Tensor Decompositions for Learning Latent Variable Models (A Survey for ALT)

Anima Anandkumar[1], Rong Ge[2], Daniel Hsu[3], Sham M. Kakade[4(✉)], and Matus Telgarsky[5]

[1] University of California, Irvine, USA
[2] Microsoft Research, New England, USA
[3] Columbia University, New York, USA
[4] Rutgers University, New Brunswick, USA
skakade@microsoft.cm
[5] University of Michigan, Ann Arbor, USA

Abstract. This note is a short version of that in [1]. It is intended as a survey for the 2015 Algorithmic Learning Theory (ALT) conference.

This work considers a computationally and statistically efficient parameter estimation method for a wide class of latent variable models—including Gaussian mixture models, hidden Markov models, and latent Dirichlet allocation—which exploits a certain tensor structure in their low-order observable moments (typically, of second- and third-order). Specifically, parameter estimation is reduced to the problem of extracting a certain (orthogonal) decomposition of a symmetric tensor derived from the moments; this decomposition can be viewed as a natural generalization of the singular value decomposition for matrices. Although tensor decompositions are generally intractable to compute, the decomposition of these specially structured tensors can be efficiently obtained by a variety of approaches, including power iterations and maximization approaches (similar to the case of matrices). A detailed analysis of a robust tensor power method is provided, establishing an analogue of Wedin's perturbation theorem for the singular vectors of matrices. This implies a robust and computationally tractable estimation approach for several popular latent variable models.

1 Introduction

The method of moments is a classical parameter estimation technique [29] from statistics which has proved invaluable in a number of application domains. The basic paradigm is simple and intuitive: (i) compute certain statistics of the data — often empirical moments such as means and correlations — and (ii) find model parameters that give rise to (nearly) the same corresponding population quantities. In a number of cases, the method of moments leads to consistent estimators which can be efficiently computed; this is especially relevant in the context of latent variable models, where standard maximum likelihood approaches are typically computationally prohibitive, and heuristic methods can be unreliable and difficult to validate with high-dimensional data. Furthermore, the method

K. Chaudhuri et al. (Eds.): ALT 2015, LNAI 9355, pp. 19–38, 2015.
DOI: 10.1007/978-3-319-24486-0_2

of moments can be viewed as complementary to the maximum likelihood approach; simply taking a single step of Newton-Ralphson on the likelihood function starting from the moment based estimator [22] often leads to the best of both worlds: a computationally efficient estimator that is (asymptotically) statistically optimal.

The primary difficulty in learning latent variable models is that the latent (hidden) state of the data is not directly observed; rather only observed variables correlated with the hidden state are observed. As such, it is not evident the method of moments should fare any better than maximum likelihood in terms of computational performance: matching the model parameters to the observed moments may involve solving computationally intractable systems of multivariate polynomial equations. Fortunately, for many classes of latent variable models, there is rich structure in low-order moments (typically second- and third-order) which allow for this inverse moment problem to be solved efficiently [2,4,6,8,9,16,18,27]. What is more is that these decomposition problems are often amenable to simple and efficient iterative methods, such as gradient descent and the power iteration method.

This survey observes that a number of important and well-studied latent variable models—including Gaussian mixture models, hidden Markov models, and Latent Dirichlet allocation—share a certain structure in their low-order moments, and this permits certain tensor decomposition approaches to parameter estimation. In particular, this decomposition can be viewed as a natural generalization of the singular value decomposition for matrices.

While much of this (or similar) structure was implicit in several previous works [2,4,9,16,18,27], here we make the decomposition explicit under a unified framework. Specifically, we express the observable moments as sums of rank-one terms, and reduce the parameter estimation task to the problem of extracting a symmetric orthogonal decomposition of a symmetric tensor derived from these observable moments. The problem can then be solved by a variety of approaches, including fixed-point and variational methods.

One approach for obtaining the orthogonal decomposition is the tensor power method of [21, Remark3]. We provide a convergence analysis of this method for orthogonally decomposable symmetric tensors, as well as a robust (and computationally tractable) variant. The perturbation analysis in [1] can be viewed as an analogue of Wedin's perturbation theorem for singular vectors of matrices [32], providing a bound on the error of the recovered decomposition in terms of the operator norm of the tensor perturbation.

1.1 Related Work

See [1] for a discussion of related work.

1.2 Organization

The rest of the survey is organized as follows. Section 2 reviews some basic definitions of tensors. Section 3 provides examples of a number of latent variable

models which, after appropriate manipulations of their low order moments, share a certain natural tensor structure. Section 4 reduces the problem of parameter estimation to that of extracting a certain (symmetric orthogonal) decomposition of a tensor. See [1] which states establishes an analogue of Wedin's perturbation theorem for the singular vectors of matrices.

2 Preliminaries

We introduce some tensor notations borrowed from [23]. A real *p-th order tensor* $A \in \bigotimes_{i=1}^{p} \mathbb{R}^{n_i}$ is a member of the tensor product of Euclidean spaces $\mathbb{R}^{n_i}$, $i \in [p]$. We generally restrict to the case where $n_1 = n_2 = \cdots = n_p = n$, and simply write $A \in \bigotimes^p \mathbb{R}^n$. For a vector $v \in \mathbb{R}^n$, we use $v^{\otimes p} := v \otimes v \otimes \cdots \otimes v \in \bigotimes^p \mathbb{R}^n$ to denote its p-th tensor power. As is the case for vectors (where $p = 1$) and matrices (where $p = 2$), we may identify a p-th order tensor with the p-way array of real numbers $[A_{i_1,i_2,\ldots,i_p} : i_1, i_2, \ldots, i_p \in [n]]$, where $A_{i_1,i_2,\ldots,i_p}$ is the $(i_1, i_2, \ldots, i_p)$-th coordinate of A (with respect to a canonical basis).

We can consider A to be a multilinear map in the following sense: for a set of matrices $\{V_i \in \mathbb{R}^{n \times m_i} : i \in [p]\}$, the $(i_1, i_2, \ldots, i_p)$-th entry in the p-way array representation of $A(V_1, V_2, \ldots, V_p) \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_p}$ is

$$[A(V_1, V_2, \ldots, V_p)]_{i_1,i_2,\ldots,i_p} \coloneqq \sum_{j_1,j_2,\ldots,j_p \in [n]} A_{j_1,j_2,\ldots,j_p} [V_1]_{j_1,i_1} [V_2]_{j_2,i_2} \cdots [V_p]_{j_p,i_p}.$$

Note that if A is a matrix ($p = 2$), then

$$A(V_1, V_2) = V_1^\top A V_2.$$

Similarly, for a matrix A and vector $v \in \mathbb{R}^n$, we can express Av as

$$A(I, v) = Av \in \mathbb{R}^n,$$

where I is the $n \times n$ identity matrix. As a final example of this notation, observe

$$A(e_{i_1}, e_{i_2}, \ldots, e_{i_p}) = A_{i_1,i_2,\ldots,i_p},$$

where $\{e_1, e_2, \ldots, e_n\}$ is the canonical basis for $\mathbb{R}^n$.

