A Spectral Algorithm for Latent Dirichlet Allocation

Anima Anandkumar
University of California
Irvine, CA
a.anandkumar@uci.edu

Dean P. Foster
University of Pennsylvania
Philadelphia, PA
dean@foster.net

Daniel Hsu
Microsoft Research
Cambridge, MA
dahsu@microsoft.com

Sham M. Kakade
Microsoft Research
Cambridge, MA
skakade@microsoft.com

Yi-Kai Liu
National Institute of Standards and Technology*
Gaithersburg, MD
yi-kai.liu@nist.gov

Abstract

Topic modeling is a generalization of clustering that posits that observations (words in a document) are generated by *multiple* latent factors (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-word distributions when only words are observed, and the topics are hidden.

This work provides a simple and efficient learning procedure that is guaranteed to recover the parameters for a wide class of topic models, including Latent Dirichlet Allocation (LDA). For LDA, the procedure correctly recovers both the topic-word distributions and the parameters of the Dirichlet prior over the topic mixtures, using only trigram statistics (*i.e.*, third order moments, which may be estimated with documents containing just three words). The method, called Excess Correlation Analysis, is based on a spectral decomposition of low-order moments via two singular value decompositions (SVDs). Moreover, the algorithm is scalable, since the SVDs are carried out only on $k \times k$ matrices, where k is the number of latent factors (topics) and is typically much smaller than the dimension of the observation (word) space.

1 Introduction

Topic models use latent variables to explain the observed (co-)occurrences of words in documents. They posit that each document is associated with a (possibly sparse) mixture of active topics, and that each word in the document is accounted for (in fact, generated) by one of these active topics. In Latent Dirichlet Allocation (LDA) [1], a Dirichlet prior gives the distribution of active topics in documents. LDA and related models possess a rich representational power because they allow for documents to be comprised of words from several topics, rather than just a single topic. 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 unsupervised estimation procedures for topic models are based on finding maximum likelihood estimates, through either local search or sampling based methods, *e.g.*, Expectation-Maximization [2], Gibbs sampling [3], and variational approaches [4]. Another body of tools is based on matrix factorization [5, 6]. For document modeling, a typical 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 [7], which attempts to match the observed moments with those posited by the model. Our approach does this efficiently through a particular decomposition of the low-order observable moments, which can be extracted using singular value decompositions (SVDs). This method is simple and efficient to implement, and is guaranteed to recover the parameters of a wide class of topic 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 called Excess Correlation Analysis (ECA) based on the low-order (cross) moments of observed variables. These observed variables are assumed to be exchangeable (and, more generally, drawn from a multi-view model). ECA differs from Principal Component Analysis and Canonical Correlation Analysis in that it is based on two singular value decompositions: the first SVD whitens the data (based on the correlation between two observed variables) and the second SVD uses higher-order moments (third- or fourth-order moments) to find directions which exhibit non-Gaussianity, *i.e.*, directions where the moments are in *excess* of those suggested by a Gaussian distribution. The SVDs are performed only on $k \times k$ matrices, where k is the number of latent factors; note that the number of latent factors (topics) k is typically much smaller than the dimension of the observed space d (number of words).

The method is applicable to a wide class of latent variable models including exchangeable and multi-view models. We first consider the class of exchangeable variables with independent latent factors. We show that the (exact) low-order moments permit a decomposition that recovers the parameters for model class, and that this decomposition can be computed using two SVD computations. We then consider LDA and show that the same decomposition of a modified third-order moment correctly recovers both the probability distribution of words under each topic, as well as the parameters of the Dirichlet prior. We note that in order to estimate third-order moments in the LDA model, it suffices for each document to contain at least three words.

While the methods described assume exact moments, it is straightforward to write down the analogue "plug-in" estimators based on empirical moments from sampled data. We provide a simple sample complexity analysis that shows that estimating the third-order moments is not as difficult as it might naïvely seem since we only need a $k \times k$ matrix to be accurate.

Finally, we remark that the moment decomposition can also be obtained using other techniques, including tensor decomposition methods and simultaneous matrix diagonalization methods. Some preliminary experiments illustrating the efficacy of one such method is given in the appendix.

Omitted proofs, and additional results and discussion are provided in the full version of the paper [8].

