:

$$Z_i := \frac{2}{(\alpha_0+2)(Wv_i)^\top \mathrm{Triples}_{\alpha_0}(Wv_i) W v_i},$$

Since $v_i = M^\top e_i$ are the singular vectors of $W^\top \mathrm{Triples}_{\alpha_0}(W\theta) W$, we have

$$(Wv_i)^\top \mathrm{Triples}_{\alpha_0}(Wv_i) W v_i = \gamma_i$$

and so:

$$Z_i = \sqrt{\frac{\alpha_i}{(\alpha_0+1)\alpha_0}}.$$

This implies $\tilde{O}_i = Z_i O_i$ and so

$$\Pi_W O_i = \frac{1}{Z_i}\Pi_W \tilde{O}_i = \frac{1}{Z_i}(W^+)^\top v_i$$

since $\Pi_W \tilde{O}_i = (W^+)^\top v_i$.

Now let us bound the reconstruction error as follows:

$$\begin{aligned}
& \|O_i - \frac{1}{\hat{Z}_i}\ (\hat{W}^+)^\top \hat{v}_i\| \\
\leq\ & \|O_i - \Pi_W O_i\| + \|\Pi_W O_i - \frac{1}{\hat{Z}_i}\ (\hat{W}^+)^\top \hat{v}_i\| \\
=\ & \|\Pi O_i - \Pi_W O_i\| + \|\frac{1}{Z_i}(W^+)^\top v_i - \frac{1}{\hat{Z}_i}\ (\hat{W}^+)^\top \hat{v}_i\| \\
\leq\ & \|\Pi - \Pi_W\|\|O_i\| + \|\frac{1}{Z_i}(W^+)^\top v_i - \frac{1}{Z_i}(W^+)^\top \hat{v}_i\| + \|\frac{1}{Z_i}(W^+)^\top \hat{v}_i - \frac{1}{\hat{Z}_i}\ (\hat{W}^+)^\top \hat{v}_i\| \\
\leq\ & \|\Pi - \Pi_W\| + \frac{\|W^+\|}{Z_i}\|v_i - \hat{v}_i\| + \|\frac{1}{Z_i}W^+ - \frac{1}{\hat{Z}_i}\ \hat{W}^+\| \\
\leq\ & \|\Pi - \Pi_W\| + \frac{\|W^+\|}{Z_i}\|v_i - \hat{v}_i\| + \|\frac{1}{Z_i}W^+ - \frac{1}{Z_i}\hat{W}^+\| + \|\frac{1}{Z_i}\hat{W}^+ - \frac{1}{\hat{Z}_i}\ \hat{W}^+\| \\
\leq\ & \|\Pi - \Pi_W\| + \frac{\|W^+\|}{Z_i}\|v_i - \hat{v}_i\| + \frac{1}{Z_i}\left\|W^+ - \ \hat{W}^+\right\| + \|\hat{W}^+\|\left|\frac{1}{Z_i} - \frac{1}{\hat{Z}_i}\right|
\end{aligned}$$

For bounding $|\frac{1}{Z_i} - \frac{1}{\hat{Z}_i}|$, first observe:

$$\begin{aligned}
& |(Wv_i)^\top \operatorname{Triples}_{\alpha_0}(Wv_i)Wv_i - (\hat{W}\hat{v}_i)^\top \widehat{\operatorname{Triples}}_{\alpha_0}(\hat{W}\hat{v}_i)\hat{W}\hat{v}_i| \\
\leq\ & |(Wv_i)^\top \operatorname{Triples}_{\alpha_0}(Wv_i)Wv_i - (W\hat{v}_i)^\top \operatorname{Triples}_{\alpha_0}(W\hat{v}_i)W\hat{v}_i| \\
& +|(W\hat{v}_i)^\top \operatorname{Triples}_{\alpha_0}(W\hat{v}_i)W\hat{v}_i - (\hat{W}\hat{v}_i)^\top \widehat{\operatorname{Triples}}_{\alpha_0}(\hat{W}\hat{v}_i)\hat{W}\hat{v}_i| \\
\leq\ & c\|v_i - \hat{v}_i\| \max_i \gamma_i + \|W^\top \operatorname{Triples}_{\alpha_0}(W\hat{v}_i)W - \hat{W}^\top \widehat{\operatorname{Triples}}_{\alpha_0}(\hat{W}\hat{v}_i)\hat{W}\|
\end{aligned}$$