Most tensors $A \in \bigotimes^p \mathbb{R}^n$ considered in this work will be *symmetric* (sometimes called *supersymmetric*), which means that their p-way array representations are invariant to permutations of the array indices: *i.e.*, for all indices $i_1, i_2, \ldots, i_p \in [n]$, $A_{i_1,i_2,\ldots,i_p} = A_{i_{\pi(1)},i_{\pi(2)},\ldots,i_{\pi(p)}}$ for any permutation π on $[p]$. It can be checked that this reduces to the usual definition of a symmetric matrix for $p = 2$.

The *rank* of a p-th order tensor $A \in \bigotimes^p \mathbb{R}^n$ is the smallest non-negative integer k such that $A = \sum_{j=1}^{k} u_{1,j} \otimes u_{2,j} \otimes \cdots \otimes u_{p,j}$ for some $u_{i,j} \in \mathbb{R}^n, i \in [p], j \in [k]$, and the *symmetric rank* of a symmetric p-th order tensor A is the smallest non-negative integer k such that $A = \sum_{j=1}^{k} u_j^{\otimes p}$ for some $u_j \in \mathbb{R}^n, j \in [k]$ (for

even p, the definition is slightly different [11]). The notion of rank readily reduces to the usual definition of matrix rank when $p = 2$, as revealed by the singular value decomposition. Similarly, for symmetric matrices, the symmetric rank is equivalent to the matrix rank as given by the spectral theorem.

The notion of tensor (symmetric) rank is considerably more delicate than matrix (symmetric) rank. For instance, it is not clear *a priori* that the symmetric rank of a tensor should even be finite [11]. In addition, removal of the best rank-1 approximation of a (general) tensor may increase the tensor rank of the residual [31].

Throughout, we use $\|v\| = (\sum_i v_i^2)^{1/2}$ to denote the Euclidean norm of a vector v, and $\|M\|$ to denote the spectral (operator) norm of a matrix. We also use $\|T\|$ to denote the operator norm of a tensor, which we define later.

3 Tensor Structure in Latent Variable Models

In this section, we give several examples of latent variable models whose low-order moments can be written as symmetric tensors of low symmetric rank; many of these examples can be deduced using the techniques developed in [25]. The basic form is demonstrated in Theorem 1 for the first example, and the general pattern will emerge from subsequent examples.

3.1 Exchangeable Single Topic Models

We first consider a simple bag-of-words model for documents in which the words in the document are assumed to be *exchangeable*. Recall that a collection of random variables $x_1, x_2, \ldots, x_\ell$ are exchangeable if their joint probability distribution is invariant to permutation of the indices. The well-known De Finetti's theorem [5] implies that such exchangeable models can be viewed as mixture models in which there is a latent variable h such that $x_1, x_2, \ldots, x_\ell$ are conditionally i.i.d. given h (see Figure 1(a) for the corresponding graphical model) and the conditional distributions are identical at all the nodes.

In our simplified topic model for documents, the latent variable h is interpreted as the (sole) topic of a given document, and it is assumed to take only a finite number of distinct values. Let k be the number of distinct topics in the corpus, d be the number of distinct words in the vocabulary, and $\ell \geq 3$ be the number of words in each document. The generative process for a document is as follows: the document's topic is drawn according to the discrete distribution specified by the probability vector $w := (w_1, w_2, \ldots, w_k) \in \Delta^{k-1}$. This is modeled as a discrete random variable h such that

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

Given the topic h, the document's ℓ words are drawn independently according to the discrete distribution specified by the probability vector $\mu_h \in \Delta^{d-1}$. It will be

convenient to represent the ℓ words in the document by d-dimensional random *vectors* $x_1, x_2, \ldots, x_\ell \in \mathbb{R}^d$. Specifically, we set

$$x_t = e_i \quad \text{if and only if} \quad \text{the } t\text{-th word in the document is } i, \quad t \in [\ell],$$

where $e_1, e_2, \ldots e_d$ is the standard coordinate basis for $\mathbb{R}^d$.

One advantage of this encoding of words is that the (cross) moments of these random vectors correspond to joint probabilities over words. For instance, observe that

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

so the (i,j)-the entry of the matrix $\mathbb{E}[x_1 \otimes x_2]$ is $\Pr[\text{1st word} = i, \text{2nd word} = j]$. More generally, the $(i_1, i_2, \ldots, i_\ell)$-th entry in the tensor $\mathbb{E}[x_1 \otimes x_2 \otimes \cdots \otimes x_\ell]$ is $\Pr[\text{1st word} = i_1, \text{2nd word} = i_2, \ldots, \ell\text{-th word} = i_\ell]$. This means that estimating cross moments, say, of $x_1 \otimes x_2 \otimes x_3$, is the same as estimating joint probabilities of the first three words over all documents. (Recall that we assume that each document has at least three words.)

The second advantage of the vector encoding of words is that the conditional expectation of x_t given $h = j$ is simply μ_j, the vector of word probabilities for topic j:

$$\mathbb{E}[x_t | h = j] \ = \ \sum_{i=1}^{d} \Pr[t\text{-th word} = i | h = j]\ e_i \ = \ \sum_{i=1}^{d} [\mu_j]_i\ e_i \ = \ \mu_j, \quad j \in [k]$$

(where $[\mu_j]_i$ is the i-th entry in the vector μ_j). Because the words are conditionally independent given the topic, we can use this same property with conditional cross moments, say, of x_1 and x_2:

$$\mathbb{E}[x_1 \otimes x_2 | h = j] \ = \ \mathbb{E}[x_1 | h = j] \otimes \mathbb{E}[x_2 | h = j] \ = \ \mu_j \otimes \mu_j, \quad j \in [k].$$

This and similar calculations lead one to the following theorem.

