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

---

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

1

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 $h = (h_1, h_2, \ldots, h_k) \in \mathbb{R}^k$ 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 $x_1, x_2, x_3, x_4, \ldots \in \mathbb{R}^d$, which we take to be the observed variables. Assume throughout that $d \geq k$; that $x_1, x_2, x_3, x_4, \ldots \in \mathbb{R}^d$ 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 $v \in \{1, 2, 3, \ldots\}$. 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 $\alpha \in \mathbb{R}_{>0}^k$ (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") 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

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

(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 Ξ be the set of left singular vectors of

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

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:

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

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

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

Roughly speaking, we can think of $\text{Triples}(\eta)$ 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 $O \leftarrow O \, \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$ (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,*

$$\text{Pairs} \;=\; \sum_{i=1}^{k} \sigma_i^2 \, O_i \otimes O_i = O \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2) O^\top,$$

$$\text{Triples} \;=\; \sum_{i=1}^{k} \mu_{i,3} \, O_i \otimes O_i \otimes O_i, \quad \text{Triples}(\eta) \;=\; O \operatorname{diag}(O^\top \eta) \operatorname{diag}(\mu_{1,3}, \mu_{2,3}, \ldots, \mu_{k,3}) O^\top.$$

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,

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

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

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. $\qquad\square$

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

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

Algorithm 2 ECA for Latent Dirichlet Allocation

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

1–3. Execute steps 1–3 of Algorithm 1 with Pairs_{α_0} and $\text{Triples}_{\alpha_0}$ 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:

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

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

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

(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:

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

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

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

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

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 $\text{Pairs}_{\alpha_0} = \tilde{O}\tilde{O}^\top$. This is akin to $\tilde{O}$ being in canonical form as per the skewed factor

7

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.

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.