where c is a constant and where that last step uses an argument similar to that of Equation 2 (along with the bounds $\|W^\top \tilde{O}\| = 1$, $\|M^\top v_i\| \leq 1$ and $\|M^\top \hat{v}_i\| \leq 1$). Continuing,

$$\begin{aligned}
& |(Wv_i)^\top \operatorname{Triples}_{\alpha_0}(Wv_i)Wv_i - (\hat{W}\hat{v}_i)^\top \widehat{\operatorname{Triples}}_{\alpha_0}(\hat{W}\hat{v}_i)\hat{W}\hat{v}_i| \\
\leq\ & c_1\|v_i - \hat{v}_i\| \max_i \gamma_i + c_2\left(\frac{E_P}{\sqrt{p_{\min}(\alpha_0+2)}\ \sigma_k(\tilde{O})^2} + \frac{E_T}{\sigma_k(\tilde{O})^3}\right) \\
\leq\ & c_3\frac{k^3}{\delta'\sqrt{p_{\min}}}\left(\frac{E_P}{\sqrt{p_{\min}(\alpha_0+2)}\ \sigma_k(\tilde{O})^2} + \frac{E_T}{\sigma_k(\tilde{O})^3}\right)
\end{aligned}$$

using that $\gamma_i \leq 2\frac{1}{\sqrt{p_{\min}(\alpha_0+2)}}$ in the last step (for constants c_1, c_2, c_3).

The fourth term is bounded as follows:

$$
\begin{aligned}
& \|\hat{W}^{+}\| \left| \frac{1}{Z_i} - \frac{1}{\hat{Z}_i} \right| \\
= \;& \|\hat{W}^{+}\| \frac{\alpha_0+2}{2} \; |(Wv_i)^\top \,\text{Triples}_{\alpha_0}(Wv_i)Wv_i - (\hat{W}\hat{v}_i)^\top \widehat{\text{Triples}}_{\alpha_0}(\hat{W}\hat{v}_i)\hat{W}\hat{v}_i| \\
\leq \;& c_1 \|\hat{W}^{+}\| \frac{k^3(\alpha_0+2)}{\delta' \sqrt{p_{\min}}} \left(\frac{E_P}{\sqrt{p_{\min}(\alpha_0+2)}\; \sigma_k(\tilde{O})^2} + \frac{E_T}{\sigma_k(\tilde{O})^3} \right) \\
\leq \;& c_2 \sigma_1(\tilde{O}) \frac{k^3(\alpha_0+2)}{\delta' \sqrt{p_{\min}}} \left(\frac{E_P}{\sqrt{p_{\min}(\alpha_0+2)}\; \sigma_k(\tilde{O})^2} + \frac{E_T}{\sigma_k(\tilde{O})^3} \right)
\end{aligned}
$$

for constants c_1, c_2, c_3.

We have:

$$
\frac{\|W^{+}\|}{Z_i} \leq c \frac{\sigma_1(\tilde{O})}{Z_i} = c\sigma_1(\tilde{O}) \sqrt{\frac{\alpha_0(\alpha_0+1)}{\alpha_i}} \leq c\sigma_1(\tilde{O}) \sqrt{\frac{\alpha_0+1}{p_{\min}}}
$$

(for a constant c), so for the second term:

$$
\frac{\|W^{+}\|}{Z_i} \|v_i - \hat{v}_i\| \leq c\sigma_1(\tilde{O}) \frac{k^3(\alpha_0+2)}{\delta' \sqrt{p_{\min}}} \left(\frac{E_P}{\sqrt{p_{\min}(\alpha_0+2)}\; \sigma_k(\tilde{O})^2} + \frac{E_T}{\sigma_k(\tilde{O})^3} \right)
$$