Theorem 1 ([4]). *If*

$$\begin{aligned}M_2 &:= \mathbb{E}[x_1 \otimes x_2] \\ M_3 &:= \mathbb{E}[x_1 \otimes x_2 \otimes x_3],\end{aligned}$$

then

$$\begin{aligned}M_2 &= \sum_{i=1}^{k} w_i\ \mu_i \otimes \mu_i \\ M_3 &= \sum_{i=1}^{k} w_i\ \mu_i \otimes \mu_i \otimes \mu_i.\end{aligned}$$

As we will see in Section 4.3, the structure of M_2 and M_3 revealed in Theorem 1 implies that the topic vectors $\mu_1, \mu_2, \ldots, \mu_k$ can be estimated by computing a certain symmetric tensor decomposition. Moreover, due to exchangeability, all triples (resp., pairs) of words in a document—and not just the first three (resp., two) words—can be used in forming M_3 (resp., M_2).

3.2 Beyond Raw Moments

In the single topic model above, the raw (cross) moments of the observed words directly yield the desired symmetric tensor structure. In some other models, the raw moments do not explicitly have this form. Here, we show that the desired tensor structure can be found through various manipulations of different moments.

Spherical Gaussian Mixtures. We now consider a mixture of k Gaussian distributions with spherical covariances. We start with the simpler case where all of the covariances are identical; this probabilistic model is closely related to the (non-probabilistic) k-means clustering problem [24]. We then consider the case where the spherical variances may differ.

Common Covariance. Let w_i be the probability of choosing component $i \in [k]$, $\{\mu_1, \mu_2, \ldots, \mu_k\} \subset \mathbb{R}^d$ be the component mean vectors, and $\sigma^2 I$ be the common covariance matrix. An observation in this model is given by

$$x := \mu_h + z,$$

where h is the discrete random variable with $\Pr[h = i] = w_i$ for $i \in [k]$ (similar to the exchangeable single topic model), and $z \sim \mathcal{N}(0, \sigma^2 I)$ is an independent multivariate Gaussian random vector in $\mathbb{R}^d$ with zero mean and spherical covariance $\sigma^2 I$.

The Gaussian mixture model differs from the exchangeable single topic model in the way observations are generated. In the single topic model, we observe multiple draws (words in a particular document) $x_1, x_2, \ldots, x_\ell$ given the same fixed h (the topic of the document). In contrast, for the Gaussian mixture model, every realization of x corresponds to a different realization of h.

Theorem 2 ([16]). *Assume $d \geq k$. The variance σ^2 is the smallest eigenvalue of the covariance matrix $\mathbb{E}[x \otimes x] - \mathbb{E}[x] \otimes \mathbb{E}[x]$. Furthermore, if*

$$M_2 := \mathbb{E}[x \otimes x] - \sigma^2 I$$

$$M_3 := \mathbb{E}[x \otimes x \otimes x] - \sigma^2 \sum_{i=1}^{d} \big(\mathbb{E}[x] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[x] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[x]\big),$$

then

$$M_2 = \sum_{i=1}^{k} w_i \ \mu_i \otimes \mu_i$$

$$M_3 = \sum_{i=1}^{k} w_i \ \mu_i \otimes \mu_i \otimes \mu_i.$$

Differing Covariances. See [1] for the case is where each component may have a *different* spherical covariance.

Independent Component Analysis (ICA). The standard model for ICA [7,10,12,19], in which independent signals are linearly mixed and corrupted with Gaussian noise before being observed, is specified as follows. Let $h \in \mathbb{R}^k$ be a latent random *vector* with independent coordinates, $A \in \mathbb{R}^{d \times k}$ the mixing matrix, and z be a multivariate Gaussian random vector. The random vectors h and z are assumed to be independent. The observed random vector is

$$x := Ah + z.$$

Let μ_i denote the i-th column of the mixing matrix A.

Theorem 3 ([12]). *Define*

$$M_4 := \mathbb{E}[x \otimes x \otimes x \otimes x] - T$$

where T is the fourth-order tensor with

$$[T]_{i_1,i_2,i_3,i_4} := \mathbb{E}[x_{i_1}x_{i_2}]\mathbb{E}[x_{i_3}x_{i_4}] + \mathbb{E}[x_{i_1}x_{i_3}]\mathbb{E}[x_{i_2}x_{i_4}] + \mathbb{E}[x_{i_1}x_{i_4}]\mathbb{E}[x_{i_2}x_{i_3}],$$

*where $1 \leq i_1, i_2, i_3, i_4 \leq k$ (*i.e.*, T is the fourth derivative tensor of the function $v \mapsto 8^{-1}\mathbb{E}[(v^\top x)^2]^2$, so M_4 is the fourth cumulant tensor). Let $\kappa_i := \mathbb{E}[h_i^4] - 3$ for each $i \in [k]$. Then*

$$M_4 = \sum_{i=1}^{k} \kappa_i \ \mu_i \otimes \mu_i \otimes \mu_i \otimes \mu_i.$$

Note that κ_i corresponds to the excess kurtosis, a measure of non-Gaussianity as $\kappa_i = 0$ if h_i is a standard normal random variable. Furthermore, note that A is not identifiable if h is a multivariate Gaussian.

We may derive forms similar to that of M_2 and M_3 from Theorem 1 using M_4 by observing that

$$M_4(I, I, u, v) = \sum_{i=1}^{k} \kappa_i (\mu_i^\top u)(\mu_i^\top v) \ \mu_i \otimes \mu_i,$$

$$M_4(I, I, I, v) = \sum_{i=1}^{k} \kappa_i (\mu_i^\top v) \ \mu_i \otimes \mu_i \otimes \mu_i$$

for any vectors $u, v \in \mathbb{R}^d$.

Latent Dirichlet Allocation (LDA). An increasingly popular class of latent variable models are *mixed membership models*, where each datum may belong to several different latent classes simultaneously. LDA is one such model for

the case of document modeling; here, each document corresponds to a mixture over topics (as opposed to just a single topic). The distribution over such topic mixtures is a Dirichlet distribution $\mathrm{Dir}(\alpha)$ with parameter vector $\alpha \in \mathbb{R}^k_{++}$ with strictly positive entries; its density over the probability simplex $\Delta^{k-1} := \{v \in \mathbb{R}^k : v_i \in [0,1] \forall i \in [k],\ \sum_{i=1}^k v_i = 1\}$ is given by

$$p_\alpha(h) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k h_i^{\alpha_i - 1}, \quad h \in \Delta^{k-1}$$

where

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

As before, the k topics are specified by probability vectors $\mu_1, \mu_2, \ldots, \mu_k \in \Delta^{d-1}$. To generate a document, first draw the topic mixture $h = (h_1, h_2, \ldots, h_k) \sim \mathrm{Dir}(\alpha)$, and then conditioned on h, we draw ℓ words $x_1, x_2, \ldots, x_\ell$ independently from the discrete distribution specified by the probability vector $\sum_{i=1}^k h_i \mu_i$ (*i.e.*, for each x_t, we independently sample a topic j according to h and then sample x_t according to μ_j). Again, we encode a word x_t by setting $x_t = e_i$ iff the t-th word in the document is i.

