Algorithmica (2015) 72:193–214
DOI 10.1007/s00453-014-9909-1

A Spectral Algorithm for Latent Dirichlet Allocation

Anima Anandkumar · Dean P. Foster · Daniel Hsu · Sham M. Kakade · Yi-Kai Liu

Received: 1 October 2013 / Accepted: 12 June 2014 / Published online: 3 July 2014

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. The increased representational power comes at the cost of a more challenging unsupervised learning problem for 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 multi-view models and 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

Preliminary versions of this article appeared as [3,4].

A. Anandkumar
University of California, Irvine, Irvine, CA, USA
e-mail: a.anandkumar@uci.edu

D. P. Foster
Yahoo! Labs, New York, NY, USA
e-mail: dean@foster.net

D. Hsu (✉)
Columbia University, New York, NY, USA
e-mail: djhsu@cs.columbia.edu

S. M. Kakade
Microsoft Research, Cambridge, MA, USA
e-mail: skakade@microsoft.com

Y.-K. Liu
National Institute of Standards and Technology, Gaithersburg, MD, USA
e-mail: yi-kai.liu@nist.gov

three words). The method is based on an efficiently computable orthogonal tensor decomposition of low-order moments.

Keywords Topic models · Mixture models · Method of moments · Latent Dirichlet allocation

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) [12], 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 [43], Gibbs sampling [22], and variational approaches [12]. Another body of tools is based on matrix factorization [26,36]. For document modeling, a typical goal is to form a sparse decomposition of a term-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 [42], 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 an orthogonal tensor decomposition. 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 a spectral approach to topic model estimation based on decomposing the low-order (cross) moments of observed variables. The approach differs from other spectral methods (e.g., those based on Principal Component Analysis and Canonical Correlation Analysis) in that it is based on an orthogonal tensor decomposition of a $k \times k \times k$ third-order moment tensor, where k is the number of latent factors (topics). In many applications, 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 low-order moment tensors possess a decomposition known as a canonical polyadic decomposition [34] (Lemma 1 and Lemma 2 in Sect. 3.1). We then consider LDA and show that, even though it does not directly possess an independent latent factor, it *nearly* does so in a rigorous sense, and hence a simple combination of lower-order moments has the required decomposition just as in the previous case (Lemma 3 in Sect. 3.2). Given these moments from either the independent latent factors model or LDA, a simple and computationally efficient algebraic procedure [23] recovers the model parameters exactly under a mild rank condition (Theorem 2 in Sect. 3.3). Informally, we have the following.

Theorem 1 (Informal) *Given low-order (*i.e.*, order* ≤ 3 *or* ≤ 4*) moments from either the independent latent factors model or LDA satisfying a rank condition, there is a polynomial time randomized algorithm that returns the model parameters up to scaling and permutation.*

We describe the parameter recovery procedures assuming exact moments as input, but it is straightforward to write down analogous "plug-in" estimators that use estimates of these moments based on sampled data. We do just this using a particular tensor decomposition procedure from [5], and analyze the error of the resulting parameter estimates in terms of the errors in the moment estimates (Theorem 3).

1.2 Related Work

Under the assumption that a single active topic occurs in each document, the work of [41] provides the first theoretical 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 learnability result [19] was under a rather strong separation condition; subsequent results relaxed [1,10,16,17,20,33,45] or removed these conditions [11,28,32,38]; 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, [6] 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 [8] provides the first theoretical 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 [41]. Under this assumption, [8] provides the first rigorously analyzed 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 technique, though, prior to this work, formal guarantees as such were rather limited). Furthermore, [8] provides results for certain correlated topic models. Our approach makes further progress on this problem by relaxing the need for this separation condition and providing a simpler parameter estimation procedure.

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 algebraic tensor decomposition technique of [23] (see also [35,37])[1] was rediscovered by [15] for parameter estimation in discrete Markov models. 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 [6,29,39]). For such single topic models, the work in [6] 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 (see also [28]).

Another set of related techniques is the body of algebraic methods used for the problem of blind source separation [14]. These approaches are tailored for independent source separation with additive noise (usually Gaussian) [18]. 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 ideas are also used by [21,40] for learning a linear transformation (in a noiseless setting) and provides a different algorithm, based on a certain ascent algorithm (rather than joint diagonalization approach, as in [14]), and a provably correct algorithm for the noisy case was only recently obtained [9].

The models we consider, including LDA, are distinguished by the presence of exchangeable (or multi-view) variables (e.g., multiple words in a document), drawn independently conditioned on the same hidden state. This allows us to exploit ideas from the aforementioned works [14,15,23]. In particular, we show that the topic modeling problem exhibits a similarly simple algebraic solution as in previous works. 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 mixture models and independent latent factors models.

The multi-view approach has been exploited in previous works for semi-supervised learning and for learning mixtures of well-separated distributions (e.g., [7,16,17,31]). These previous works essentially use variants of canonical correlation analysis [27] between the two views. This present work follows [6] in showing that having a third view of the data permits rather simple estimation procedures for parameter recovery without separation conditions.

[1] The technique of [23] is actually attributed to Robert Jennrich.

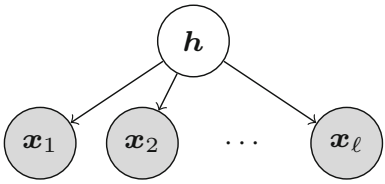

Fig. 1 Directed graphical model representation of the multi-view models

A preliminary version of this work [3,4] proposed a method based on using two singular value decompositions—essentially a symmetrized version of the algorithm from [6]. However, that method is not particularly robust, as it depends on randomization to create separation between eigenvalues of a particular matrix. In this article, we propose to use the robust orthogonal tensor decomposition method from [5], which is more robust than the previous method and still computationally efficient.

2 The Multi-view Models

Let $\boldsymbol{h} = (h_1, h_2, \ldots, h_k) \in \mathbb{R}^k$ be a random vector specifying the latent factors (i.e., $\boldsymbol{h}$ is the hidden state) in a model, where h_i is the value of the ith factor. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell \in \mathbb{R}^d$ be random vectors which we take to be observable. Throughout, we assume $\ell \geq 3$, and the reason will become clear in the sequel. The primary modeling assumption is that the observable random vectors and latent factors satisfy the following conditional independence and linearity condition; we assume this condition holds the remainder of this paper.

Condition 1 [Multi-view model] *The observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell$ are conditionally independent given $\boldsymbol{h}$. Moreover, for each $v \in [\ell] := \{1, 2, \ldots, \ell\}$, there exists a matrix $\boldsymbol{O}^{(v)} \in \mathbb{R}^{d \times k}$ such that*

$$\mathbb{E}[\boldsymbol{x}_v | \boldsymbol{h}] = \boldsymbol{O}^{(v)} \boldsymbol{h}.$$

The conditional independence assumption from Condition 1 is depicted in the directed graphical model in Fig. 1. This model generalizes the multi-view model from [6] in that $\boldsymbol{h}$ is *not* assumed to only take values in $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k\}$, where $\boldsymbol{e}_i \in \{0, 1\}^k$ is the ith coordinate basis vector. We note that at this stage, we have not made any assumptions on the noise model; it need not be additive nor even independent of $\boldsymbol{h}$.[2]

We also assume throughout that the following rank condition on the matrices $\boldsymbol{O}^{(v)}$ from Condition 1.

Condition 2 (Rank condition) *Each $\boldsymbol{O}^{(v)}$ has full column rank.*

Condition 2 allows for the identifiability of the columns of $\boldsymbol{O}^{(v)}$ and was used in previous works on parameter estimation [6,8,9,15,21,28,39,40].

As shown in [6] (for the case where $\boldsymbol{h}$ has support only on k points), the multi-view model captures a number of interesting statistical models, including certain Gaussian

[2] By additive noise, we mean a model in which $\boldsymbol{x}_v = \boldsymbol{O}^{(v)} \boldsymbol{h} + \boldsymbol{\eta}_v$ where $\boldsymbol{\eta}_v$ is a zero-mean random vector independent of $\boldsymbol{h}$.

mixture models and HMMs. In our setting, $\boldsymbol{h}$ is allowed to have a more general distribution, thus enhancing the flexibility of the model. Our goal in this work is to given estimators of the matrices $\boldsymbol{O}^{(v)}$ (and possibly parameters related to the latent factor $\boldsymbol{h}$), solely from repeated observations (independent copies of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell$).

We now consider two specializations of this multi-view model in which the different views $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell$ are naturally *exchangeable*. For these cases, the conditional means of the different views are the same. That is, the following condition holds.

Condition 3 $\boldsymbol{O}^{(v)} = \boldsymbol{O}$ *for all* $v \in [\ell]$.

In the context of topic models, the common matrix $\boldsymbol{O}$ is referred to as the *topic matrix*, as it specifies parameters associated with each topic in the model. Borrowing this terminology, we generally refer to all $\boldsymbol{O}^{(v)}$ as topic matrices.

2.1 Independent Latent Factors Model

In the independent latent factors model, we assume $\boldsymbol{h}$ has a product distribution, *i.e.*, $h_1, h_2, \ldots, h_k$ are independent. In the case where the $\boldsymbol{x}_v$ are deterministic linear functions of $\boldsymbol{h}$, (*i.e.*, $\boldsymbol{x}_v = \boldsymbol{O}\boldsymbol{h}$), the model reduces to the noiseless ICA model of [30], which was reinterpreted as a parallelepiped learning problem in [21,40].

Two important examples of this setting with noise are as follows.

Mixture of Gaussians Suppose $\boldsymbol{x}_v = \boldsymbol{O}\boldsymbol{h} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is zero-mean Gaussian noise and $\boldsymbol{h}$ has a product distribution over $\{0, 1\}^k$. Here, the ith column $\boldsymbol{o}_i$ of $\boldsymbol{O}$ can be considered to be the mean of the ith Gaussian component. This generalizes the classic mixture of k Gaussians, as the model now permits any number of Gaussians to be responsible for generating an observation (i.e., $\boldsymbol{h}$ is permitted to be any of the 2^k vectors on the hypercube, while in the classic mixture problem, only one component is responsible). Alternatively, the model can be viewed as a classical mixture of 2^k Gaussians, where the mean of a component $S \in 2^{[k]}$ is $\sum_{i \in S} \boldsymbol{o}_i$ and its mixing weight is $\Pr(h_i = 1 \forall i \in S)$ [9]. Note that $\boldsymbol{\eta}$ is allowed to be heteroskedastic (i.e., the noise may depend on $\boldsymbol{h}$), so the Gaussians need not have the same covariance.