The remaining terms can be show to be of lower order, so that:

$$
\begin{aligned}
\|O_i - \frac{1}{\hat{Z}_i}\, (\hat{W}^{+})^\top \hat{v}_i\| \;&\leq\; c\sigma_1(\tilde{O}) \frac{k^3(\alpha_0+1)}{\delta' \sqrt{p_{\min}}} \left(\frac{E_P}{\sqrt{p_{\min}(\alpha_0+2)}\; \sigma_k(\tilde{O})^2} + \frac{E_T}{\sigma_k(\tilde{O})^3} \right) \\
&\leq\; c_2 \frac{k^3\sqrt{\alpha_0+2}}{\delta' \sqrt{p_{\min}}} \left(\frac{(\alpha_0+2)^{1/2} E_P}{p_{\min}^{3/2}\; \sigma_k(O)^2} + \frac{(\alpha_0+2)^{3/2} E_T}{p_{\min}^{3/2} \sigma_k(O)^3} \right) \\
&=\; c_2 \frac{k^3(\alpha_0+2)}{p_{\min}^2 \delta'} \left(\frac{E_P}{\sigma_k(O)^2} + \frac{(\alpha_0+2)E_T}{\sigma_k(O)^3} \right)
\end{aligned}
$$

using that $\sigma_k(\tilde{O}) \geq \sigma_k(O)\sqrt{\frac{p_{\min}}{\alpha_0+1}}$ and $\sigma_1(\tilde{O}) \leq \frac{1}{\sqrt{\alpha_0+1}}$. □

C.4 Completing the proof

Proof of Theorem 5.1. Lemma D.1 and the definition of Pairs_{α_0} and $\text{Triples}_{\alpha_0}$ imply that:

$$
\begin{aligned}
\|\widehat{\text{Pairs}}_{\alpha_0} - \text{Pairs}_{\alpha_0}\| \;&\leq\; 3\frac{1+\sqrt{\ln(3/\delta)}}{\sqrt{N}} \\
\|\,\text{Triples}_{\alpha_0}(\eta) - \widehat{\text{Triples}}_{\alpha_0}(\eta)\| \;&\leq\; c\frac{\|\eta\|_2(1+\sqrt{\ln(3/\delta)})}{\sqrt{N}}
\end{aligned}
$$

for a constant c (by expanding out the terms and by using $\delta/3$ results in a total error probability of δ). Hence, we can take $E_P = E_T = c\frac{1+\sqrt{\ln(1/\delta)}}{\sqrt{N}}$. Since $N \geq \left(\frac{(\alpha_0+1)(6+6\sqrt{\ln(3/\delta)})}{p_{\min}\sigma_k(O)^2}\right)^2 \geq$

$\left(\frac{6+6\sqrt{\ln(3/\delta)}}{\sigma_k(\text{Pairs}_{\alpha_0})}\right)^2$, the condition $E_P \le \sigma_k(\text{Pairs}_{\alpha_0})/2$ is satisfied. The proof is completed using Lemma C.6. □

D Tail Inequalities

Lemma D.1 (Lemma A.1 in Anandkumar et al. (2012))**.** *Fix $\delta \in (0,1)$. Let x_1, x_2, x_3 are random variables in which $\|x_1\|, \|x_2\|, \|x_3\|$ are bounded by 1, almost surely. Let $\hat{E}[x_1]$ be the empirical average of N independent copies of x_1; let $\hat{E}[x_1 x_2^\top]$ be the empirical average of N independent copies of $x_1 x_2^\top$; let $\hat{E}[x_1 x_2^\top \langle \eta, x_3\rangle]$. be the empirical average of N independent copies of $x_1 x_2^\top \langle \eta, x_3 \rangle$. Then*