1.2 Related work

Under the assumption that a single active topic occurs in each document, the work of [9] provides the first provable guarantees for recovering the topic distributions (*i.e.*, the distribution of words under each topic), albeit with a rather stringent separation condition (where the words in each topic are essentially non-overlapping). Understanding what separation conditions permit efficient learning is a natural question; in the clustering literature, a line of work has focussed on understanding the relationship between the separation of the mixture components and the complexity of learning. For clustering, the first provable learnability result [10] was under a rather strong separation condition; subsequent results relaxed [11–18] or removed these conditions [19–21]; roughly speaking, learning under a weaker separation condition is more challenging, both computationally and statistically. For the topic modeling problem in which only a single active topic is present per document, [22] provides an algorithm for learning topics with no separation requirement, but under a certain full rank assumption on the topic probability matrix.

For the case of LDA (where each document may be about multiple topics), the recent work of [23] provides the first provable result under a natural separation condition. The condition requires that each topic be associated with "anchor words" that only occur in documents about that topic. This is a significantly milder assumption than the one in [9]. Under this assumption, [23] provide the first provably correct algorithm for learning the topic distributions. Their work also justifies the use of non-negative matrix (NMF) as a provable 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, [23] provides results for certain correlated topic models. Our approach makes further progress on this problem by relaxing the need for this separation condition and establishing a much simpler procedure for parameter estimation.

The underlying approach we take is a certain diagonalization technique of the observed moments. We know of at least three different settings which use this idea for parameter estimation.

The work in [24] uses eigenvector methods for parameter estimation in discrete Markov models involving multinomial distributions. The idea has been extended to other discrete mixture models such as discrete hidden Markov models (HMMs) and mixture models with a single active topic in each document (see [22, 25, 26]). For such single topic models, the work in [22] demonstrates the generality of the eigenvector method and the irrelevance of the noise model for the observations, making it applicable to both discrete models like HMMs as well as certain Gaussian mixture models.

Another set of related techniques is the body of algebraic methods used for the problem of blind source separation [27]. These approaches are tailored for independent source separation with additive noise (usually Gaussian) [28]. Much of the literature focuses on understanding the effects of measurement noise, which often requires more sophisticated algebraic tools (typically, knowledge of noise statistics or the availability of multiple views of the latent factors is not assumed). These algebraic ideas are also used by [29, 30] for learning a linear transformation (in a noiseless setting) and provides a different provably correct algorithm, based on a certain ascent algorithm (rather than joint diagonalization approach, as in [27]), and a provably correct algorithm for the noisy case was recently obtained by [31].

The underlying insight exploited by our method is the presence of exchangeable (or multi-view) variables (*e.g.*, multiple words in a document), which are drawn independently conditioned on the same hidden state. This allows us to exploit ideas both from [24] and from [27]. In particular, we show that the "topic" modeling problem exhibits a rather simple algebraic solution, where only two SVDs suffice for parameter estimation.

Furthermore, the exchangeability assumption permits us to have an *arbitrary* noise model (rather than an additive Gaussian noise, which is not appropriate for multinomial and other discrete distributions). A key technical contribution is that we show how the basic diagonalization approach can be adapted for Dirichlet models, through a rather careful construction. This construction bridges the gap between the single topic models (as in [22, 24]) and the independent latent factors 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.*, [16, 18, 32, 33]). These previous works essentially use variants of canonical correlation analysis [34] between the two views. This work follows [22] in showing that having a third view of the data permits rather simple estimation procedures with guaranteed parameter recovery.

2 The independent latent factors and LDA models

Let $h = (h_1, h_2, \dots, h_k) \in \mathbb{R}^k$ be a random vector specifying the latent factors (*i.e.*, the hidden state) of a model, where h_i is the value of the i-th factor. Consider a sequence of *exchangeable* random vectors $x_1, x_2, x_3, x_4, \dots \in \mathbb{R}^d$, which we take 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. Furthermore, assume there exists a matrix $O \in \mathbb{R}^{d \times k}$ such that

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

for each $v \in \{1, 2, 3, \dots\}$. Throughout, we assume the following condition.

Condition 2.1. O has full column 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*. Note that at this stage, we have not made any assumptions on the noise model; it need not be additive nor even independent of h.