Poisson topic model This is a model proposed by Canny [13]. Suppose $[\boldsymbol{O}\boldsymbol{h}]_j$ specifies the Poisson rate of counts for $[\boldsymbol{x}_v]_j$. For example, $\boldsymbol{x}_v$ could be a vector of word counts in the vth sentence of a document. Here, $\boldsymbol{O}$ 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 ith column of $\boldsymbol{O}$). The linearity assumption in Condition 1 is satisfied as $\mathbb{E}[\boldsymbol{x}_v|\boldsymbol{h}] = \boldsymbol{O}\boldsymbol{h}$ (note the noise is not additive in this case, in contrast to the mixture of Gaussians example). Here, multiple topics may be responsible for generating the words in each sentence. This model similar to LDA, except for the fact that here, $\boldsymbol{h}$ has a product distribution (whereas in LDA, $\boldsymbol{h}$ is a probability vector).

2.2 The Dirichlet Model

Now suppose the hidden state $\boldsymbol{h}$ is a probability distribution itself, with a density specified by the Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k) \in$

$\mathbb{R}^k_{>0}$ ($\boldsymbol{\alpha}$ is a strictly positive real vector); this assumption is written as $\boldsymbol{h} \sim \mathrm{Dir}(\boldsymbol{\alpha})$. We think of $\boldsymbol{h} \in \Delta^{k-1}$ as a distribution over topics; here, Δ^{k-1} denotes the probability simplex of distributions over k outcomes. The density of $\boldsymbol{h}$, the Dirichlet density, is

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

where $Z(\boldsymbol{\alpha}) := \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$ 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 $\boldsymbol{h}$ is 1, and the rest are 0).

If the $\boldsymbol{x}_v$ are deterministic linear functions of $\boldsymbol{h}$ (i.e., $\boldsymbol{x}_v = \boldsymbol{O}\boldsymbol{h}$), then the model can be viewed as the problem of learning a certain class of distributions over a simplex with vertices $\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_k$. The special case of a uniform distribution over a simplex (a problem suggested in [21]) is obtained when $\boldsymbol{\alpha} = (1, 1, \ldots, 1)$.

Latent Dirichlet Allocation LDA makes the further assumption that each random vector $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell$ takes on discrete values out of d outcomes (e.g., $\boldsymbol{x}_v$ represents what the vth word in a document is, so d represents the number of words in a vocabulary). The ith column $\boldsymbol{o}_i$ of $\boldsymbol{O}$ is a probability vector representing the distribution over words for the ith topic. The sampling process for a document is as follows. First, the topic mixture $\boldsymbol{h}$ is drawn from the Dirichlet distribution. Then, the vth word in the document (for $v \in [\ell] := \{1, 2, \ldots, \ell\}$) is generated by: (i) drawing $t \in [k]$ according to the discrete distribution specified by h, then (ii) drawing $\boldsymbol{x}_v$ according to the discrete distribution specified by $\boldsymbol{o}_t$. Note that $\boldsymbol{x}_v$ is independent of $\boldsymbol{h}$ given t. For this model to fit in our setting, we use the "one-hot" encoding for $\boldsymbol{x}_v$ (also used by [6] in this context): $\boldsymbol{x}_v \in \{0, 1\}^d$ with

$$[\boldsymbol{x}_v]_j = 1 \quad \Leftrightarrow \quad \text{the } v\text{th word in the document is the } j\text{th vocabulary word.}$$

Observe that

$$\mathbb{E}[\boldsymbol{x}_v|\boldsymbol{h}] = \sum_{i=1}^{k} \Pr[t = i|\boldsymbol{h}] \cdot \mathbb{E}[\boldsymbol{x}_v|t = i, \boldsymbol{h}] = \sum_{i=1}^{k} h_i \cdot \boldsymbol{o}_i = \boldsymbol{O}\boldsymbol{h}$$

as required in Condition 1. Again, note that the noise model is not additive.

3 Moment Structure in Multi-view Models

In this section, we describe the structure of moments for the multi-view models, focusing primarily on the exchangeable models (the independent latent factors model and the Dirichlet model). Recall that for these exchangeable models, we assume $\boldsymbol{O}^{(v)} = \boldsymbol{O}$ for all $v \in [\ell]$ (Condition 3). In all cases, the pth-order (not necessarily raw) moments are of the form

$$\sum_{i=1}^{k} c_{i,p} \underbrace{\boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \cdots \otimes \boldsymbol{o}_i}_{p \text{ times}},$$

where $c_{1,p}, c_{2,p}, \ldots, c_{k,p} \in \mathbb{R}$ are scalars possibly depending on p. This is known as a symmetric canonical polyadic decomposition [24,25,34] (or a Tucker decomposition with a diagonal core tensor [44]). We show that if the $c_{i,p}$ are non-zero, then the $\boldsymbol{o}_i$ (up to some scaling and permutation) can be extracted from these moments using simple algebraic techniques such as that from [23]. We also give a reduction from the general multi-view setting to the exchangeable setting.

3.1 Moments of Skewed and Kurtotic Independent Factors

Let $m_{i,p} := \mathbb{E}[(h_i - \mathbb{E}[h_i])^p]$ denote the pth central moment of h_i. The *variance*, *skewness*, and *kurtosis* of h_i are given by $\sigma_i^2 := m_{i,2}$, $\gamma_i := m_{i,3}/\sigma_i^3$, and $\kappa_i := m_{i,4}/\sigma_i^4 - 3$, respectively.

Define the following moments of the observable random vectors:

$$\begin{aligned} \boldsymbol{\mu} &:= \mathbb{E}[\boldsymbol{x}_1] \in \mathbb{R}^d, \\ \textbf{Pairs} &:= \mathbb{E}[(\boldsymbol{x}_1 - \boldsymbol{\mu}) \otimes (\boldsymbol{x}_2 - \boldsymbol{\mu})] \in \mathbb{R}^{d \times d}, \\ \textbf{Triples} &:= \mathbb{E}[(\boldsymbol{x}_1 - \boldsymbol{\mu}) \otimes (\boldsymbol{x}_2 - \boldsymbol{\mu}) \otimes (\boldsymbol{x}_3 - \boldsymbol{\mu})] \in \mathbb{R}^{d \times d \times d}. \end{aligned} \tag{1}$$

Here, we use $\otimes$ to denote the tensor product. For instance, given vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$, the tensor product $\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \in \mathbb{R}^{d \times d \times d}$ is the third-order tensor whose (i, j, k)th entry is $u_i v_j w_k$.

Lemma 1 (Independent latent factors moments) *Assume Conditions 1 and 3 and that* $\boldsymbol{h}$ *has a product distribution. Then*

$$\boldsymbol{\mu} = \sum_{i=1}^{k} m_{i,1} \boldsymbol{o}_i, \quad \textbf{Pairs} = \sum_{i=1}^{k} m_{i,2} \boldsymbol{o}_i \otimes \boldsymbol{o}_i, \quad \textbf{Triples} = \sum_{i=1}^{k} m_{i,3} \boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_i.$$

Proof Since $\mathbb{E}[\boldsymbol{x}_v | \boldsymbol{h}] = \boldsymbol{O}\boldsymbol{h}$ (by Conditions 1 and 3) we have

$$\boldsymbol{\mu} = \boldsymbol{O}\mathbb{E}[\boldsymbol{h}] = \sum_{i=1}^{k} m_{i,1} \boldsymbol{o}_i.$$

This also implies

$$\mathbb{E}[(\boldsymbol{x}_v - \boldsymbol{\mu}) | \boldsymbol{h}] = \boldsymbol{O}(\boldsymbol{h} - \mathbb{E}[\boldsymbol{h}]) = \sum_{i=1}^{k} (h_i - \mathbb{E}[h_i]) \boldsymbol{o}_i.$$

Since $\boldsymbol{x}_1 - \boldsymbol{\mu}$ and $\boldsymbol{x}_2 - \boldsymbol{\mu}$ are conditionally independent given $\boldsymbol{h}$ (under Condition 1), we may write **Pairs** as the expectation of a tensor product:

$$\textbf{Pairs} = \mathbb{E}\left[\left(\sum_{i=1}^{k}(h_i - \mathbb{E}[h_i])\boldsymbol{o}_i\right) \otimes \left(\sum_{i=1}^{k}(h_i - \mathbb{E}[h_i])\boldsymbol{o}_i\right)\right].$$

Expanding out the tensor product and applying linearity leaves only the diagonal terms with $\mathbb{E}[(h_i - \mathbb{E}[h_i])^2]$. This is because $\boldsymbol{h}$ has a product distribution by assumption, and thus $\mathbb{E}[(h_i - \mathbb{E}[h_i])(h_j - \mathbb{E}[h_j])] = 0$ for $i \neq j$. With only the diagonal terms in the product remaining, we have

$$\textbf{Pairs} = \sum_{i=1}^{k}\mathbb{E}[(h_i - \mathbb{E}[h_i])^2]\boldsymbol{o}_i \otimes \boldsymbol{o}_i = \sum_{i=1}^{k} m_{i,2}\boldsymbol{o}_i \otimes \boldsymbol{o}_i.$$

An analogous argument gives the claim for **Triples**. □

Lemma 1 shows that $\boldsymbol{\mu}$, **Pairs**, and **Triples** possess the structure needed to apply the tensor decomposition technique of [5] to extract the $\boldsymbol{o}_i$. However, this is only possible if the scalar factors are non-zero; in Lemma 1, the scalar factors are the central moments $m_{i,p}$. If $\boldsymbol{h}$ has a distribution symmetric about its mean, the third central moment is zero, and hence **Triples** cannot be used. One recourse comes from the literature on independent component analysis; if the distribution of $\boldsymbol{h}$ is kurtotic (i.e., $\kappa_i \neq 0$), then one may use fourth-order moments in the form of the fourth cumulant tensor [14]. In the next lemma, we show that this can also be applied with multi-view/exchangeable models.