1. $\Pr\left[\|\hat{E}[x_1] - E[x_1]\|_\text{F} \le \frac{1+\sqrt{\ln(1/\delta)}}{\sqrt{N}}\right] \ge 1-\delta$
2. $\Pr\left[\|\hat{E}[x_1 x_2^\top] - E[x_1 x_2^\top]\|_\text{F} \le \frac{1+\sqrt{\ln(1/\delta)}}{\sqrt{N}}\right] \ge 1-\delta$
3. $\Pr\left[\forall \eta \in \mathbb{R}^d,\ \|\hat{E}[x_1 x_2^\top \langle \eta, x_3\rangle] - E[x_1 x_2^\top \langle \eta, x_3\rangle]\|_\text{F} \le \frac{\|\eta\|_2(1+\sqrt{\ln(1/\delta)})}{\sqrt{N}}\right] \ge 1-\delta.$

Lemma D.2 (Dasgupta and Gupta (2003))**.** *Let $\theta \in \mathbb{R}^n$ be a random vector distributed uniformly over $\mathcal{S}^{n-1}$, and fix a vector $v \in \mathbb{R}^n$.*

1. *If $\beta \in (0,1)$, then*
$$\Pr\left[|\langle \theta, v\rangle| \le \|v\| \cdot \frac{1}{\sqrt{n}} \cdot \beta\right] \le \exp\left(\frac{1}{2}(1-\beta^2+\ln\beta^2)\right).$$
2. *If $\beta > 1$, then*
$$\Pr\left[|\langle \theta, v\rangle| \ge \|v\| \cdot \frac{1}{\sqrt{n}} \cdot \beta\right] \le \exp\left(\frac{1}{2}(1-\beta^2+\ln\beta^2)\right).$$

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

E Matrix Perturbation Lemmas

Lemma E.1 (Weyl's theorem; Theorem 4.11, p. 204 in Stewart and Sun (1990))**.** *Let $A, E \in \mathbb{R}^{m \times n}$ with $m \ge n$ be given. Then*
$$\max_{i \in [n]} |\sigma_i(A+E) - \sigma_i(A)| \le \|E\|.$$

Lemma E.2 (Wedin's theorem; Theorem 4.1, p. 260 in Stewart and Sun (1990)))**.** *Let $A, E \in \mathbb{R}^{m \times n}$ with $m \ge n$ be given. Let A have the singular value decomposition*
$$\begin{bmatrix} U_1^\top \\ U_2^\top \\ U_3^\top \end{bmatrix} A \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{bmatrix}.$$

Here, we do not suppose Σ_1 and Σ_2 have singular values in any order. Let $\tilde{A} := A + E$, with analogous singular value decomposition $(\tilde{U}_1, \tilde{U}_2, \tilde{U}_3, \tilde{\Sigma}_1, \tilde{\Sigma}_2, \tilde{V}_1 \tilde{V}_2)$ (again with no ordering to the *singular values). Let* Φ *be the matrix of canonical angles between* $\mathrm{range}(U_1)$ *and* $\mathrm{range}(\tilde{U}_1)$*, and* Θ *be the matrix of canonical angles between* $\mathrm{range}(V_1)$ *and* $\mathrm{range}(\tilde{V}_1)$*. Suppose there exists a* δ *such that:*

$$\min_{i,j} |[\Sigma_1]_{i,i} - [\Sigma_2]_{j,j}| > \delta \ \text{ and } \ \min_{i,i} |[\Sigma_1]_{i,i}| > \delta,$$

then

$$\|\sin \Phi\|_F^2 + \|\sin \Theta\|_F^2 \leq \frac{2\|E\|_F^2}{\delta^2}.$$

Lemma E.3 (Wedin's theorem; Theorem 4.4, p. 262 in Stewart and Sun (1990).)**.** *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* $\mathrm{range}(U_1)$ *and* $\mathrm{range}(\tilde{U}_1)$*, and* Θ *be the matrix of canonical angles between* $\mathrm{range}(V_1)$ *and* $\mathrm{range}(\tilde{V}_1)$*. 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*