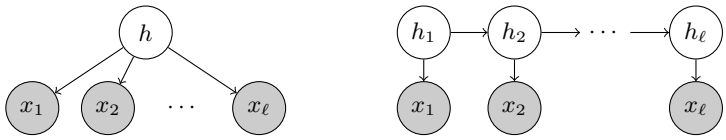
The parameter α_0 (the sum of the "pseudo-counts") characterizes the concentration of the distribution. As $\alpha_0 \to 0$, the distribution degenerates to a single topic model (*i.e.*, the limiting density has, with probability 1, exactly one entry of h being 1 and the rest are 0). At the other extreme, if $\alpha = (c, c, \ldots, c)$ for some scalar $c > 0$, then as $\alpha_0 = ck \to \infty$, the distribution of h becomes peaked around the uniform vector $(1/k, 1/k, \ldots, 1/k)$ (furthermore, the distribution behaves like a product distribution). We are typically interested in the case where α_0 is small (*e.g.*, a constant independent of k), whereupon h typically has only a few large entries. This corresponds to the setting where the documents are mainly comprised of just a few topics.

Theorem 4 ([2]). *Define*

$$\begin{aligned}
M_1 &:= \mathbb{E}[x_1] \\
M_2 &:= \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0 + 1} M_1 \otimes M_1 \\
M_3 &:= \mathbb{E}[x_1 \otimes x_2 \otimes x_3] \\
&\quad - \frac{\alpha_0}{\alpha_0 + 2} \Big(\mathbb{E}[x_1 \otimes x_2 \otimes M_1] + \mathbb{E}[x_1 \otimes M_1 \otimes x_2] + \mathbb{E}[M_1 \otimes x_1 \otimes x_2] \Big) \\
&\quad + \frac{2\alpha_0^2}{(\alpha_0 + 2)(\alpha_0 + 1)} M_1 \otimes M_1 \otimes M_1.
\end{aligned}$$

Then

$$\begin{aligned}
M_2 &= \sum_{i=1}^k \frac{\alpha_i}{(\alpha_0 + 1)\alpha_0}\ \mu_i \otimes \mu_i \\
M_3 &= \sum_{i=1}^k \frac{2\alpha_i}{(\alpha_0 + 2)(\alpha_0 + 1)\alpha_0}\ \mu_i \otimes \mu_i \otimes \mu_i.
\end{aligned}$$



(a) Multi-view models (b) Hidden Markov model

Fig. 1. Examples of latent variable models.

Note that α_0 needs to be known to form M_2 and M_3 from the raw moments. This, however, is a much weaker than assuming that the entire distribution of h is known (*i.e.*, knowledge of the whole parameter vector α).

3.3 Multi-view Models

Multi-view models (also sometimes called naïve Bayes models) are a special class of Bayesian networks in which observed variables $x_1, x_2, \ldots, x_\ell$ are conditionally independent given a latent variable h. This is similar to the exchangeable single topic model, but here we do not require the conditional distributions of the $x_t, t \in [\ell]$ to be identical. Techniques developed for this class can be used to handle a number of widely used models including hidden Markov models (HMMs) [4,27], phylogenetic tree models [9,27], certain tree mixtures [3], and certain probabilistic grammar models [17].

As before, we let $h \in [k]$ be a discrete random variable with $\Pr[h = j] = w_j$ for all $j \in [k]$. Now consider random vectors $x_1 \in \mathbb{R}^{d_1}$, $x_2 \in \mathbb{R}^{d_2}$, and $x_3 \in \mathbb{R}^{d_3}$ which are conditionally independent given h, and

$$\mathbb{E}[x_t | h = j] = \mu_{t,j}, \quad j \in [k],\ t \in \{1, 2, 3\}$$

where the $\mu_{t,j} \in \mathbb{R}^{d_t}$ are the conditional means of the x_t given $h = j$. Thus, we allow the observations $x_1, x_2, \ldots, x_\ell$ to be random vectors, parameterized only by their conditional means. Importantly, these conditional distributions may be discrete, continuous, or even a mix of both.

We first note the form for the raw (cross) moments.

Proposition 1. *We have that:*

$$\mathbb{E}[x_t \otimes x_{t'}] = \sum_{i=1}^{k} w_i\ \mu_{t,i} \otimes \mu_{t',i}, \quad \{t, t'\} \subset \{1, 2, 3\}, t \neq t'$$

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{i=1}^{k} w_i\ \mu_{1,i} \otimes \mu_{2,i} \otimes \mu_{3,i}.$$

The cross moments do not possess a symmetric tensor form when the conditional distributions are different. Nevertheless, the moments can be "symmetrized" via a simple linear transformation of x_1 and x_2 (roughly speaking,

this relates x_1 and x_2 to x_3); this leads to an expression from which the conditional means of x_3 (*i.e.*, $\mu_{3,1}, \mu_{3,2}, \ldots, \mu_{3,k}$) can be recovered. For simplicity, we assume $d_1 = d_2 = d_3 = k$; the general case (with $d_t \geq k$) is easily handled using low-rank singular value decompositions.

Theorem 5 ([2]). *Assume that the vectors* $\{\mu_{v,1}, \mu_{v,2}, \ldots, \mu_{v,k}\}$ *are linearly independent for each* $v \in \{1, 2, 3\}$. *Define*

$$\begin{aligned}
\tilde{x}_1 &:= \mathbb{E}[x_3 \otimes x_2]\mathbb{E}[x_1 \otimes x_2]^{-1} x_1 \\
\tilde{x}_2 &:= \mathbb{E}[x_3 \otimes x_1]\mathbb{E}[x_2 \otimes x_1]^{-1} x_2 \\
M_2 &:= \mathbb{E}[\tilde{x}_1 \otimes \tilde{x}_2] \\
M_3 &:= \mathbb{E}[\tilde{x}_1 \otimes \tilde{x}_2 \otimes x_3].
\end{aligned}$$

Then

$$\begin{aligned}
M_2 &= \sum_{i=1}^{k} w_i \ \mu_{3,i} \otimes \mu_{3,i} \\
M_3 &= \sum_{i=1}^{k} w_i \ \mu_{3,i} \otimes \mu_{3,i} \otimes \mu_{3,i}.
\end{aligned}$$

Hidden Markov Models. Our last example is the time-homogeneous HMM for sequences of vector-valued observations $x_1, x_2, \ldots \in \mathbb{R}^d$. Consider a Markov chain of discrete hidden states $y_1 \to y_2 \to y_3 \to \cdots$ over k possible states $[k]$; given a state y_t at time t, the observation x_t at time t (a random vector taking values in $\mathbb{R}^d$) is independent of all other observations and hidden states. See Figure 1(b).