2.1 Independent latent factors model

In the independent latent factors model, we assume h has a product distribution, *i.e.*, $h_1, h_2, \ldots, h_k$ are independent. 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 generalizes 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). 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. 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 $\alpha \in \mathbb{R}^k_{>0}$ (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") characterizes the concentration 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, almost surely, exactly 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). The i-th column O_i of O is a probability vector representing the distribution over words for the i-th topic. The sampling process for a document is as follows. First, the topic mixture h is drawn from the Dirichlet distribution. Then, the v-th word in the document (for $v = 1, 2, \ldots$) is generated by: (i) drawing $t \in [k] := \{1, 2, \ldots k\}$ according to the discrete distribution specified by h, then (ii) drawing x_v according to the discrete distribution specified by O_t (the t-th column of O). Note that x_v is independent of h given t. For this model to fit in our setting, we use the "one-hot" encoding for x_v from [22]: $x_v \in \{0,1\}^d$ with $[x_v]_j = 1$ iff the v-th word in the document is the j-th word in the vocabulary. Observe that

$$\mathbb{E}[x_v|h] = \sum_{i=1}^{k} \Pr[t = i|h] \cdot \mathbb{E}[x_v|t = i, h] = \sum_{i=1}^{k} h_i \cdot O_i = Oh$$

as required. Again, note that the noise model is not additive.

3 Excess Correlation Analysis (ECA)

We now present efficient algorithms for exactly recovering O from low-order moments of the observed variables. 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

Algorithm 1 ECA, with skewed factors

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

1. **Dimensionality reduction:** Find a matrix $U \in \mathbb{R}^{d \times k}$ such that
$$\mathrm{range}(U) = \mathrm{range}(\mathrm{Pairs}).$$
(See Remark 1 for a fast procedure.)
2. **Whiten:** Find $V \in \mathbb{R}^{k \times k}$ so $V^\top (U^\top \,\mathrm{Pairs}\, U) V$ is the $k \times k$ identity matrix. Set:
$$W = UV.$$
3. **SVD:** Let Ξ be the set of left singular vectors of
$$W^\top \,\mathrm{Triples}(W\theta) W$$
corresponding to *non-repeated* singular values (*i.e.*, singluar values with multiplicity one).
4. **Reconstruct:** Return the set
$$\widehat{O} := \{(W^+)^\top \xi \ : \ \xi \in \Xi\}.$$

out on higher-order moments. We start with the case of independent factors, as these algorithms make the basic diagonalization approach clear.

Throughout, we use A^+ to denote the Moore-Penrose pseudo-inverse.

3.1 Independent and skewed latent factors

Define the following moments:

$$\mu := \mathbb{E}[x_1], \quad \mathrm{Pairs} := \mathbb{E}[(x_1 - \mu) \otimes (x_2 - \mu)], \quad \mathrm{Triples} := \mathbb{E}[(x_1 - \mu) \otimes (x_2 - \mu) \otimes (x_3 - \mu)]$$

(here $\otimes$ denotes the tensor product, so $\mu \in \mathbb{R}^d$, $\mathrm{Pairs} \in \mathbb{R}^{d \times d}$, and $\mathrm{Triples} \in \mathbb{R}^{d \times d \times d}$). It is convenient to project Triples to matrices as follows:

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

Roughly speaking, we can think of $\mathrm{Triples}(\eta)$ as a re-weighting of a cross covariance (by $\langle \eta, x_3 - \mu \rangle$).

Note that the matrix O is only identifiable up to permutation and scaling of columns. To see the latter, observe the distribution of any x_v is unaltered if, for any $i \in [k]$, we multiply the i-th column of O by a scalar $c \neq 0$ and divide the variable h_i by the same scalar c. Without further assumptions, we can only hope to recover a certain canonical form of O, defined as follows.

Definition 1 (Canonical form)**.** We say O is in a *canonical form* (relative to h) if, for each $i \in [k]$,

$$\sigma_i^2 := \mathbb{E}[(h_i - \mathbb{E}[h_i])^2] = 1.$$

The transformation $O \leftarrow O \,\mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$ (and a rescaling of h) places O in canonical form relative to h, and the distribution over $x_1, x_2, x_3, \ldots$ is unaltered. In canonical form, O is unique up to a signed column permutation.

Let $\mu_{i,p} := \mathbb{E}[(h_i - \mathbb{E}[h_i])^p]$ denote the p-th central moment of h_i, so the variance and skewness of h_i are given by $\sigma_i^2 := \mu_{i,2}$ and $\gamma_i := \mu_{i,3}/\sigma_i^3$. The first result considers the case when the skewness is non-zero.