Define

$$\textbf{Quadruples} := \mathbb{E}\big[(\boldsymbol{x}_1 - \boldsymbol{\mu}) \otimes (\boldsymbol{x}_2 - \boldsymbol{\mu}) \otimes (\boldsymbol{x}_3 - \boldsymbol{\mu}) \otimes (\boldsymbol{x}_4 - \boldsymbol{\mu})\big] - \boldsymbol{T} \tag{2}$$

where $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d\times d}$ is the fourth-order tensor whose (i, j, m, n)th entry is

$$T_{i,j,m,n} = [\textbf{Pairs}]_{i,j}[\textbf{Pairs}]_{m,n} + [\textbf{Pairs}]_{i,m}[\textbf{Pairs}]_{j,n} + [\textbf{Pairs}]_{i,n}[\textbf{Pairs}]_{j,m}.$$

Lemma 2 (Independent latent factors moments, fourth-order) *Under the same setting as Lemma 1,*

$$\textbf{Quadruples} = \sum_{i=1}^{k}(m_{i,4} - 3m_{i,2}^2)\boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_i.$$

Proof By the same argument as in Lemma 1, we have that

$$\mathbb{E}\big[(\boldsymbol{x}_1 - \boldsymbol{\mu}) \otimes (\boldsymbol{x}_2 - \boldsymbol{\mu}) \otimes (\boldsymbol{x}_3 - \boldsymbol{\mu}) \otimes (\boldsymbol{x}_4 - \boldsymbol{\mu})\big] = \mathbb{E}[\boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v}]$$

where

$$\boldsymbol{v} := \sum_{i=1}^{k} (h_i - \mathbb{E}[h_i])\boldsymbol{o}_i .$$

Expanding the tensor product and applying linearity of expectation leaves only terms involving $\mathbb{E}[(h_i - \mathbb{E}[h_i])^4]$ and $\mathbb{E}[(h_i - \mathbb{E}[h_i])^2 (h_j - \mathbb{E}[h_j])^2]$, again due to the product distribution of $\boldsymbol{h}$. Thus

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v}] \\
&\quad = \sum_{i=1}^{k} m_{i,4} \boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_i \\
&\qquad + \sum_{i \neq j} m_{i,2} m_{j,2} \left(\boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_j \otimes \boldsymbol{o}_j + \boldsymbol{o}_i \otimes \boldsymbol{o}_j \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_j + \boldsymbol{o}_i \otimes \boldsymbol{o}_j \otimes \boldsymbol{o}_j \otimes \boldsymbol{o}_i \right) \\
&\quad = \sum_{i=1}^{k} (m_{i,4} - 3m_{i,2}^2) \boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_i \\
&\qquad + \sum_{i,j} m_{i,2} m_{j,2} \left(\boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_j \otimes \boldsymbol{o}_j + \boldsymbol{o}_i \otimes \boldsymbol{o}_j \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_j + \boldsymbol{o}_i \otimes \boldsymbol{o}_j \otimes \boldsymbol{o}_j \otimes \boldsymbol{o}_i \right) \\
&\quad = \sum_{i=1}^{k} (m_{i,4} - m_{i,2}^2) \boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_i \otimes \boldsymbol{o}_i + \boldsymbol{T}
\end{aligned}$$

where the step uses the identity for **Pairs** from Lemma 1. Rearranging the terms gives the desired identity. □

We note that if $\boldsymbol{h}$ is an isotropic Gaussian random vector, then both $m_{i,3}$ and $m_{i,4} - 3m_{i,2}^2$ are zero for all $i \in [k]$, causing both **Triples** and **Quadruples** to vanish. As expected, these higher-order moments cannot help with the identification of $\boldsymbol{O}$ without further assumptions: this is simply because $\boldsymbol{g} := \boldsymbol{U}^\top \boldsymbol{h}$ has the same distribution as $\boldsymbol{h}$ when $\boldsymbol{U}$ is orthogonal, and $\mathbb{E}[\boldsymbol{x}_v | \boldsymbol{h}] = \boldsymbol{O}\boldsymbol{h} = (\boldsymbol{O}\boldsymbol{U})\boldsymbol{g}$.

3.2 Moments of Dirichlet Factors

The Dirichlet distribution is not a product distribution, and therefore Lemma 1 does not apply to models where $\boldsymbol{h} \sim \mathrm{Dir}(\boldsymbol{\alpha})$. However, it is *nearly* a product distribution. Indeed, if $\boldsymbol{h} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, then $\boldsymbol{h}$ has the same distribution $\boldsymbol{x} / \sum_{i=1}^{k} x_i$, where $x_i \sim \mathrm{Gamma}(\alpha_i, 1)$ (i.e., x_i follows a Gamma distribution with shape parameter α_i and scale parameter 1), and $x_1, x_2, \ldots, x_k$ are independent. Thus, the essential reason why the Dirichlet distribution is not a product distribution is because it is restricted to the probability simplex.

Lemma 3 (below) nevertheless shows that if the concentration parameter α_0 is known, the correlations introduced by restricting to the probability simplex can be

accounted for by slightly tweaking the moments; this tweaking causes the same kind of structure as in Lemma 1 to manifest.

Define the following moments of the observable random vectors:

$$\begin{aligned}
\boldsymbol{\mu} &:= \mathbb{E}[\boldsymbol{x}_1] \in \mathbb{R}^d, \\
\mathbf{Pairs}_{\alpha_0} &:= \mathbb{E}[\boldsymbol{x}_1 \otimes \boldsymbol{x}_2] - \frac{\alpha_0}{\alpha_0+1}\boldsymbol{\mu}\otimes\boldsymbol{\mu} \in \mathbb{R}^{d\times d}, \\
\mathbf{Triples}_{\alpha_0} &:= \mathbb{E}[\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3] \\
&\quad - \frac{\alpha_0}{\alpha_0+2}\Big(\mathbb{E}[\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{\mu}] + \mathbb{E}[\boldsymbol{x}_1 \otimes \boldsymbol{\mu} \otimes \boldsymbol{x}_3] + \mathbb{E}[\boldsymbol{\mu} \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3]\Big) \\
&\quad + \frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}\boldsymbol{\mu}\otimes\boldsymbol{\mu}\otimes\boldsymbol{\mu} \in \mathbb{R}^{d\times d\times d}. \qquad (3)
\end{aligned}$$

Lemma 3 (Dirichlet factors moments) *Assume Conditions 1 and 3 and that* $\boldsymbol{h} \sim \mathrm{Dir}(\boldsymbol{\alpha})$. *Then*

$$\boldsymbol{\mu} = \sum_{i=1}^k \frac{\alpha_i}{\alpha_0}\boldsymbol{o}_i, \quad \mathbf{Pairs}_{\alpha_0} = \sum_{i=1}^k \frac{\alpha_i}{(\alpha_0+1)\alpha_0}\boldsymbol{o}_i\otimes\boldsymbol{o}_i,$$
$$\mathbf{Triples}_{\alpha_0} = \sum_{i=1}^k \frac{2\alpha_i}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\boldsymbol{o}_i\otimes\boldsymbol{o}_i\otimes\boldsymbol{o}_i.$$

Proof We directly calculate univariate, bivariate, and trivariate moments of $\boldsymbol{h}$: for any distinct $i, j, l \in [k]$,

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

Putting these in vector, matrix, and third-order tensor form,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{h}] &= \frac{1}{\alpha_0}\sum_{i=1}^k \alpha_i \boldsymbol{e}_i \\
\mathbb{E}[\boldsymbol{h}\otimes\boldsymbol{h}] &= \frac{1}{(\alpha_0+1)\alpha_0}\left(\sum_{i=1}^k \alpha_i \boldsymbol{e}_i\otimes\boldsymbol{e}_i + \boldsymbol{\alpha}\otimes\boldsymbol{\alpha}\right) \\
\mathbb{E}[\boldsymbol{h}\otimes\boldsymbol{h}\otimes\boldsymbol{h}] &= \frac{1}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\left(2\sum_{i=1}^k \alpha_i\left(\boldsymbol{e}_i\otimes\boldsymbol{e}_i\otimes\boldsymbol{e}_i\right)\right.
\end{aligned}$$

$$+\sum_{i=1}^{k}\alpha_i\left(\boldsymbol{e}_i\otimes\boldsymbol{e}_i\otimes\boldsymbol{\alpha}\right)+\sum_{i=1}^{k}\alpha_i\left(\boldsymbol{\alpha}\otimes\boldsymbol{e}_i\otimes\boldsymbol{e}_i\right)$$
$$+\sum_{i=1}^{k}\alpha_i\left(\boldsymbol{e}_i\otimes\boldsymbol{\alpha}\otimes\boldsymbol{e}_i\right)+\boldsymbol{\alpha}\otimes\boldsymbol{\alpha}\otimes\boldsymbol{\alpha}\Big).$$

Since $\mathbb{E}[\boldsymbol{x}_v|\boldsymbol{h}] = \boldsymbol{O}\boldsymbol{h}$ by Conditions 1 and 3 (as in the proof of Lemma 1), and because the $\boldsymbol{x}_v$ are conditionally independent given $\boldsymbol{h}$, the claim follows by linearity and rearranging. □

3.3 Identifiability from Low-order Moments

We now provide a simple argument that establishes the identifiability of the columns of the topic matrix $\boldsymbol{O}$ (up to scaling and permutation) from the low-order moments considered in Lemmas 1, 2, and 3. Here, Condition 2 plays an essential role, as does the non-Gaussianity of $\boldsymbol{h}$ in the case of the independent latent factors model.

Theorem 2 (Identifiability from low-order moments) *Assume Conditions 1, 2, and 3 hold.*

1. *If $\boldsymbol{h}$ has a product distribution and each h_i has positive variance σ_i^2 and non-zero skew γ_i, then there is a randomized algorithm that, given* **Pairs** *and* **Triples** *from (1), returns the set of vectors returns $\{(s_i\sigma_i\boldsymbol{o}_i, s_i\gamma_i) : i \in [k]\}$ for some $\{s_1, s_2, \ldots, s_k\} \subseteq \{\pm 1\}$.*
2. *If $\boldsymbol{h}$ has a product distribution and each h_i has positive variance σ_i^2 and non-zero kurtosis κ_i, then there is a randomized algorithm that, given* **Pairs** *and* **Quadruples** *from (1) and (2), returns $\{(s_i\sigma_i\boldsymbol{o}_i, s_i\kappa_i) : i \in [k]\}$ for some $\{s_1, s_2, \ldots, s_k\} \subseteq \{\pm 1\}$.*
3. *If $\boldsymbol{h} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ for some $\boldsymbol{\alpha} \in \mathbb{R}^k_{>0}$, then there is a randomized algorithm that, given* $\mathbf{Pairs}_{\alpha_0}$ *and* $\mathbf{Triples}_{\alpha_0}$ *from (3), returns $\{(s_i\sqrt{c_{i,2}}\boldsymbol{o}_i, s_i c_{i,3}/c_{i,2}^{3/2}) : i \in [k]\}$ for some $\{s_1, s_2, \ldots, s_k\} \subseteq \{\pm 1\}$, where $c_{i,2} = \alpha_i/((\alpha_0+1)\alpha_0)$ and $c_{i,3} = 2\alpha_i/((\alpha_0+2)(\alpha_0+1)\alpha_0)$.*

Proof The theorem follows from the structural results from Lemma 1, Lemma 2, and Lemma 3, combined with Lemma 4 (below). □

Lemma 4 is a variant of a result from [23], which establishes the uniqueness of (symmetric) canonical polyadic decompositions under a rank condition. Our variant uses a symmetrized version of procedure from [23], and is proved using the following observations. First, if the second-order moment matrix (called $\boldsymbol{P}$ in Lemma 4) has rank k, then it defines an inner product system under which the $\boldsymbol{o}_i$ are orthogonal. Moreover, under this inner product, the $\boldsymbol{o}_i$ are orthogonal eigenvectors of a generic flattening of higher-order moment tensors ($\boldsymbol{T}$ or $\boldsymbol{Q}$ in Lemma 4).