Let $\pi \in \Delta^{k-1}$ be the initial state distribution (*i.e.*, the distribution of y_1), and $T \in \mathbb{R}^{k \times k}$ be the stochastic transition matrix for the hidden state Markov chain: for all times t,

$$\Pr[y_{t+1} = i | y_t = j] = T_{i,j}, \quad i, j \in [k].$$

Finally, let $O \in \mathbb{R}^{d \times k}$ be the matrix whose j-th column is the conditional expectation of x_t given $y_t = j$: for all times t,

$$\mathbb{E}[x_t | y_t = j] = O e_j, \quad j \in [k].$$

Proposition 2 ([4]). *Define* $h := y_2$, *where* y_2 *is the second hidden state in the Markov chain. Then*

- x_1, x_2, x_3 *are conditionally independent given* h;
- *the distribution of* h *is given by the vector* $w := T\pi \in \Delta^{k-1}$;
- *for all* $j \in [k]$,

$$\begin{aligned}
\mathbb{E}[x_1 | h = j] &= O \operatorname{diag}(\pi) T^\top \operatorname{diag}(w)^{-1} e_j \\
\mathbb{E}[x_2 | h = j] &= O e_j \\
\mathbb{E}[x_3 | h = j] &= O T e_j.
\end{aligned}$$

Note the matrix of conditional means of x_t has full column rank, for each $t \in \{1, 2, 3\}$, provided that: (i) O has full column rank, (ii) T is invertible, and (iii) π and $T\pi$ have positive entries.

4 Orthogonal Tensor Decompositions

We now show how recovering the μ_i's in our aforementioned problems reduces to the problem of finding a certain orthogonal tensor decomposition of a symmetric tensor. We start by reviewing the spectral decomposition of symmetric matrices, and then discuss a generalization to the higher-order tensor case. Finally, we show how orthogonal tensor decompositions can be used for estimating the latent variable models from the previous section.

4.1 Review: The Matrix Case

We first build intuition by reviewing the matrix setting, where the desired decomposition is the eigendecomposition of a symmetric rank-k matrix $M = V\Lambda V^\top$, where $V = [v_1|v_2|\cdots|v_k] \in \mathbb{R}^{n\times k}$ is the matrix with orthonormal eigenvectors as columns, and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k) \in \mathbb{R}^{k\times k}$ is diagonal matrix of non-zero eigenvalues. In other words,

$$M = \sum_{i=1}^{k} \lambda_i \; v_i v_i^\top = \sum_{i=1}^{k} \lambda_i \; v_i^{\otimes 2}. \tag{1}$$

Such a decomposition is guaranteed to exist for every symmetric matrix.

Recovery of the v_i's and λ_i's can be viewed at least two ways. First, each v_i is fixed under the mapping $u \mapsto Mu$, up to a scaling factor λ_i:

$$Mv_i = \sum_{j=1}^{k} \lambda_j (v_j^\top v_i) v_j = \lambda_i v_i$$

as $v_j^\top v_i = 0$ for all $j \neq i$ by orthogonality. The v_i's are not necessarily the only such fixed points. For instance, with the multiplicity $\lambda_1 = \lambda_2 = \lambda$, then any linear combination of v_1 and v_2 is similarly fixed under M. However, in this case, the decomposition in (1) is not unique, as $\lambda_1 v_1 v_1^\top + \lambda_2 v_2 v_2^\top$ is equal to $\lambda(u_1 u_1^\top + u_2 u_2^\top)$ for any pair of orthonormal vectors, u_1 and u_2 spanning the same subspace as v_1 and v_2. Nevertheless, the decomposition is unique when $\lambda_1, \lambda_2, \ldots, \lambda_k$ are distinct, whereupon the v_j's are the only directions fixed under $u \mapsto Mu$ up to non-trivial scaling.

The second view of recovery is via the variational characterization of the eigenvalues. Assume $\lambda_1 > \lambda_2 > \cdots > \lambda_k$; the case of repeated eigenvalues again leads to similar non-uniqueness as discussed above. Then the *Rayleigh quotient*

$$u \mapsto \frac{u^\top M u}{u^\top u}$$

is maximized over non-zero vectors by v_1. Furthermore, for any $s \in [k]$, the maximizer of the Rayleigh quotient, subject to being orthogonal to $v_1, v_2, \ldots, v_{s-1}$, is v_s. Another way of obtaining this second statement is to consider the *deflated* Rayleigh quotient

$$u \mapsto \frac{u^\top \left(M - \sum_{j=1}^{s-1} \lambda_j v_j v_j^\top \right) u}{u^\top u}$$

and observe that v_s is the maximizer.

Efficient algorithms for finding these matrix decompositions are well studied [15, Section8.2.3], and iterative power methods are one effective class of algorithms.

We remark that in our multilinear tensor notation, we may write the maps $u \mapsto Mu$ and $u \mapsto u^\top Mu/\|u\|_2^2$ as

$$u \mapsto Mu \equiv u \mapsto M(I, u), \tag{2}$$

$$u \mapsto \frac{u^\top Mu}{u^\top u} \equiv u \mapsto \frac{M(u, u)}{u^\top u}. \tag{3}$$

4.2 The Tensor Case

Decomposing general tensors is a delicate issue; tensors may not even have unique decompositions. Fortunately, the orthogonal tensors that arise in the aforementioned models have a structure which permits a unique decomposition under a mild non-degeneracy condition. We focus our attention to the case $p = 3$, *i.e.*, a third order tensor; the ideas extend to general p with minor modifications.

An *orthogonal decomposition* of a symmetric tensor $T \in \bigotimes^3 \mathbb{R}^n$ is a collection of orthonormal (unit) vectors $\{v_1, v_2, \ldots, v_k\}$ together with corresponding positive scalars $\lambda_i > 0$ such that

$$T = \sum_{i=1}^{k} \lambda_i v_i^{\otimes 3}. \tag{4}$$

Note that since we are focusing on odd-order tensors ($p = 3$), we have added the requirement that the λ_i be positive. This convention can be followed without loss of generality since $-\lambda_i v_i^{\otimes p} = \lambda_i (-v_i)^{\otimes p}$ whenever p is odd. Also, it should be noted that orthogonal decompositions do not necessarily exist for every symmetric tensor.