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

Lemma E.4. *Let* Θ *be the matrix of canonical angles between* $\mathrm{range}(X)$ *and* $\mathrm{range}(Y)$*. Let* Π_X *and* Π_Y *be the orthogonal projections onto* $\mathrm{range}(X)$ *and* $\mathrm{range}(Y)$*, respectively. We have:*

$$\|\Pi_X - \Pi_Y\| = \|\sin \Theta\|$$

Proof. See Theorem 4.5, p. 92, and Corollary 4.6, p. 93, in Stewart and Sun (1990). □

F Illustrative empirical results

We applied Algorithm 5 to the UCI "Bag of Words" dataset comprised of New York Times articles. This data set has 300000 articles and a vocabular of size $d = 102660$; we set $k = 50$ and $\alpha_0 = 0$. Instead of using a single random θ and obtaining singular vectors of $\hat{W}^\top \operatorname{Triples}_{\alpha_0}(\hat{W}\theta)\hat{W}$, we used the following power iteration to obtain the singular vectors $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k\}$:

$\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k\} \leftarrow$ random orthonormal basis for $\mathbb{R}^k$.
Repeat:

1. For $i = 1, 2, \ldots, k$:

$$\hat{v}_i \leftarrow \hat{W}^\top \operatorname{Triples}_{\alpha_0}(\hat{W}\hat{v}_i)\hat{W}\hat{v}_i.$$

2. Orthonormalize $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k\}$.

The top 25 words (ordered by estimated conditional probability value) from each topic are shown below.

zzz_held	premature	las	sales	million	com	run	school	women
send	guard	como	economic	shares	question	inning	student	team
advisory	zzz_held	los	consumer	public	information	hit	teacher	woman
publication	released	zzz_latin_trade	major	offering	zzz_eastern	game	program	job
released	publication	articulo	home	source	sport	season	official	sport
guard	advisory	telefono	indicator	initial	daily	home	public	cancer
zzz_attn_editor	send	transmiten	weekly	debt	commentary	right	children	look
undatelined	undatelined	fax	order	bond	business	games	high	company
night	zzz_washington_datelined	una	claim	billion	newspaper	zzz_dodger	education	group
advance	zzz_istanbul	del	scheduled	share	separate	left	district	percent
zzz_andrew_pollack	zzz_attn_editor	articulos	listed	quarter	spot	team	parent	girl
zzz_douglas_frantz	zzz_seth_mydan	espanol	dates	revenue	marked	start	college	study
billion	nyt	paises	jobless	market	today	yankees	money	game
zzz_jennifer	zzz_johannesburg	sobre	prices	zzz_calif	zzz_tom_oder	pitcher	test	games
zzz_dirk_johnson	zzz_afghanistan	financial	price	school	holiday	ball	percent	female
zzz_leslie	zzz_jane_perlez	zzz_america_latina	market	zzz_new_york	need	pitch	system	american
cell	zzz_john_broder	notas	leading	cash	staffed	manager	kid	number
zzz_linda	zzz_warren	prohibitivo	retailer	stock	development	lead	federal	season
games	zzz_melbourne	con	economy	percent	toder	night	law	breast
zzz_lee	zzz_lexington	revista	index	securities	client	homer	need	play
zzz_james_brooke	zzz_erik_eckholm	tiene	retail	zzz_credit_suisse_first_boston	eta	field	help	zzz_taliban
zzz_winnipeg	zzz_bernard_simon	economia	spending	deal	directed	play	class	right
deal	substitute	costo	product	contract	additional	ranger	group	part
husband	close	otros	cost	president	reach	win	plan	male
zzz_usc	point	zzz_paris	producer	expected	washington	hitter	black	high