Below, for a vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_d) \in \mathbb{R}^d$, let $\boldsymbol{T}\{\boldsymbol{v}\} \in \mathbb{R}^{d\times d}$ denote the matrix whose $(i, j)^{\text{th}}$ entry is $\sum_{l=1}^{d} v_l[\boldsymbol{T}]_{i,j,l}$.

Lemma 4 *Let $\boldsymbol{O} := [\boldsymbol{o}_1|\boldsymbol{o}_2|\cdots|\boldsymbol{o}_k] \in \mathbb{R}^{d\times k}$, $\boldsymbol{P} := \sum_{i=1}^{k} c_{i,2}\boldsymbol{o}_i\otimes\boldsymbol{o}_i \in \mathbb{R}^{d\times d}$, $\boldsymbol{T} := \sum_{i=1}^{k} c_{i,3}\boldsymbol{o}_i\otimes\boldsymbol{o}_i\otimes\boldsymbol{o}_i \in \mathbb{R}^{d\times d\times d}$, and $\boldsymbol{Q} := \sum_{i=1}^{k} c_{i,4}\boldsymbol{o}_i\otimes\boldsymbol{o}_i\otimes\boldsymbol{o}_i\otimes\boldsymbol{o}_i \in \mathbb{R}^{d\times d\times d\times d}$.*

Algorithm 1 Identification from exact moments

input positive integer $k \in \mathbb{N}$, $\boldsymbol{P} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{T} \in \mathbb{R}^{d\times d\times d}$ satisfying conditions in Lemma 4.
output $\{(\boldsymbol{v}_i, \lambda_i) : i \in [k]\}$.

1: Let $\sum_{i=1}^{k} \eta_i \boldsymbol{\zeta}_i \otimes \boldsymbol{\zeta}_i$ be an eigendecomposition of $\boldsymbol{P}$, where $\{\boldsymbol{\zeta}_i : i \in [k]\}$ are orthonormal eigenvectors, and $\{\eta_i : i \in [k]\}$ are corresponding positive eigenvalues.
2: Set $\boldsymbol{W} := [\boldsymbol{\zeta}_1|\boldsymbol{\zeta}_2|\cdots|\boldsymbol{\zeta}_k]\mathbf{diag}(1/\sqrt{\eta_1}, 1/\sqrt{\eta_2}, \ldots, 1/\sqrt{\eta_k})$.
3: Draw $\boldsymbol{\theta}$ uniformly at random from the unit sphere in $\mathbb{R}^k$.
4: Let $\sum_{i=1}^{k} \delta_i \boldsymbol{\xi}_i \otimes \boldsymbol{\xi}_i$ be an eigendecomposition of $\boldsymbol{W}^\top \boldsymbol{T}\{\boldsymbol{W}\boldsymbol{\theta}\}\boldsymbol{W}$, where $\{\boldsymbol{\theta}_i : i \in [k]\}$ are orthonormal eigenvectors, and $\{\delta_i : i \in [k]\}$ are corresponding non-zero eigenvalues.
5: Let $\boldsymbol{v}_i := (\boldsymbol{W}^\top)^\dagger \boldsymbol{\xi}_i$ and $\lambda_i := (\boldsymbol{W}\boldsymbol{\xi}_i)^\top \boldsymbol{T}\{\boldsymbol{W}\boldsymbol{\xi}_i\}\boldsymbol{W}\boldsymbol{\xi}_i$ for each $i \in [k]$.

Assume that $\boldsymbol{O}$ has full column rank and that $c_{i,2} > 0$ for all $i \in [k]$. If $c_{i,3} \neq 0$ for all $i \in [k]$, then there is a randomized algorithm that, given $\boldsymbol{P}$ and $\boldsymbol{T}$ as input, returns $\{(s_i c_{i,2}^{1/2} \boldsymbol{o}_i, s_i c_{i,3}/c_{i,2}^{3/2}) : i \in [k]\}$ for some $\{s_1, s_2, \ldots, s_k\} \subseteq \{\pm 1\}$. If $c_{i,4} \neq 0$ for all $i \in [k]$, then the same holds with $\boldsymbol{Q}$ in place of $\boldsymbol{T}$ and $c_{i,4}/c_{i,2}^2$ in place of $c_{i,3}/c_{i,2}^{3/2}$.

Proof Under the assumptions on $\boldsymbol{O}$ and the $c_{i,2}$, the matrix $\boldsymbol{P}$ is positive definite. By rescaling the columns of $\boldsymbol{O}$ (replacing $\boldsymbol{o}_i$ by $\sqrt{c_{i,2}}\boldsymbol{o}_i$), we may assume that $c_{i,2} = 1$ for all $i \in [k]$. Therefore there exists a matrix $\boldsymbol{W} \in \mathbb{R}^{d\times k}$, which can be obtained from the singular value decomposition of $\boldsymbol{P}$, such that $\boldsymbol{W}^\top \boldsymbol{P}\boldsymbol{W} = (\boldsymbol{W}^\top \boldsymbol{O})(\boldsymbol{W}^\top \boldsymbol{O})^\top = \boldsymbol{I}_k$. This in turn implies that $\boldsymbol{M} := \boldsymbol{W}^\top \boldsymbol{O}$ is orthogonal. Then, for any vector $\boldsymbol{\theta} \in \mathbb{R}^k$,

$$\begin{aligned}\boldsymbol{W}^\top \boldsymbol{T}(\boldsymbol{W}\boldsymbol{\theta})\boldsymbol{W} &= \boldsymbol{W}^\top \boldsymbol{O}\mathbf{diag}(\mathbf{diag}(c_{1,3}, c_{2,3}, \ldots, c_{k,3})\boldsymbol{O}^\top \boldsymbol{W}\boldsymbol{\theta})\boldsymbol{O}^\top \boldsymbol{W} \\ &= \boldsymbol{M}\mathbf{diag}(\mathbf{diag}(c_{1,3}, c_{2,3}, \ldots, c_{k,3})\boldsymbol{M}^\top \boldsymbol{\theta})\boldsymbol{M}^\top.\end{aligned}$$

Since $\boldsymbol{M}$ is orthogonal, the above displayed equation gives an eigendecomposition of $\boldsymbol{W}^\top \boldsymbol{T}\{\boldsymbol{W}\boldsymbol{\theta}\}\boldsymbol{W}$, and the set of eigenvectors corresponding to non-repeated eigenvalues are uniquely defined up to sign. Each such unit-norm eigenvector $\boldsymbol{\xi}_i$ (after appropriate reordering) is of the form $s_i \boldsymbol{M}\boldsymbol{e}_i = s_i \boldsymbol{W}^\top \boldsymbol{o}_i$ for some $s_i \in \{\pm 1\}$. Therefore $\boldsymbol{v}_i := (\boldsymbol{W}^\dagger)^\top \boldsymbol{\xi}_i = s_i \boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{o}_i = s_i \boldsymbol{o}_i$, since $\mathrm{range}(\boldsymbol{W}) = \mathrm{range}(\boldsymbol{O})$. Finally,

$$\begin{aligned}\lambda_i := (\boldsymbol{W}\boldsymbol{\xi}_i)^\top \boldsymbol{T}\{\boldsymbol{W}\boldsymbol{\xi}_i\}\boldsymbol{W}\boldsymbol{\xi}_i &= \sum_{j=1}^{k} c_{j,3}(s_i o_j^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{o}_i)^3 \\ &= s_i \sum_{j=1}^{k} c_{j,3}(\boldsymbol{e}_j^\top \boldsymbol{O}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{O}\boldsymbol{e}_i)^3 = s_i c_{i,3}.\end{aligned}$$

Assume that $c_{i,3} \neq 0$ for all $i \in [k]$. If $\boldsymbol{\theta}$ is drawn uniformly at random from $\mathbb{S}^{k-1}$, then so is $\boldsymbol{M}^\top \boldsymbol{\theta}$. In this case, almost surely, the entries of $\mathbf{diag}(\mathbf{diag}(c_{1,3}, c_{2,3}, \ldots, c_{k,3})\boldsymbol{M}^\top \boldsymbol{\theta})$ are unique. Hence, no eigenvalue of $\boldsymbol{W}^\top \boldsymbol{T}\{\boldsymbol{W}\boldsymbol{\theta}\}\boldsymbol{W}$ is repeated, so the set $\{(\boldsymbol{v}_i, \lambda_i) : i \in [k]\}$ has the property claimed by the lemma and is returned by Algorithm 1.

An analogous argument proves the claim assuming $c_{i,4} \neq 0$ for all $i \in [k]$ using $\boldsymbol{Q}$ in place of $\boldsymbol{T}$. □

The critical aspect in Theorem 2 is that only low-order observable moments are needed to identify the columns of $\boldsymbol{O}$. The observability implies that they can be estimated from data, and used in a plug-in estimator. The low-order nature of the moments mean that they can be estimated reliably (relative to higher-order moments).

From Theorem 2, we may also conclude that since $\boldsymbol{\mu}$, $\mathbf{Pairs}_{\alpha_0}$, and $\mathbf{Triples}_{\alpha_0}$ involve only $\boldsymbol{x}_1, \boldsymbol{x}_2$, and $\boldsymbol{x}_3$, a random document under the LDA model need only have three words (conditionally independent given $\boldsymbol{h}$). This is the same conclusion as in the mixture of multinomial model for documents studied in [6], even though in that model, the words in documents were assumed to only be generated from a single topic.