Theorem 3.1 (Independent and skewed factors)**.** *Assume Condition 2.1 and $\sigma_i^2 > 0$ for each $i \in [k]$. Under the independent latent factor model, the following hold.*

- No False Positives*: For all $\theta \in \mathbb{R}^k$, Algorithm 1 returns a subset of the columns of O, in canonical form up to sign.*
- Exact Recovery*: Assume $\gamma_i \neq 0$ for each $i \in [k]$. If $\theta \in \mathbb{R}^k$ is drawn uniformly at random from the unit sphere $\mathcal{S}^{k-1}$, then with probability 1, Algorithm 1 returns all columns of O, in canonical form up to sign.*

The proof of this theorem relies on the following lemma.

Lemma 3.1 (Independent latent factors moments)**.** *Under the independent latent factor model,*

$$\begin{aligned} \text{Pairs} &= \sum_{i=1}^{k} \sigma_i^2 \, O_i \otimes O_i = O \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2) O^\top, \\ \text{Triples} &= \sum_{i=1}^{k} \mu_{i,3} \, O_i \otimes O_i \otimes O_i, \quad \text{Triples}(\eta) \;=\; O \operatorname{diag}(O^\top \eta) \operatorname{diag}(\mu_{1,3}, \mu_{2,3}, \ldots, \mu_{k,3}) O^\top. \end{aligned}$$

Proof. The model assumption $\mathbb{E}[x_v|h] = Oh$ implies $\mu = O\mathbb{E}[h]$. Therefore $\mathbb{E}[(x_v - \mu)|h] = O(h - \mathbb{E}[h])$. Using the conditional independence of x_1 and x_2 given h, and the fact that h has a product distribution,

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

An analogous argument gives the claims for Triples and $\text{Triples}(\eta)$. □

Proof of Theorem 3.1. Assume O is in canonical form with respect to h. By Condition 2.1, $U^\top \text{Pairs}\, U \in \mathbb{R}^{k \times k}$ is full rank and hence positive definite. Thus the whitening step is possible, and $M := W^\top O$ is orthogonal. Observe that $W^\top \text{Triples}(W\theta) W = MDM^\top$, where $D := \operatorname{diag}(M^\top \theta) \operatorname{diag}(\gamma_1, \gamma_2, \ldots, \gamma_k)$. Since M is orthogonal, the above is an eigendecomposition of $W^\top \text{Triples}(W\theta) W$, and hence the set of left singular vectors corresponding to non-repeated singular values are uniquely defined up to sign. Each such singular vector ξ is of the form $s_i M e_i = s_i W^\top O e_i = s_i W^\top O_i$ for some $i \in [k]$ and $s_i \in \{\pm 1\}$, so $(W^+)^\top \xi = s_i W (W^\top W)^{-1} W^\top O_i = s_i O_i$ (because $\operatorname{range}(W) = \operatorname{range}(U) = \operatorname{range}(O)$).

If θ is drawn uniformly at random from $\mathcal{S}^{k-1}$, then so is $M^\top \theta$. In this case, almost surely, the diagonal entries of D are unique (provided that each $\gamma_i \neq 0$), and hence every singular value of $W^\top \text{Triples}(W\theta) W$ is non-repeated. □

Remark 1 (Finding $\operatorname{range}(\text{Pairs})$ efficiently). Let $\Theta \in \mathbb{R}^{d \times k}$ be a random matrix with entries sampled independently from the standard normal distribution, and set $U := \text{Pairs}\, \Theta$. Then with probability 1, $\operatorname{range}(U) = \operatorname{range}(\text{Pairs})$.

It is easy to extend Algorithm 1 to kurtotic sources where $\kappa_i := (\mu_{i,4}/\sigma_i^4) - 3 \neq 0$ for each $i \in [k]$, simply by using fourth-order cumulants in places of $\text{Triples}(\eta)$. The details are given in the full version of the paper.

3.2 Latent Dirichlet Allocation

Now we turn to LDA where h has a Dirichlet density. Even though the distribution on h is proportional to the product $h_1^{\alpha_1 - 1} h_2^{\alpha_2 - 1} \cdots h_k^{\alpha_k - 1}$, the h_i are not independent because h is constrained to live in the simplex. These mild dependencies suggest using a certain correction of the moments with ECA.

We assume α_0 is known. Knowledge of $\alpha_0 = \alpha_1 + \alpha_2 + \cdots + \alpha_k$ is significantly weaker than having full knowledge of the entire parameter vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$. A common practice is to specify the entire parameter vector α in a homogeneous manner, with each component being identical (see [35]). Here, we need only specify the sum, which allows for arbitrary inhomogeneity in the prior.