drug	player	article	palestinian	tax	cup	point	yard	percent
patient	zzz_tiger_wood	zzz_new_york	zzz_israel	cut	minutes	game	game	stock
million	won	misstated	zzz_israeli	percent	oil	team	play	market
company	shot	zzz_boston_globe	zzz_yasser_arafat	zzz_bush	water	shot	season	fund
doctor	play	zzz_united_states	peace	billion	add	play	team	investor
companies	round	company	israeli	plan	tablespoon	zzz_laker	touchdown	companies
percent	win	president	israelis	bill	food	season	quarterback	analyst
cost	tournament	campaign	leader	taxes	teaspoon	half	coach	money
program	tour	zzz_clinton	official	million	pepper	lead	defense	investment
health	right	surname	attack	zzz_congress	sugar	games	quarter	economy
care	par	player	zzz_bush	zzz_george_bush	large	quarter	ball	point
billion	final	incorrectly	zzz_west_bank	economy	fat	minutes	field	company
plan	playing	point	zzz_palestinian	money	butter	night	pass	quarter
medical	major	film	violence	income	sauce	left	run	price
treatment	ball	director	security	government	serving	goal	offense	billion
zzz_aid	hit	office	killed	spending	hour	king	line	earning
disease	lead	school	talk	federal	fresh	final	running	prices
cancer	golf	home	military	pay	pan	played	defensive	firm
hospital	guy	misspelled	jewish	republican	taste	scored	zzz_nfl	index
prescription	hole	died	zzz_jerusalem	zzz_white_house	bowl	zzz_kobe_bryant	football	growth
federal	course	information	soldier	zzz_senate	cream	rebound	receiver	zzz_nasdaq
government	game	misidentified	zzz_clinton	zzz_democrat	onion	right	left	shares
product	played	referred	zzz_sharon	sales	serve	win	win	rates
zzz_medicare	night	zzz_washington	minister	zzz_social_security	medium	percent	player	rate
study	set	son	fire	proposal	pound	ball	zzz_giant	interest

zzz_al_gore	zzz_george_bush	car	book	zzz_taliban	com	zzz_bush	court	percent
campaign	president	race	children	attack	www	percent	case	number
president	zzz_al_gore	driver	ages	zzz_afghanistan	site	campaign	law	group
zzz_george_bush	campaign	team	author	official	web	zzz_enron	lawyer	rate
zzz_bush	republican	won	read	military	sites	administration	federal	million
zzz_clinton	zzz_john_mccain	win	newspaper	zzz_u_s	information	president	government	sales
vice	election	racing	web	zzz_united_states	online	zzz_white_house	decision	survey
presidential	zzz_texas	track	writer	terrorist	mail	money	trial	according
million	presidential	season	written	war	internet	plan	zzz_microsoft	study
democratic	political	lap	sales	bin	telegram	republican	right	quarter
night	zzz_enron	point	find	laden	visit	company	judge	average
voter	governor	sport	history	zzz_american	find	million	legal	economy
election	administration	seat	list	zzz_bush	zzz_internet	zzz_republican	ruling	american
vote	democratic	races	word	government	computer	official	attorney	increase
plan	zzz_white_house	road	published	group	org	zzz_texas	death	rose
zzz_bill_bradley	voter	run	school	forces	newspaper	election	system	black
ballot	nation	look	zzz_new_york	zzz_pakistan	offer	show	company	student
zzz_governor_bush	public	right	right	country	free	political	zzz_supreme_court	level
republican	zzz_clinton	zzz_nascar	boy	leader	services	zzz_mccain	election	school
zzz_florida	zzz_republican	drive	writing	american	company	energy	cases	season
right	candidate	zzz_winston_cup	american	afghan	official	zzz_washington	prosecutor	poll
votes	point	owner	reading	troop	list	zzz_united_states	public	newspaper
poll	question	start	game	terrorism	user	voter	zzz_florida	job
court	percent	big	reader	nation	companies	fund	ballot	consumer
candidates	zzz_party	ago	won	zzz_pentagon	customer	zzz_al_gore	states	government