It is worth noting that $\mathbf{Pairs}_{\alpha_0}$ and $\mathbf{Triples}_{\alpha_0}$ depend on the value $\alpha_0 = \sum_{i=1}^{k} \alpha_i$, so it must be known in order to form the appropriate moment estimates. In some applications, one may have prior knowledge of α_0, as it characterizes the concentration of the Dirichlet distribution (and, indeed, having prior knowledge of α_0 is indeed much weaker than knowing the entire parameter vector $\boldsymbol{\alpha}$). The particular case where $\boldsymbol{\alpha} = (1, 1, \ldots, 1)$ is essentially the problem of learning the vertices of a simplex when given access to (noisy) samples drawn from the uniform distribution over this simplex; therefore our result resolves this open problem posed by [21].

In the case of the independent latent factors model, the third- and fourth-order moments are exploited to take advantage of the non-Gaussianity of $\boldsymbol{h}$ (which is not possible with only first- and second-order moments, without further assumptions).

3.4 Reducing the General Multi-view Setting to the Exchangeable Setting

To conclude this section, we show how to reduce the general multi-view setting (where the different $\boldsymbol{O}^{(v)}$ are not necessarily the same) to the case where the $\boldsymbol{O}^{(v)} = \boldsymbol{O}$ for all $v \in [\ell]$. Here, we must assume that the number of views is at least three (i.e., $\ell \geq 3$, just as in the independent latent factors and Dirichlet models, since **Triples**, **Quadruples**, and $\mathbf{Triples}_{\alpha_0}$ all depend on at least $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$). This reduction is based on a proof technique used in [2].

Define $\mathbf{Pairs}_{u,v} := \mathbb{E}[\boldsymbol{x}_u \otimes \boldsymbol{x}_v]$ for $u, v \in [\ell]$. Fix any $v_0 \in [\ell]$, which we call the *target view*. Then, for each $v \in [\ell] \backslash \{v_0\}$, pick any $u = u(v) \in [\ell] \backslash \{v, v_0\}$ (which is possible since $\ell \geq 3$) and define

$$\boldsymbol{C}_{v \to v_0} := \mathbf{Pairs}_{v_0,u} \mathbf{Pairs}_{v,u}^{\dagger}, \qquad \boldsymbol{C}_{v_0 \to v} := \mathbf{Pairs}_{v,u} \mathbf{Pairs}_{v_0,u}^{\dagger}.$$

Proposition 1 (View symmetrization) *Assume Conditions 1 and 2 hold, and also that* $\mathbb{E}[\boldsymbol{h} \otimes \boldsymbol{h}]$ *is invertible. Fix a target view* $v_0 \in [\ell]$*, and let* $\tilde{\boldsymbol{O}} := \boldsymbol{O}^{(v_0)}$*. Let* $\boldsymbol{C}_{v_0 \to v_0} := \boldsymbol{I}_d$*, and for each view* $v \in [\ell]$*, let* $\tilde{\boldsymbol{x}}_v := \boldsymbol{C}_{v \to v_0} \boldsymbol{x}_v$ *where* $\boldsymbol{C}_{v \to v_0}$ *is defined above for* $v \neq v_0$*. For each* $v \in [\ell]$*,*

$$\mathbb{E}[\tilde{\boldsymbol{x}}_v | \boldsymbol{h}] = \tilde{\boldsymbol{O}} \boldsymbol{h}, \qquad \boldsymbol{O}^{(v)} = \boldsymbol{C}_{v_0 \to v} \tilde{\boldsymbol{O}}.$$

Proof Assume without loss of generality that $v_0 := 1$, $v := 2$, and $u = u(v) := 3$ (in the definition of $\boldsymbol{C}_{2 \to 1}$). By Condition 1, $\mathbf{Pairs}_{1,3} = \boldsymbol{O}^{(1)} \mathbb{E}[\boldsymbol{h} \otimes \boldsymbol{h}] \boldsymbol{O}^{(3)\top}$ and $\mathbf{Pairs}_{2,3} = \boldsymbol{O}^{(2)} \mathbb{E}[\boldsymbol{h} \otimes \boldsymbol{h}] \boldsymbol{O}^{(3)\top}$. By Condition 2 and the assumption that

$\mathbb{E}[\boldsymbol{h} \otimes \boldsymbol{h}]$ is invertible, both $\mathbf{Pairs}_{1,3}$ and $\mathbf{Pairs}_{2,3}$ have rank k. Moreover, $\mathbf{Pairs}_{2,3}^{\dagger} = (\boldsymbol{O}^{(3)\top})^{\dagger}\mathbb{E}[\boldsymbol{h} \otimes \boldsymbol{h}]^{-1}\boldsymbol{O}^{(2)\dagger}$, so

$$\boldsymbol{C}_{2\to 1} = \left(\boldsymbol{O}^{(1)}\mathbb{E}[\boldsymbol{h} \otimes \boldsymbol{h}]\boldsymbol{O}^{(3)\top}\right)\left((\boldsymbol{O}^{(3)\top})^{\dagger}\mathbb{E}[\boldsymbol{h} \otimes \boldsymbol{h}]^{-1}\boldsymbol{O}^{(2)\dagger}\right) = \boldsymbol{O}^{(1)}\boldsymbol{O}^{(2)\dagger}.$$

Since $\boldsymbol{O}^{(2)}$ has full column rank (by Condition 2),

$$\mathbb{E}[\tilde{\boldsymbol{x}}_2|\boldsymbol{h}] = \boldsymbol{C}_{2\to 1}\mathbb{E}[\boldsymbol{x}_2|\boldsymbol{h}] = \boldsymbol{O}^{(1)}\boldsymbol{O}^{(2)\dagger}\boldsymbol{O}^{(2)}\boldsymbol{h} = \boldsymbol{O}^{(1)}\boldsymbol{h} = \tilde{\boldsymbol{O}}\boldsymbol{h}.$$

By the same argument, $\boldsymbol{C}_{1\to 2} = \boldsymbol{O}^{(2)}\boldsymbol{O}^{(1)\dagger} = \boldsymbol{O}^{(2)}\tilde{\boldsymbol{O}}^{\dagger}$; therefore $\boldsymbol{C}_{1\to 2}\tilde{\boldsymbol{O}} = \boldsymbol{O}^{(2)}\tilde{\boldsymbol{O}}^{\dagger}\tilde{\boldsymbol{O}} = \boldsymbol{O}^{(2)}$ since $\tilde{\boldsymbol{O}}$ has full column rank. □

4 Estimation from Moments via Orthogonal Tensor Decomposition

In addition to establishing identifiability, Theorem 2 provides an efficient randomized algorithm for recovering the columns of $\boldsymbol{O}$ up to permutation and scaling, and it indeed can be shown to work using only estimates of the required moments using techniques similar to [6,39]. However, the resulting sample complexity is rather high (i.e., a high-degree polynomial) on account of the randomization technique used to ensure the uniqueness of an eigendecomposition. Indeed, the randomization collapses moment tensors into matrices, but the resulting spacing between eigenvalues is small; hence, such procedures may not be very robust to errors in the moment estimates. It turns out this is only an artifact of the algorithmic technique, as a different technique based on orthogonal tensor decomposition has both a better analysis and better empirical support. In this section, we recall the tensor decomposition technique from [5] and show how it can be applied to the estimation problem for the multi-view models considered in this work.

4.1 Multi-way Arrays, Multi-linear Forms, and Tensor Algebras

Recall that low-order moments from Sect. 3 have the form $\sum_{i=1}^{k} c_{i,p}\boldsymbol{o}_i^{\otimes p}$, where $\boldsymbol{v}^{\otimes p} := \boldsymbol{v} \otimes \boldsymbol{v} \otimes \cdots \otimes \boldsymbol{v}$ denotes the p-fold tensor product of a vector $\boldsymbol{v}$ with itself. We assume that

$$\boldsymbol{P} := \sum_{i=1}^{k} c_{i,2}\boldsymbol{o}_i^{\otimes 2} \quad \text{and} \quad \boldsymbol{T} := \sum_{i=1}^{k} c_{i,p}\boldsymbol{o}_i^{\otimes p}$$

for some $p \geq 3$ are given, with $c_{i,2} > 0$ and $c_{i,p} > 0$ for all $i \in [k]$, and $\{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_k\} \subset \mathbb{R}^d$ linearly independent. We note that it is also possible to handle the case where $c_{i,2} < 0$ and $c_{i,p} < 0$ for some $i \in [k]$ (using techniques from [46]), though we avoid here this case for simplicity. In our examples, we will specialize to the $p = 3$ case.

From the second-order moment matrix $\boldsymbol{P}$, as discussed in the proof of Lemma 4, we may extract a so-called *whitening matrix* $\boldsymbol{W} \in \mathbb{R}^{d\times k}$ using the singular value decomposition of $\boldsymbol{P}$ such that $\boldsymbol{W}^\top \boldsymbol{P}\boldsymbol{W} = \boldsymbol{I}_k$. Put another way, $\boldsymbol{P}$ defines a k-dimensional inner product subspace of $\mathbb{R}^d$ in which $\mathcal{B} := \{\sqrt{c_{1,2}}\boldsymbol{o}_1, \sqrt{c_{2,2}}\boldsymbol{o}_2, \ldots, \sqrt{c_{k,2}}\boldsymbol{o}_k\}$ are orthonormal. To see this, we define the inner product by

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle := \boldsymbol{u}^\top \boldsymbol{P}^\dagger \boldsymbol{v}$$

Since the $\boldsymbol{o}_i$ are linearly independent and $c_{i,2} > 0$,

$$\boldsymbol{P}^\dagger = (\boldsymbol{O}^\top)^\dagger \mathbf{diag}(c_{1,2}, c_{2,2}, \ldots, c_{k,2})^{-1}\boldsymbol{O}^\dagger,$$

where $\boldsymbol{O}^\dagger\boldsymbol{O} = \boldsymbol{I}_k$. This implies that

$$\langle \sqrt{c_{i,2}}\boldsymbol{o}_i, \sqrt{c_{j,2}}\boldsymbol{o}_j\rangle = \begin{cases} 0 & \text{if } \ i \neq j \\ 1 & \text{if } \ i = j \end{cases}$$

as claimed.