In analogy to the matrix setting, we consider two ways to view this decomposition: a fixed-point characterization and a variational characterization. Related characterizations based on optimal rank-1 approximations can be found in [33].

Fixed-Point Characterization. For a tensor T, consider the vector-valued map

$$u \mapsto T(I, u, u) \tag{5}$$

which is the third-order generalization of (2). This can be explicitly written as

$$T(I,u,u) = \sum_{i=1}^{d} \sum_{1 \le j,l \le d} T_{i,j,l}(e_j^\top u)(e_l^\top u)e_i.$$

Observe that (5) is *not* a linear map, which is a key difference compared to the matrix case.

An eigenvector u for a matrix M satisfies $M(I,u) = \lambda u$, for some scalar λ. We say a unit vector $u \in \mathbb{R}^n$ is an *eigenvector* of T, with corresponding *eigenvalue* $\lambda \in \mathbb{R}$, if

$$T(I,u,u) = \lambda u.$$

(To simplify the discussion, we assume throughout that eigenvectors have unit norm; otherwise, for scaling reasons, we replace the above equation with $T(I,u,u) = \lambda \|u\| u$.) This concept was originally introduced in [23,30]. For orthogonally decomposable tensors $T = \sum_{i=1}^{k} \lambda_i v_i^{\otimes 3}$,

$$T(I,u,u) = \sum_{i=1}^{k} \lambda_i (u^\top v_i)^2 v_i \ .$$

By the orthogonality of the v_i, it is clear that $T(I,v_i,v_i) = \lambda_i v_i$ for all $i \in [k]$. Therefore each (v_i, λ_i) is an eigenvector/eigenvalue pair.

There are a number of subtle differences compared to the matrix case that arise as a result of the non-linearity of (5). First, even with the multiplicity $\lambda_1 = \lambda_2 = \lambda$, a linear combination $u := c_1 v_1 + c_2 v_2$ may *not* be an eigenvector. In particular,

$$T(I,u,u) = \lambda_1 c_1^2 v_1 + \lambda_2 c_2^2 v_2 = \lambda(c_1^2 v_1 + c_2^2 v_2)$$

may not be a multiple of $c_1 v_1 + c_2 v_2$. This indicates that the issue of repeated eigenvalues does not have the same status as in the matrix case. Second, even if all the eigenvalues are distinct, it turns out that the v_i's are not the only eigenvectors. For example, set $u := (1/\lambda_1)v_1 + (1/\lambda_2)v_2$. Then,

$$T(I,u,u) = \lambda_1 (1/\lambda_1)^2 v_1 + \lambda_2 (1/\lambda_2)^2 v_2 = u,$$

so $u/\|u\|$ is an eigenvector. More generally, for any subset $S \subseteq [k]$, we have that $\sum_{i \in S} (1/\lambda_i) v_i$ is (proportional to) an eigenvector.

As we now see, these additional eigenvectors can be viewed as spurious. We say a unit vector u is a *robust eigenvector* of T if there exists an $\epsilon > 0$ such that for all $\theta \in \{u' \in \mathbb{R}^n : \|u' - u\| \le \epsilon\}$, repeated iteration of the map

$$\bar{\theta} \mapsto \frac{T(I,\bar{\theta},\bar{\theta})}{\|T(I,\bar{\theta},\bar{\theta})\|} \ , \tag{6}$$

starting from θ converges to u. Note that the map (6) rescales the output to have unit Euclidean norm. Robust eigenvectors are also called attracting fixed points of (6) (see, *e.g.*, [20]).

The following theorem implies that if T has an orthogonal decomposition as given in (4), then the set of robust eigenvectors of T are precisely the set $\{v_1, v_2, \ldots v_k\}$, implying that the orthogonal decomposition is unique. (For even order tensors, the uniqueness is true up to sign-flips of the v_i.)

Theorem 6. *Let T have an orthogonal decomposition as given in (4).*

1. *The set of $\theta \in \mathbb{R}^n$ which do not converge to some v_i under repeated iteration of (6) has measure zero.*
2. *The set of robust eigenvectors of T is equal to $\{v_1, v_2, \ldots, v_k\}$.*

The proof of Theorem 6 is given in [1] and follows readily from simple orthogonality considerations. Note that *every* v_i in the orthogonal tensor decomposition is robust, whereas for a symmetric matrix M, for almost all initial points, the map $\bar{\theta} \mapsto \frac{M\bar{\theta}}{\|M\bar{\theta}\|}$ converges only to an eigenvector corresponding to the largest magnitude eigenvalue. Also, since the tensor order is odd, the signs of the robust eigenvectors are fixed, as each $-v_i$ is mapped to v_i under (6).

Variational Characterization. We now discuss a variational characterization of the orthogonal decomposition. The *generalized Rayleigh quotient* [33] for a third-order tensor is

$$u \mapsto \frac{T(u,u,u)}{(u^\top u)^{3/2}},$$

which can be compared to (3). For an orthogonally decomposable tensor, the following theorem shows that a non-zero vector $u \in \mathbb{R}^n$ is an *isolated local maximizer* [28] of the generalized Rayleigh quotient if and only if $u = v_i$ for some $i \in [k]$.

Theorem 7. *Assume $n \geq 2$. Let T have an orthogonal decomposition as given in (4), and consider the optimization problem*

$$\max_{u \in \mathbb{R}^n} T(u,u,u) \ \ s.t. \ \|u\| = 1.$$

1. *The stationary points are eigenvectors of T.*
2. *A stationary point u is an isolated local maximizer if and only if $u = v_i$ for some $i \in [k]$.*

The proof of Theorem 7 is given in [1]. It is similar to local optimality analysis for ICA methods using fourth-order cumulants (*e.g.*, [13,14]).

Again, we see similar distinctions to the matrix case. In the matrix case, the only local maximizers of the Rayleigh quotient are the eigenvectors with the largest eigenvalue (and these maximizers take on the globally optimal value). For the case of orthogonal tensor forms, the robust eigenvectors are precisely the isolated local maximizers.

An important implication of the two characterizations is that, for orthogonally decomposable tensors T, (i) the local maximizers of the objective function

$u \mapsto T(u,u,u)/(u^\top u)^{3/2}$ correspond precisely to the vectors v_i in the decomposition, and (ii) these local maximizers can be reliably identified using a simple fixed-point iteration (*i.e.*, the tensor analogue of the matrix power method). Moreover, a second-derivative test based on $T(I,I,u)$ can be employed to test for local optimality and rule out other stationary points.