Denote the mean and a modified second moment by

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

and a modified third moment as

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

Algorithm 2 ECA for Latent Dirichlet Allocation

Input: vector $\theta \in \mathbb{R}^k$; the modified moments $\mathrm{Pairs}_{\alpha_0}$ and $\mathrm{Triples}_{\alpha_0}$.

1–3. Execute steps 1–3 of Algorithm 1 with $\mathrm{Pairs}_{\alpha_0}$ and $\mathrm{Triples}_{\alpha_0}$ in place of Pairs and Triples.

4. **Reconstruct and normalize:** Return the set

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

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

Remark 2 (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 (*i.e.*, each document is about just one topic). In this case, the modified moments tend to the raw (cross) moments:

$$\lim_{\alpha_0 \to 0} \mathrm{Pairs}_{\alpha_0} = \mathbb{E}[x_1 \otimes x_2], \quad \lim_{\alpha_0 \to 0} \mathrm{Triples}_{\alpha_0} = \mathbb{E}[x_1 \otimes x_2 \otimes x_3].$$

Note that the one-hot encoding of words in x_v implies that

$$\mathbb{E}[x_1 \otimes x_2] = \sum_{1 \le i,j \le d} \Pr[x_1 = e_i, x_2 = e_j]\, e_i \otimes e_j = \sum_{1 \le i,j \le d} \Pr[\text{1st word} = i, \text{2nd word} = j]\, e_i \otimes e_j,$$

(and a similar expression holds for $\mathbb{E}[x_1 \otimes x_2 \otimes x_3]$), so these raw moments in the limit $\alpha_0 \to 0$ are precisely the joint probabilitiy tables of words across all documents.

At the other extreme $\alpha_0 \to \infty$, the modified moments tend to the central moments:

$$\lim_{\alpha_0 \to \infty} \mathrm{Pairs}_{\alpha_0} = \mathbb{E}[(x_1 - \mu) \otimes (x_2 - \mu)], \quad \lim_{\alpha_0 \to \infty} \mathrm{Triples}_{\alpha_0} = \mathbb{E}[(x_1 - \mu) \otimes (x_2 - \mu) \otimes (x_3 - \mu)]$$

(to see this, expand the central moment and use 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).

Theorem 3.2 (Latent Dirichlet Allocation)**.** *Assume Condition 2.1 holds. Under the LDA model, the following hold.*

- No False Positives*: For all $\theta \in \mathbb{R}^k$, Algorithm 2 returns a subset of the columns of O.*
- Topic Recovery*: If $\theta \in \mathbb{R}^k$ is drawn uniformly at random from the unit sphere $\mathcal{S}^{k-1}$, then with probability* 1*, Algorithm 2 returns all columns of O.*
- Parameter Recovery*: The Dirichlet parameter α satisfies $\alpha = \alpha_0(\alpha_0 + 1) O^+ \mathrm{Pairs}_{\alpha_0} (O^+)^\top \vec{1}$, where $\vec{1} \in \mathbb{R}^k$ is a vector of all ones.*

The proof relies on the following lemma.

Lemma 3.2 (LDA moments)**.** *Under the LDA model,*

$$\begin{aligned}
\mathrm{Pairs}_{\alpha_0} &= \frac{1}{(\alpha_0 + 1)\alpha_0} O \operatorname{diag}(\alpha) O^\top, \\
\mathrm{Triples}_{\alpha_0}(\eta) &= \frac{2}{(\alpha_0 + 2)(\alpha_0 + 1)\alpha_0} O \operatorname{diag}(O^\top \eta) \operatorname{diag}(\alpha) O^\top.
\end{aligned}$$

The proof of Lemma 3.2 is similar to that of Lemma 3.1, except here we must use the specific properties of the Dirichlet distribution to show that the corrections to the raw (cross) moments have the desired effect.

Proof of Theorem 3.2. Note that with the rescaling $\tilde{O} := \frac{1}{\sqrt{(\alpha_0+1)\alpha_0}} O \operatorname{diag}(\sqrt{\alpha_1}, \sqrt{\alpha_2}, \ldots, \sqrt{\alpha_k})$, we have that $\mathrm{Pairs}_{\alpha_0} = \tilde{O}\tilde{O}^\top$. This is akin to $\tilde{O}$ being in canonical form as per the skewed factor model of Theorem 3.1. Now the proof of the first two claims is the same as that of Theorem 3.1; 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 $\mathrm{Pairs}_{\alpha_0}$. $\square$

Remark 3 (Limiting behaviors). ECA seamlessly interpolates between the single topic model ($\alpha_0 \to 0$) of [22] and the skewness-based ECA, Algorithm 1 ($\alpha_0 \to \infty$).