The pth-order tensor $\boldsymbol{T}$ can be viewed in a number of different ways. The first is as a p-way array of real numbers $T_{i_1,i_2,\ldots,i_p}$ for $i_1, i_2, \ldots, i_p \in [d]$. The second is as a p-linear form $\boldsymbol{T} : \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$ (generalizing the bilinear form view of a matrix): for any $\boldsymbol{u} = (u_1, u_2, \ldots, u_d) \in \mathbb{R}^d$, $\boldsymbol{v} = (v_1, v_2, \ldots, v_d) \in \mathbb{R}^d$, and $\boldsymbol{w} = (w_1, w_2, \ldots, w_d) \in \mathbb{R}^d$,

$$\boldsymbol{T}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \sum_{i,j,l} T_{i,j,l} u_i v_j w_l,$$

which connects the p-way array with the p-linear form. We will generally denote, for $\boldsymbol{U} = [\boldsymbol{u}_1|\boldsymbol{u}_2|\cdots|\boldsymbol{u}_d] \in \mathbb{R}^{m_1\times d}$, $\boldsymbol{V} = [\boldsymbol{v}_1|\boldsymbol{v}_2|\cdots|\boldsymbol{v}_d] \in \mathbb{R}^{m_2\times d}$, and $\boldsymbol{W} = [\boldsymbol{w}_1|\boldsymbol{w}_2|\cdots|\boldsymbol{w}_d] \in \mathbb{R}^{m_3\times d}$, the tensor $\boldsymbol{T}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \in \mathbb{R}^{m_1\times m_2\times m_3}$ given by

$$\boldsymbol{T}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) = \sum_{i,j,l} T_{i,j,l}\boldsymbol{u}_i \otimes \boldsymbol{v}_j \otimes \boldsymbol{w}_l.$$

(Note that the notation "$\boldsymbol{T}\{\boldsymbol{v}\}$" from Lemma 4 can be written as $\boldsymbol{T}(\boldsymbol{I}_d, \boldsymbol{I}_d, \boldsymbol{v})$.)

The third view of $\boldsymbol{T}$ is as a member of $(\mathbb{R}^d)^{\otimes p}$, the pth-order tensor product of $\mathbb{R}^d$. In this view, we may pick a basis for this vector space, and represent $\boldsymbol{T}$ in that basis. A natural choice is derived from the standard coordinate basis for $\mathbb{R}^d$, giving $\{\boldsymbol{e}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_l : i, j, l \in [d]\}$. In this basis, we identify the p-way array as a tensor algebra element:

$$\boldsymbol{T} = \sum_{i,j,l} T_{i,j,l}\boldsymbol{e}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_l.$$

Alternatively, if $\mathcal{B}' := \{\sqrt{c_{1,2}}\boldsymbol{o}_1, \sqrt{c_{2,2}}\boldsymbol{o}_2, \ldots \sqrt{c_{d,2}}\boldsymbol{o}_d\}$ is a completion of $\mathcal{B}$ from above, we may derive a different basis $\{\boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} : \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathcal{B}'\}$ for $(\mathbb{R}^d)^{\otimes p}$, under

which $\boldsymbol{T}$ has the following *diagonal representation*:

$$\boldsymbol{T} = \sum_{i=1}^{k} \frac{c_{i,3}}{c_{i,2}^{3/2}} (\sqrt{c_{i,2}}\boldsymbol{o}_i) \otimes (\sqrt{c_{i,2}}\boldsymbol{o}_i) \otimes (\sqrt{c_{i,2}}\boldsymbol{o}_i).$$

Note that $\boldsymbol{T}$ as a p-linear form can be expressed in terms of the standard inner product $\boldsymbol{u}^\top \boldsymbol{v} = \sum_{i=1}^{d} u_i v_i$ giving

$$\boldsymbol{T}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \sum_{i,j,l} T_{i,j,l} (\boldsymbol{e}_i^\top \boldsymbol{u})(\boldsymbol{e}_j^\top \boldsymbol{v})(\boldsymbol{e}_l^\top \boldsymbol{w}).$$

The proposed orthogonal tensor decomposition algorithm from [5] exploits the p-linear form based on the inner product $\langle \cdot, \cdot \rangle$, as it exploits the orthogonality of the $\sqrt{c_{i,2}}\boldsymbol{o}_i$:

$$\begin{aligned} \boldsymbol{T}(\boldsymbol{P}^\dagger \boldsymbol{u}, \boldsymbol{P}^\dagger \boldsymbol{v}, \boldsymbol{P}^\dagger \boldsymbol{w}) &= \sum_{i,j,l} T_{i,j,l} \langle \boldsymbol{e}_i, \boldsymbol{u} \rangle \langle \boldsymbol{e}_j, \boldsymbol{v} \rangle \langle \boldsymbol{e}_l, \boldsymbol{w} \rangle \\ &= \sum_{i=1}^{k} \frac{c_{i,3}}{c_{i,2}^{3/2}} \langle \sqrt{c_{i,2}}\boldsymbol{o}_i, \boldsymbol{u} \rangle \langle \sqrt{c_{i,2}}\boldsymbol{o}_i, \boldsymbol{v} \rangle \langle \sqrt{c_{i,2}}\boldsymbol{o}_i, \boldsymbol{w} \rangle \end{aligned}$$

Indeed, it is shown that the map

$$\boldsymbol{v} \mapsto \frac{\boldsymbol{T}(\boldsymbol{I}_d, \boldsymbol{P}^\dagger \boldsymbol{v}, \boldsymbol{P}^\dagger \boldsymbol{v})}{\sqrt{\langle \boldsymbol{T}(\boldsymbol{I}_d, \boldsymbol{P}^\dagger \boldsymbol{v}, \boldsymbol{P}^\dagger \boldsymbol{v}), \boldsymbol{T}(\boldsymbol{I}_d, \boldsymbol{P}^\dagger \boldsymbol{v}, \boldsymbol{P}^\dagger \boldsymbol{v}) \rangle}} \tag{4}$$

has non-zero stable fixed points only at $\sqrt{c_{i,2}}\boldsymbol{o}_i$ for $i \in [k]$. Moreover, repeated application of the above map starting from a random $\boldsymbol{v} \in \text{range}(\boldsymbol{P}) \cap \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 = 1\}$ converges at a quadratic rate to one of the $\sqrt{c_{i,2}}\boldsymbol{o}_i$ (Lemma 5.1 from [5]). Finally, it is easy to check that $\boldsymbol{T}(\boldsymbol{P}^\dagger \sqrt{c_{i,2}}\boldsymbol{o}_i, \boldsymbol{P}^\dagger \sqrt{c_{i,2}}\boldsymbol{o}_i, \boldsymbol{P}^\dagger \sqrt{c_{i,2}}\boldsymbol{o}_i) = c_{i,3}/c_{i,2}^{3/2}$.

4.2 Efficient Estimation Algorithm

In an actual implementation, we must estimate the moment matrices and tensors $\boldsymbol{P}$ and $\boldsymbol{T}$ from data. Let $S := \{(\boldsymbol{x}_1^{(j)}, \boldsymbol{x}_2^{(j)}, \ldots, \boldsymbol{x}_\ell^{(j)}) : j \in [n]\}$ be an i.i.d. sample comprised of n independent copies of $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell)$. We may form estimates, $\widehat{\boldsymbol{P}}$ and $\widehat{\boldsymbol{T}}$, of $\boldsymbol{P}$ and $\boldsymbol{T}$, respectively, using empirical averages with respect to S. For instance, in LDA, we may we let

$$\widehat{\boldsymbol{\mu}} := \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_1^{(j)},$$

$$\widehat{\mathbf{Pairs}}_{\alpha_0} := \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_1^{(j)} \otimes \boldsymbol{x}_2^{(j)} - \frac{\alpha_0}{\alpha_0 + 1} \widehat{\boldsymbol{\mu}} \otimes \widehat{\boldsymbol{\mu}},$$

$$\widehat{\mathbf{Triples}}_{\alpha_0} := \frac{1}{n}\sum_{j=1}^{n}\Bigg(\boldsymbol{x}_1^{(j)} \otimes \boldsymbol{x}_2^{(j)} \otimes \boldsymbol{x}_3^{(j)} - \frac{\alpha_0}{\alpha_0+2}\Big(\boldsymbol{x}_1^{(j)} \otimes \boldsymbol{x}_2^{(j)} \otimes \widehat{\boldsymbol{\mu}} + \boldsymbol{x}_1^{(j)} \otimes \widehat{\boldsymbol{\mu}} \otimes \boldsymbol{x}_3^{(j)} + \widehat{\boldsymbol{\mu}} \otimes \boldsymbol{x}_2^{(j)} \otimes \boldsymbol{x}_3^{(j)}\Big)\Bigg) + \frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}\widehat{\boldsymbol{\mu}} \otimes \widehat{\boldsymbol{\mu}} \otimes \widehat{\boldsymbol{\mu}},$$

and set $\widehat{\boldsymbol{P}} := \widehat{\mathbf{Pairs}}_{\alpha_0}$ and $\widehat{\boldsymbol{T}} := \widehat{\mathbf{Triples}}_{\alpha_0}$. Note that we may also ensure that $\widehat{\boldsymbol{P}}$ and $\widehat{\boldsymbol{T}}$ are symmetric (so $[\boldsymbol{T}]_{i,j,l} = [\boldsymbol{T}]_{i,l,j} = [\boldsymbol{T}]_{j,i,l}$ *etc.*). This gives consistent estimates of $\boldsymbol{P}$ and $\boldsymbol{T}$.

Using $\widehat{\boldsymbol{P}}$ and $\widehat{\boldsymbol{T}}$, we propose a plug-in approach to estimation for $\boldsymbol{O}$, given as Algorithm 2 (Lines 5–13 make up the robust tensor power method of [5]). This is essentially a robust version of the iteration given by (4) which finds an approximate orthogonal tensor decomposition of $\boldsymbol{T}$ given the nearby estimate $\widehat{\boldsymbol{T}}$. The algorithm from [5] is applied to the tensor $\widehat{\boldsymbol{T}}(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}})$, and the outputs $\hat{\boldsymbol{v}}_i$ and $\hat{\lambda}_i$ should be interpreted as estimates of the $\sqrt{c_{j,2}}\boldsymbol{o}_j$ and $c_{j,3}/c_{j,2}^{3/2}$ for some $j \in [k]$.