company	show	game	computer	film	team	bill	cell	election
percent	network	games	system	movie	player	zzz_senate	patient	ballot
million	season	season	program	director	season	law	human	vote
business	zzz_nbc	play	zzz_microsoft	play	game	right	research	voter
companies	zzz_cb	goal	mail	character	coach	zzz_white_house	group	campaign
billion	program	king	software	actor	play	zzz_congress	scientist	political
analyst	television	team	window	show	games	vote	zzz_enron	votes
stock	series	won	web	movies	right	member	study	official
quarter	night	player	company	million	league	president	disease	zzz_florida
executive	zzz_new_york	coach	million	part	million	legislation	information	democratic
deal	zzz_abc	played	information	zzz_hollywood	deal	zzz_clinton	found	race
sales	tonight	period	need	look	manager	group	team	zzz_republican
share	hour	left	technology	big	need	zzz_house	public	recount
zzz_enron	look	playing	user	young	contract	republican	doctor	republican
chief	zzz_fox	night	security	music	guy	campaign	government	won
market	air	win	zzz_internet	set	point	federal	death	leader
employees	viewer	right	problem	screen	played	money	cancer	candidate
customer	rating	com	internet	writer	baseball	election	researcher	zzz_al_gore
president	game	playoff	money	television	agent	support	stem	zzz_party
product	early	power	home	making	fan	zzz_republican	official	poll
executives	big	guy	network	love	playing	measure	problem	candidates
financial	talk	zzz_new_york	product	played	job	issue	called	party
earning	event	record	called	producer	free	passed	medical	presidential
operation	hit	shot	help	guy	sport	percent	director	win
cent	award	minutes	number	kind	basketball	billion	question	result

money	police	team	air	family	music	official	companies	president
million	officer	game	water	children	song	government	job	program
fund	official	win	million	home	group	zzz_united_states	worker	zzz_bush
zzz_enron	president	won	high	father	part	zzz_china	company	group
campaign	government	zzz_u_s	building	mother	zzz_new_york	zzz_u_s	business	game
program	attack	play	power	son	company	zzz_american	firm	member
group	case	games	plant	parent	million	country	zzz_new_york	zzz_clinton
plan	told	official	plan	child	band	administration	attack	care
government	office	point	cost	friend	show	zzz_clinton	president	leader
firm	member	run	hour	school	album	million	employees	health
company	public	home	system	boy	companies	nation	plan	zzz_white_house
pay	death	zzz_united_states	wind	wife	record	countries	need	vice
worker	group	sport	part	house	play	president	law	plan
help	zzz_new_york	zzz_new_york	weather	told	right	economic	percent	job
job	chief	attack	area	daughter	business	foreign	customer	children
political	black	tournament	home	kid	look	power	industry	patient
lawyer	lawyer	american	rain	night	artist	chinese	number	executive
member	prosecutor	percent	shower	help	home	zzz_russia	cost	worker
account	security	minutes	front	care	industry	political	terrorist	doctor
effort	building	zzz_olympic	program	left	member	plan	security	school
billion	campaign	final	billion	official	black	meeting	market	decision
employees	night	player	night	room	sound	leader	information	director
financial	hour	company	feet	money	night	trade	help	zzz_congress
question	home	lead	low	hour	called	percent	official	administration
need	found	zzz_washington	miles	job	fan	right	economy	chief

government	season	right	test	file
companies	team	zzz_united_states	zzz_seattle_post_intelligencer	onlytest
political	won	american	zzz_hearst_news_service	sport
country	race	war	zzz_kansas_city	notebook
president	win	student	look	zzz_los_angeles
campaign	attack	look	testing	onlyendpar
leader	home	need	houston	zzz_joe_haakenson_san_gabriel_valley_tribune
business	record	show	ellipses	zzz_anaheim_angel
election	games	home	anthrax	frontend
zzz_bush	zzz_u_s	question	student	zzz_seattle_pi
win	final	black	glories	zzz_seattle_post_intelligencer
war	zzz_clinton	military	mark	zzz_chuck
company	night	left	night	zzz_abcdefg_test
zzz_internet	million	country	rare	added
billion	zzz_olympic	com	zzz_texas	zzz_los_angeles_dodger
race	winning	women	result	read
power	coach	word	risk	zzz_calif
support	championship	put	exam	output
market	patient	zzz_american	system	email
team	playoff	help	scores	internet
democratic	victory	room	missile	zzz_brian_dohn
won	american	zzz_u_s	zzz_washington	files
public	trial	zzz_america	body	zzz_scott_wolf
web	medal	percent	according	wrote
industry	series	job	scientist	consumer