4 Discussion

4.1 Sample complexity

It is straightforward to derive a "plug-in" variant of Algorithm 2 based on empirical moments rather than exact population moments. The empirical moments are formed using the word co-occurrence statistics for documents in a corpus. The following theorem shows that the empirical version of ECA returns accurate estimates of the topics. The details and proof are left to the full version of the paper.

Theorem 4.1 (Sample complexity for LDA)**.** *There exist universal constants $C_1, C_2 > 0$ such that the following hold. 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 $N \geq C_1 \cdot ((\alpha_0+1)/(p_{\min}\sigma_k(O)^2))^2$ independent samples of x_1, x_2, x_3 in the LDA model, which are used to form empirical moments $\widehat{\mathrm{Pairs}}_{\alpha_0}$ and $\widehat{\mathrm{Triples}}_{\alpha_0}$. With high probability, the plug-in variant of Algorithm 2 returns a set $\{\hat{O}_1, \hat{O}_2, \ldots \hat{O}_k\}$ such that, for some permutation σ of $[k]$,*

$$\|O_i - \hat{O}_{\sigma(i)}\|_2 \leq C_2 \cdot \frac{(\alpha_0+1)^2 k^3}{p_{\min}^2 \sigma_k(O)^3 \sqrt{N}}, \quad \forall i \in [k].$$

4.2 Alternative decomposition methods

Algorithm 1 is a theoretically efficient and simple-to-state method for obtaining the desired decomposition of the tensor $\mathrm{Triples} = \sum_{i=1}^k \mu_{i,3} O_i \otimes O_i \otimes O_i$ (a similar tensor form for $\mathrm{Triples}_{\alpha_0}$ in the case of LDA can also be given). However, in practice the method is not particularly stable, due to the use of internal randomization to guarantee strict separation of singular values. It should be noted that there are other methods in the literature for obtaining these decompositions, for instance, methods based on simultaneous diagonalizations of matrices [36] as well as direct tensor decomposition methods [37]; and that these methods can be significantly more stable than Algorithm 1. In particular, very recent work in [37] shows that the structure revealed in Lemmas 3.1 and 3.2 can be exploited to derive very efficient estimation algorithms for all the models considered here (and others) based on a tensor power iteration. We have used a simplified version of this tensor power iteration in preliminary experiments for estimating topic models, and found the results (Appendix A) to be very encouraging, especially due to the speed and robustness of the algorithm.

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 preliminary results (in [23]) and early insights into this problem with us. Part of this work was completed while all authors were at Microsoft Research New England. AA is supported in part by the NSF Award CCF-1219234, AFOSR Award FA9550-10-1-0310 and the ARO Award W911NF-12-1-0404.

References

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

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

[3] 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.

[4] M.D. Hoffman, D.M. Blei, and F. Bach. Online learning for latent dirichlet allocation. In *NIPS*, 2010.

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

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

[7] K. Pearson. Contributions to the mathematical theory of evolution. *Phil. Trans. of the Royal Society, London, A.*, 1894.

[8] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y.-K. Liu. Two svds suffice: spectral decompositions for probabilistic topic models and latent dirichlet allocation, 2012. arXiv:1204.6703.

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

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

[11] S. Dasgupta and L. Schulman. A two-round variant of em for gaussian mixtures. In *UAI*, 2000.

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

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

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

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

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

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

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

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

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

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

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

[23] S. Arora, , R. Ge, and A. Moitra. Learning topic models — going beyond svd. In *FOCS*, 2012.

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

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

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

[27] 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.

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

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

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

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

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

[33] Sham M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysis. In *COLT*, 2007.

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

[35] 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.

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

[37] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and T. Telgarsky. Tensor decompositions for learning latent variable models, 2012. arXiv:1210.7559.

A Illustrative empirical results

We applied a variant of Algorithm 2 to the UCI "Bag of Words" dataset comprised of New York Times articles. This data set has 300000 articles and a vocabulary of size $d = 102660$; we set $k = 50$ and $\alpha_0 = 0$. Following [37], 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