4.3 Analysis of Algorithm 2

In this section, we give a simple error analysis for Algorithm 2. Let $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_k > 0$ be the non-zero singular values of $\boldsymbol{P}$. Also, let $\lambda_i := c_{i,3}/c_{i,2}^{3/2}$, ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$. We define the operator norm of a symmetric third-order tensor $\boldsymbol{E} \in \mathbb{R}^{m\times m\times m}$ by

$$\|\boldsymbol{E}\|_2 := \max_{\|\boldsymbol{v}\|_2=1} |\boldsymbol{E}(\boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v})|.$$

Define $\epsilon_{\boldsymbol{P}} := \|\widehat{\boldsymbol{P}} - \boldsymbol{P}\|_2/\sigma_k$ and $\epsilon_{\boldsymbol{T}} := \|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_2$. The error bound is given in terms of $\epsilon_{\boldsymbol{P}}$ and $\epsilon_{\boldsymbol{T}}$.

Theorem 3 (Error analysis of Algorithm 2) *There exist universal constants $C, C', c, c' > 0$ such that the following holds. Pick any $\delta \in (0, 1)$. If*

$$\epsilon_{\boldsymbol{P}} \leq c \cdot \frac{\lambda_k/\lambda_1}{k}, \qquad \epsilon_{\boldsymbol{T}} \leq c' \cdot \frac{\lambda_k \sigma_k^{3/2}}{k},$$
$$N \geq C\left(\log(k) + \log\left(\log\left(\frac{\lambda_1\sigma_k^{3/2}}{\epsilon_{\boldsymbol{T}}} + \frac{1}{\epsilon_{\boldsymbol{P}}}\right)\right)\right),$$
$$L \geq \mathrm{poly}(k)\log(1/\delta)$$

(for some fixed polynomial $\mathrm{poly}(k)$ *specified in Theorem 5.1 of [5]), then with probability at least* $1-\delta$*, Algorithm 2 returns* $\{(\hat{\boldsymbol{v}}_i, \lambda_i) : i \in [k]\}$ *satisfying, after appropriate reordering,*

Algorithm 2 Plug-in estimator based on orthogonal tensor decomposition

input positive integer $k \in \mathbb{N}$, symmetric matrix $\widehat{\boldsymbol{P}} \in \mathbb{R}^{d\times d}$, symmetric tensor $\widehat{\boldsymbol{T}} \in \mathbb{R}^{d\times d\times d}$, number of iterations L and N.
output $\{(\hat{\boldsymbol{v}}_i, \hat{\lambda}_i) : i \in [k]\}$.
1: Compute eigendecomposition of $\widehat{\boldsymbol{P}}$, let $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_k$ be the top k eigenvalues, and let $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots, \boldsymbol{\zeta}_k \in \mathbb{R}^d$ be the corresponding orthonormal eigenvectors.
(Halt and signal failure if $\eta_k \leq 0$.)
2: Set $\widehat{\boldsymbol{W}} := [\boldsymbol{\zeta}_1|\boldsymbol{\zeta}_2|\cdots|\boldsymbol{\zeta}_k]\mathbf{diag}(1/\sqrt{\eta_1}, 1/\sqrt{\eta_2}, \ldots, 1/\sqrt{\eta_k})$.
($\widehat{\boldsymbol{W}}\widehat{\boldsymbol{W}}^\top$ is the pseudoinverse of a rank-k approximation to $\widehat{\boldsymbol{P}}$.)
3: Initialize $\widetilde{\boldsymbol{T}} := \widehat{\boldsymbol{T}}$.
4: **for** $i = 1$ to k **do**
5: **for** $\tau = 1$ to L **do**
6: Draw $\boldsymbol{\theta}_0^{(\tau)}$ uniformly at random from the unit sphere in $\mathbb{R}^k$.
7: **for** $t = 1$ to N **do**
8: Compute power iteration update:

$$\boldsymbol{u} := \widetilde{\boldsymbol{T}}(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{W}}\theta_{t-1}^{(\tau)}, \widehat{\boldsymbol{W}}\theta_{t-1}^{(\tau)}), \qquad \boldsymbol{\theta}_t^{(\tau)} := \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2}. \tag{$\star$}$$

9: **end for**
10: **end for**
11: Let

$$\tau^* := \arg\max_{\tau\in[L]} \widetilde{\boldsymbol{T}}(\widehat{\boldsymbol{W}}\boldsymbol{\theta}_N^{(\tau)}, \widehat{\boldsymbol{W}}\boldsymbol{\theta}_N^{(\tau)}, \widehat{\boldsymbol{W}}\boldsymbol{\theta}_N^{(\tau)})$$

and execute N more power iteration updates ($\star$) starting from $\boldsymbol{\theta}_N^{(\tau^*)}$ to obtain $\hat{\boldsymbol{\theta}}_i$.
12: Let

$$\hat{\boldsymbol{v}}_i := (\widehat{\boldsymbol{W}}^\top)^\dagger\hat{\boldsymbol{\theta}}_i, \qquad \hat{\lambda}_i := \widetilde{\boldsymbol{T}}(\widehat{\boldsymbol{W}}\hat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{W}}\hat{\boldsymbol{\theta}}_i, \widehat{\boldsymbol{W}}\hat{\boldsymbol{\theta}}_i).$$

13: Deflate $\widetilde{\boldsymbol{T}}$: set $\widetilde{\boldsymbol{T}} := \widetilde{\boldsymbol{T}} - \hat{\lambda}_i\hat{\boldsymbol{v}}_i \otimes \hat{\boldsymbol{v}}_i \otimes \hat{\boldsymbol{v}}_i$.
14: **end for**

$$\|\hat{\boldsymbol{v}}_i - \boldsymbol{o}_i\|_2 \leq C' \cdot \left(\frac{1}{\lambda_i}\cdot\frac{1}{\sigma_k^2}\cdot\epsilon_{\boldsymbol{T}} + \left(\frac{\lambda_1}{\lambda_i}\cdot\frac{1}{\sqrt{\sigma_k}} + 1\right)\cdot\epsilon_{\boldsymbol{P}}\right),$$
$$|\hat{\lambda}_i - \lambda_i| \leq C' \cdot \left(\frac{1}{\sigma_k^{3/2}}\cdot\epsilon_{\boldsymbol{T}} + \lambda_1\epsilon_{\boldsymbol{P}}\right)$$

for all $i \in [k]$.

Proof We assume without loss of generality that $c_{i,2} = 1$ for all $i \in [k]$, so $\boldsymbol{P} = \boldsymbol{O}\boldsymbol{O}^\top$, and $\lambda_i = c_{i,3}$ for all $i \in [k]$. By Lemma 8 from [28], if $\widehat{\boldsymbol{\Pi}}$ is the orthogonal projection to $\mathrm{span}\{\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots, \boldsymbol{\zeta}_k\}$ (w.r.t. the standard inner product), and $\boldsymbol{\Pi}$ is the orthogonal projection to $\mathrm{range}(\boldsymbol{O})$ (again w.r.t. the standard inner product), then

$$\|(\boldsymbol{I}_d - \widehat{\boldsymbol{\Pi}})\boldsymbol{\Pi}\|_2 \leq 1.5\epsilon_{\boldsymbol{P}}. \tag{5}$$

Note that $(\widehat{\boldsymbol{W}}^\top)^\dagger = [\boldsymbol{\zeta}_1|\boldsymbol{\zeta}_2|\cdots|\boldsymbol{\zeta}_k]\mathbf{diag}(\sqrt{\eta_1}, \sqrt{\eta_2}, \ldots, \sqrt{\eta_k})$ and therefore $\widehat{\boldsymbol{\Pi}} = (\widehat{\boldsymbol{W}}^\top)^\dagger\widehat{\boldsymbol{W}}^\top$.

Let $\boldsymbol{W} := \widehat{\boldsymbol{W}}(\widehat{\boldsymbol{W}}^\top \boldsymbol{P}\widehat{\boldsymbol{W}})^{\dagger 1/2}$. By Lemma 9 and Lemma 11 from [28],

$$\|\widehat{\boldsymbol{W}}\|_2 \le \frac{1}{\sqrt{(1-\epsilon_{\boldsymbol{P}})\sigma_k}}, \tag{6}$$

$$\|(\widehat{\boldsymbol{W}} - \boldsymbol{W})^\top \boldsymbol{O}\|_2 \le \sqrt{1+1.5\epsilon_{\boldsymbol{P}}}\,1.5\epsilon_{\boldsymbol{P}}, \tag{7}$$

$$\|\widehat{\boldsymbol{T}}(\widehat{\boldsymbol{W}},\widehat{\boldsymbol{W}},\widehat{\boldsymbol{W}}) - \boldsymbol{T}(\boldsymbol{W},\boldsymbol{W},\boldsymbol{W})\|_2 \le \frac{2}{\sigma_k^{3/2}}\cdot\epsilon_{\boldsymbol{T}} + 6\lambda_1\cdot\epsilon_{\boldsymbol{P}} =: \varepsilon. \tag{8}$$

Note that

$$\boldsymbol{T}(\boldsymbol{W},\boldsymbol{W},\boldsymbol{W}) = \sum_{i=1}^{k} \lambda_i (\boldsymbol{W}^\top \boldsymbol{o}_i)\otimes(\boldsymbol{W}^\top \boldsymbol{o}_i)\otimes(\boldsymbol{W}^\top \boldsymbol{o}_i)$$

where $\{\boldsymbol{W}^\top \boldsymbol{o}_i : i \in [k]\}$ are orthonormal. Using (8), Theorem 5.1 from [5] implies the following: if $\varepsilon \le C_1\lambda_k/k$, $N \ge C_2(\log(k) + \log\log(\lambda_1/\varepsilon))$, and $L \ge \mathrm{poly}(k)\log(1/\delta)$, then with probability at least $1-\delta$, the robust tensor power method returns $\hat{\boldsymbol{\theta}}_i$ and $\hat{\lambda}_i$ satisfying (after appropriate reordering)

$$\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{W}^\top \boldsymbol{o}_i\|_2 \le \frac{8\varepsilon}{\lambda_i}, \qquad |\hat{\lambda}_i - \lambda_i| \le 5\varepsilon. \tag{9}$$