4.3 Estimation via Orthogonal Tensor Decompositions

We now demonstrate how the moment tensors obtained for various latent variable models in Section 3 can be reduced to an orthogonal form. For concreteness, we take the specific form from the exchangeable single topic model (Theorem 1):

$$M_2 = \sum_{i=1}^{k} w_i \ \mu_i \otimes \mu_i,$$

$$M_3 = \sum_{i=1}^{k} w_i \ \mu_i \otimes \mu_i \otimes \mu_i.$$

(The more general case allows the weights w_i in M_2 to differ in M_3, but for simplicity we keep them the same in the following discussion.) We now show how to reduce these forms to an orthogonally decomposable tensor from which the w_i and μ_i can be recovered. See [1] for a discussion as to how previous approaches [2,4,16,27] achieved this decomposition through a certain simultaneous diagonalization method.

Throughout, we assume the following non-degeneracy condition.

Condition 41 (Non-degeneracy). *The vectors $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ are linearly independent, and the scalars $w_1, w_2, \ldots, w_k > 0$ are strictly positive.*

Observe that Condition 41 implies that $M_2 \succeq 0$ is positive semidefinite and has rank-k. This is a mild condition; furthermore, when this condition is not met, learning is conjectured to be hard for both computational [27] and information-theoretic reasons [26].

The Reduction. First, let $W \in \mathbb{R}^{d \times k}$ be a linear transformation such that

$$M_2(W,W) \ = \ W^\top M_2 W \ = \ I$$

where I is the $k \times k$ identity matrix (*i.e.*, W whitens M_2). Since $M_2 \succeq 0$, we may for concreteness take $W := UD^{-1/2}$, where $U \in \mathbb{R}^{d \times k}$ is the matrix of orthonormal eigenvectors of M_2, and $D \in \mathbb{R}^{k \times k}$ is the diagonal matrix of positive eigenvalues of M_2. Let

$$\tilde{\mu}_i := \sqrt{w_i}\ W^\top \mu_i.$$

Observe that

$$M_2(W,W) \ = \ \sum_{i=1}^{k} W^\top (\sqrt{w_i}\mu_i)(\sqrt{w_i}\mu_i)^\top W \ = \ \sum_{i=1}^{k} \tilde{\mu}_i \tilde{\mu}_i^\top \ = \ I,$$

so the $\tilde{\mu}_i \in \mathbb{R}^k$ are orthonormal vectors.

Now define $\widetilde{M}_3 := M_3(W, W, W) \in \mathbb{R}^{k \times k \times k}$, so that

$$\widetilde{M}_3 \;=\; \sum_{i=1}^{k} w_i \, (W^\top \mu_i)^{\otimes 3} \;=\; \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \, \tilde{\mu}_i^{\otimes 3}.$$

As the following theorem shows, the orthogonal decomposition of $\widetilde{M}_3$ can be obtained by identifying its robust eigenvectors, upon which the original parameters w_i and μ_i can be recovered. For simplicity, we only state the result in terms of robust eigenvector/eigenvalue pairs; one may also easily state everything in variational form using Theorem 7.

Theorem 8. *Assume Condition 41 and take $\widetilde{M}_3$ as defined above.*

1. *The set of robust eigenvectors of $\widetilde{M}_3$ is equal to $\{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_k\}$.*
2. *The eigenvalue corresponding to the robust eigenvector $\tilde{\mu}_i$ of $\widetilde{M}_3$ is equal to $1/\sqrt{w_i}$, for all $i \in [k]$.*
3. *If $B \in \mathbb{R}^{d \times k}$ is the Moore-Penrose pseudoinverse of $W^\top$, and (v, λ) is a robust eigenvector/eigenvalue pair of $\widetilde{M}_3$, then $\lambda B v = \mu_i$ for some $i \in [k]$.*

The theorem follows by combining the above discussion with the robust eigenvector characterization of Theorem 6. Recall that we have taken as convention that eigenvectors have unit norm, so the μ_i are exactly determined from the robust eigenvector/eigenvalue pairs of $\widetilde{M}_3$ (together with the pseudoinverse of $W^\top$); in particular, the scale of each μ_i is correctly identified (along with the corresponding w_i). Relative to previous works on moment-based estimators for latent variable models (*e.g.*, [2,4,16]), Theorem 8 emphasizes the role of the special tensor structure, which in turn makes transparent the applicability of methods for orthogonal tensor decomposition.

5 Tensor Power Method

In this section, we consider the tensor power method of [21, Remark 3] for orthogonal tensor decomposition. We first state a simple convergence analysis for an orthogonally decomposable tensor T.

When only an approximation $\hat{T}$ to an orthogonally decomposable tensor T is available (*e.g.*, when empirical moments are used to estimate population moments), an orthogonal decomposition need not exist for this perturbed tensor (unlike for the case of matrices), and a more robust approach is required to extract the approximate decomposition. Here, we propose such a variant in Algorithm 1 and provide a detailed perturbation analysis. We note that alternative approaches such as simultaneous diagonalization can also be employed (see [1]).

5.1 Convergence Analysis for Orthogonally Decomposable Tensors

The following lemma establishes the quadratic convergence of the tensor power method (*i.e.*, repeated iteration of (6)) for extracting a single component of

the orthogonal decomposition. Note that the initial vector θ_0 determines which robust eigenvector will be the convergent point. Computation of subsequent eigenvectors can be computed with deflation, *i.e.*, by subtracting appropriate terms from T.

Lemma 1. *Let* $T \in \bigotimes^3 \mathbb{R}^n$ *have an orthogonal decomposition as given in* (4)*. For a vector* $\theta_0 \in \mathbb{R}^n$*, suppose that the set of numbers* $|\lambda_1 v_1^\top \theta_0|, |\lambda_2 v_2^\top \theta_0|, \ldots, |\lambda_k v_k^\top \theta_0|$ *has a unique largest element. Without loss of generality, say* $|\lambda_1 v_1^\top \theta_0|$ *is this largest value and* $|\lambda_2 v_2^\top \theta_0|$ *is the second largest value. For* $t = 1, 2, \ldots,$ *let*

$$\theta_t := \frac{T(I, \theta_{t-1}, \theta_{t-1})}{\|T(I, \theta_{t-1}, \theta_{t-1})\|}.$$

Then

$$\|v_1 - \theta_t\|^2 \leq \left(2\lambda_1^2 \sum_{i=2}^{k} \lambda_i^{-2}\right) \cdot \left|\frac{\lambda_2 v_2^\top \theta_0}{\lambda_1 v_1^\top \theta_0}\right|^{2^{t+1}}.$$

That is, repeated iteration of (6) *starting from* θ_0 *converges to* v_1 *at a quadratic rate.*

To obtain all eigenvectors, we may simply proceed iteratively using deflation, executing the power method on $T - \sum_j \lambda_j v_j^{\otimes 3}$ after having obtained robust eigenvector / eigenvalue pairs $\{(v_j, \lambda_j)\}$.