Therefore,

$$\begin{aligned}
\|\hat{\boldsymbol{v}}_i - \boldsymbol{o}_i\|_2 &= \Big\| \big((\widehat{\boldsymbol{W}}^\top)^\dagger \hat{\boldsymbol{\theta}}_i - (\widehat{\boldsymbol{W}}^\top)^\dagger \boldsymbol{W}^\top \boldsymbol{o}_i\big) \\
&\quad + \big((\widehat{\boldsymbol{W}}^\top)^\dagger \boldsymbol{W}^\top \boldsymbol{o}_i - (\widehat{\boldsymbol{W}}^\top)^\dagger \widehat{\boldsymbol{W}}^\top \boldsymbol{o}_i\big) + \big((\widehat{\boldsymbol{W}}^\top)^\dagger \widehat{\boldsymbol{W}}^\top \boldsymbol{o}_i - \boldsymbol{o}_i\big)\Big\|_2 \\
&\le \big\|(\widehat{\boldsymbol{W}}^\top)^\dagger \hat{\boldsymbol{\theta}}_i - (\widehat{\boldsymbol{W}}^\top)^\dagger \boldsymbol{W}^\top \boldsymbol{o}_i\big\|_2 \\
&\quad + \big\|(\widehat{\boldsymbol{W}}^\top)^\dagger \boldsymbol{W}^\top \boldsymbol{o}_i - (\widehat{\boldsymbol{W}}^\top)^\dagger \widehat{\boldsymbol{W}}^\top \boldsymbol{o}_i\big\|_2 + \big\|(\widehat{\boldsymbol{W}}^\top)^\dagger \widehat{\boldsymbol{W}}^\top \boldsymbol{o}_i - \boldsymbol{o}_i\big\|_2 \\
&\le \|(\widehat{\boldsymbol{W}}^\top)^\dagger\|_2 \Big(\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{W}^\top \boldsymbol{o}_i\|_2 + \|(\widehat{\boldsymbol{W}} - \boldsymbol{W})^\top \boldsymbol{O}\|_2 \Big) \\
&\quad + \|(\widehat{\boldsymbol{\Pi}} - \boldsymbol{I}_d)\boldsymbol{\Pi}\boldsymbol{o}_i\|_2 \\
&\le \frac{1}{\sqrt{(1-\epsilon_{\boldsymbol{P}})\sigma_k}} \left(\frac{8\varepsilon}{\lambda_i} + \sqrt{1+1.5\epsilon_{\boldsymbol{P}}}\,1.5\epsilon_{\boldsymbol{P}} \right) + 1.5\epsilon_{\boldsymbol{P}}
\end{aligned}$$

using (5), (6), (9), and (7). □

We remark that Theorem 3 immediately implies estimation consistency under appropriate assumptions on the σ_i and λ_i, and it is straightforward to obtain finite sample guarantees using concentration arguments to bound $\|\widehat{\boldsymbol{P}} - \boldsymbol{P}\|_2$ and $\|\widehat{\boldsymbol{T}} - \boldsymbol{T}\|_2$. We leave this as an exercise for the reader.

Acknowledgments We thank Kamalika Chaudhuri, Adam Kalai, Percy Liang, Chris Meek, David Sontag, and Tong Zhang for valuable insights. We also thank Rong Ge for sharing preliminary results (in [8]) and the anonymous reviewers for their comments, suggestions, and pointers to references. Part of this work was completed while DH was a postdoctoral researcher at Microsoft Research New England, and while DPF, YKL, and AA were visiting the same lab. AA is supported in part by Microsoft Faculty Fellowship, NSF Career award CCF-1254106, NSF Award CCF-1219234, NSF BIGDATA IIS-1251267 and ARO YIP Award W911NF-13-1-0084.

References

1. Achlioptas, D., McSherry, F.: On spectral learning of mixtures of distributions. Eighteenth Annual Conference on Learning Theory, pp. 458–469. Springer, Bertinoro (2005)
2. Anandkumar, A., Chaudhuri, K., Hsu, D., Kakade, S.M., Song, L., Zhang, T.: Spectral methods for learning multivariate latent tree structure. Adv. Neural Inf. Process. Syst. **24**, 2025–2033 (2011)
3. Anandkumar, A., Foster, D.P., Hsu, D., Kakade, S.M., Liu, Y.K.: A spectral algorithm for latent Dirichlet allocation. Adv. Neural Inf. Process. Syst. **25**, 917–925 (2012)
4. Anandkumar, A., Foster, D.P., Hsu, D., Kakade, S.M., Liu, Y.K.: Two SVDs suffice: spectral decompositions for probabilistic topic models and latent Dirichlet allocation (2012). arXiv:1204.6703v1
5. Anandkumar, A., Ge, R., Hsu, D., Kakade, S.M., Telgarsky, M.: Tensor decompositions for learning latent variable models. J. Mach. Learn. Res. (2014). To appear.
6. Anandkumar, A., Hsu, D., Kakade, S.M.: A method of moments for mixture models and hidden Markov models. In: Twenty-Fifth Annual Conference on Learning Theory, vol. 23, pp. 33.1-33.34 (2012)
7. Ando, R., Zhang, T.: Two-view feature generation model for semi-supervised learning. In: Twenty-Fourth International Conference on Machine Learning, pp. 25–32 (2007)
8. Arora, S., Ge, R., Moitra, A.: Learning topic models – going beyond SVD. In: Fifty-Third IEEE Annual Symposium on Foundations of Computer Science, pp. 1–10 (2012)
9. Arora, S., Ge, R., Moitra, A., Sachdeva, S.: Provable ICA with unknown Gaussian noise, with implications for Gaussian mixtures and autoencoders. Adv. Neural Inf. Process. Syst. **25**, 2375–2383 (2012)
10. Arora, S., Kannan, R.: Learning mixtures of separated nonspherical Gaussians. Ann. Appl. Probab. **15**(1A), 69–92 (2005)
11. Belkin, M., Sinha, K.: Polynomial learning of distribution families. In: Fifty-First Annual IEEE Symposium on Foundations of Computer Science, pp. 103–112 (2010)
12. Blei, D.M., Ng, A., Jordan, M.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
13. Canny, J.: GaP: A factor model for discrete data. In: Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 122–129 (2004)
14. Cardoso, J.F., Comon, P.: Independent component analysis, a survey of some algebraic methods. In: IEEE International Symposium on Circuits and Systems, pp. 93–96 (1996)
15. Chang, J.T.: Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. **137**, 51–73 (1996)
16. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: Twenty-Sixth Annual International Conference on Machine Learning, pp. 129–136 (2009)
17. Chaudhuri, K., Rao, S.: Learning mixtures of product distributions using correlations and independence. In: Twenty-First Annual Conference on Learning Theory, pp. 9–20 (2008)
18. Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press, Waltham (2010)
19. Dasgupta, S.: Learning mixutres of Gaussians. In: Fortieth Annual IEEE Symposium on Foundations of Computer Science, pp. 634–644 (1999)
20. Dasgupta, S., Schulman, L.: A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. J. Mach. Learn. Res. **8**, 203–226 (2007)
21. Frieze, A.M., Jerrum, M., Kannan, R.: Learning linear transformations. In: Thirty-Seventh Annual Symposium on Foundations of Computer Science, pp. 359–368 (1996)
22. Griffiths, T.: Gibbs sampling in the generative model of latent Dirichlet allocation. Tech. rep., Stanford University (2002)

23. Harshman, R.: Foundations of the PARAFAC procedure: model and conditions for an 'explanatory' multi-mode factor analysis. Tech. rep., UCLA Working Papers in Phonetics (1970)
24. Hitchcock, F.: The expression of a tensor or a polyadic as a sum of products. J. Math. Phys. **6**, 164–189 (1927)
25. Hitchcock, F.: Multiple invariants and generalized rank of a p-way matrix or tensor. J. Math. Phys. **7**, 39–79 (1927)
26. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the Twenty-Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999)
27. Hotelling, H.: The most predictable criterion. J. Educ. Psychol. **26**(2), 139–142 (1935)
28. Hsu, D., Kakade, S.M.: Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In: Fourth Innovations in Theoretical Computer Science (2013)
29. Hsu, D., Kakade, S.M., Zhang, T.: A spectral algorithm for learning hidden Markov models. J. Comput. Syst. Sci. **78**(5), 1460–1480 (2012). http://www.sciencedirect.com/science/article/pii/S0022000012000244
30. Jutten, C., Herault, J.: Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. Signal Process. **24**, 1–10 (1991)
31. Kakade, S.M., Foster, D.P.: Multi-view regression via canonical correlation analysis. In: Twentieth Annual Conference on Learning Theory, pp. 82–96 (2007)
32. Kalai, A.T., Moitra, A., Valiant, G.: Efficiently learning mixtures of two Gaussians. In: Forty-second ACM Symposium on Theory of Computing, pp. 553–562 (2010)
33. Kannan, R., Salmasian, H., Vempala, S.: The spectral method for general mixture models. SIAM J. Comput. **38**(3), 1141–1156 (2008)
34. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009)
35. Kruskal, J.B.: Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. Linear Algebra Its Appl. **18**(2), 95–138 (1977)
36. Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. Nature **401**, 788–791 (1999)
37. Leurgans, S., Ross, R., Abel, R.: A decomposition for three-way arrays. SIAM J. Matrix Anal. Appl. **14**(4), 1064–1083 (1993)
38. Moitra, A., Valiant, G.: Settling the polynomial learnability of mixtures of Gaussians. In: Fifty-First Annual IEEE Symposium on Foundations of Computer Science, pp. 93–102 (2010)
39. Mossel, E., Roch, S.: Learning nonsingular phylogenies and hidden Markov models. Ann. Appl. Probab. **16**(2), 583–614 (2006)
40. Nguyen, P.Q., Regev, O.: Learning a parallelepiped: cryptanalysis of GGH and NTRU signatures. J. Cryptol. **22**(2), 139–160 (2009)
41. Papadimitriou, C.H., Raghavan, P., Tamaki, H., Vempala, S.: Latent semantic indexing: a probabilistic analysis. J. Comput. Syst. Sci. **61**(2), 217–235 (2000)
42. Pearson, K.: Contributions to the mathematical theory of evolution. Philos. Trans. R. Soc. Lond. A **185**, 71–110 (1894)
43. Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. **26**(2), 195–239 (1984)
44. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**(3), 279–311 (1966)
45. Vempala, S., Wang, G.: A spectral algorithm for learning mixtures models. J. Comput. Syst. Sci. **68**(4), 841–860 (2004)
46. Zou, J., Hsu, D., Parkes, D., Adams, R.: Contrastive learning using spectral methods. Adv. Neural Inf. Process. Syst. **26**, 2238–2246 (2013)