Proof. Let $\overline{\theta}_0, \overline{\theta}_1, \overline{\theta}_2, \ldots$ be the sequence given by $\overline{\theta}_0 := \theta_0$ and $\overline{\theta}_t := T(I, \theta_{t-1}, \theta_{t-1})$ for $t \geq 1$. Let $c_i := v_i^\top \theta_0$ for all $i \in [k]$. It is easy to check that (i) $\theta_t = \overline{\theta}_t / \|\overline{\theta}_t\|$, and (ii) $\overline{\theta}_t = \sum_{i=1}^k \lambda_i^{2^t - 1} c_i^{2^t} v_i$. (Indeed, $\overline{\theta}_{t+1} = \sum_{i=1}^k \lambda_i (v_i^\top \overline{\theta}_t)^2 v_i = \sum_{i=1}^k \lambda_i (\lambda_i^{2^t - 1} c_i^{2^t})^2 v_i = \sum_{i=1}^k \lambda_i^{2^{t+1} - 1} c_i^{2^{t+1}} v_i$.) Then

$$1 - (v_1^\top \theta_t)^2 = 1 - \frac{\lambda_1^{2^{t+1}-2} c_1^{2^{t+1}}}{\sum_{i=1}^k \lambda_i^{2^{t+1}-2} c_i^{2^{t+1}}} \leq \frac{\sum_{i=2}^k \lambda_i^{2^{t+1}-2} c_i^{2^{t+1}}}{\sum_{i=1}^k \lambda_i^{2^{t+1}-2} c_i^{2^{t+1}}} \leq \lambda_1^2 \sum_{i=2}^{k} \lambda_i^{-2} \cdot \left|\frac{\lambda_2 c_2}{\lambda_1 c_1}\right|^{2^{t+1}}.$$

Since $\lambda_1 > 0$, we have $v_1^\top \theta_t > 0$ and hence $\|v_1 - \theta_t\|^2 = 2(1 - v_1^\top \theta_t) \leq 2(1 - (v_1^\top \theta_t)^2)$ as required.

5.2 Perturbation Analysis of a Robust Tensor Power Method

Now we summarize the case where we have an approximation $\hat{T}$ to an orthogonally decomposable tensor T. Here, a more robust approach is required to extract an approximate decomposition. We propose such an algorithm in Algorithm 1, and provide a detailed perturbation analysis. For simplicity, we assume the tensor $\hat{T}$ is of size $k \times k \times k$ as per the reduction from Section 4.3. In some applications, it may be preferable to work directly with a $n \times n \times n$ tensor of rank $k \leq n$ (as in Lemma 1); our results apply in that setting with little modification.

Algorithm 1. Robust tensor power method

input symmetric tensor $\tilde{T} \in \mathbb{R}^{k \times k \times k}$, number of iterations L, N.
output the estimated eigenvector/eigenvalue pair; the deflated tensor.

for $\tau = 1$ to L **do**
  Draw $\theta_0^{(\tau)}$ uniformly at random from the unit sphere in $\mathbb{R}^k$.
  for $t = 1$ to N **do**
    Compute power iteration update

$$\theta_t^{(\tau)} := \frac{\tilde{T}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})}{\|\tilde{T}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})\|} \tag{7}$$

  end for
end for
Let $\tau^* := \arg\max_{\tau \in [L]} \{\tilde{T}(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)})\}$.
Do N power iteration updates (7) starting from $\theta_N^{(\tau^*)}$ to obtain $\hat{\theta}$, and set $\hat{\lambda} := \tilde{T}(\hat{\theta}, \hat{\theta}, \hat{\theta})$.
return the estimated eigenvector/eigenvalue pair $(\hat{\theta}, \hat{\lambda})$; the deflated tensor $\tilde{T} - \hat{\lambda}\, \hat{\theta}^{\otimes 3}$.

Assume that the symmetric tensor $T \in \mathbb{R}^{k \times k \times k}$ is orthogonally decomposable, and that $\hat{T} = T + E$, where the perturbation $E \in \mathbb{R}^{k \times k \times k}$ is a symmetric tensor with small operator norm:

$$\|E\| := \sup_{\|\theta\|=1} |E(\theta, \theta, \theta)|.$$

In our latent variable model applications, $\hat{T}$ is the tensor formed by using empirical moments, while T is the orthogonally decomposable tensor derived from the population moments for the given model. In the context of parameter estimation (as in Section 4.3), E must account for any error amplification throughout the reduction, such as in the whitening step.

[1] provides a perturbation analysis which is similar to Wedin's perturbation theorem for singular vectors of matrices [32] in that it bounds the error of the (approximate) decomposition returned by Algorithm 1 on input $\hat{T}$ in terms of the size of the perturbation, provided that the perturbation is small enough.

Acknowledgments. We thank Boaz Barak, Dean Foster, Jon Kelner, and Greg Valiant for helpful discussions. This work was completed while DH was a postdoctoral researcher at Microsoft Research New England, and partly while AA, RG, and MT were visiting the same lab. AA is supported in part by the NSF Award CCF-1219234, AFOSR Award FA9550-10-1-0310 and the ARO Award W911NF-12-1-0404.

References

1. Anandkumar, A., Ge, R., Hsu, D., Kakade, S., Telgarsky, M.: Tensor decompositions for learning latent variable models. Journal of Machine Learning Research **15**, (2014)

2. Anandkumar, A., Foster, D.P., Hsu, D., Kakade, S.M., Liu, Y.-K.: A spectral algorithm for latent Dirichlet allocation. In: Advances in Neural Information Processing Systems 25, (2012)
3. Anandkumar, A., Hsu, D., Huang, F., Kakade, S.M.: Learning mixtures of tree graphical models. In: Advances in Neural Information Processing Systems 25 (2012)
4. 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)
5. Austin, T.: On exchangeable random variables and the statistics of large graphs and hypergraphs. Probab. Survey **5**, 80–145 (2008)
6. Cardoso, J.-F.: Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors. In: ICASSP-91, 1991 International Conference on Acoustics, Speech, and Signal Processing, pp. 3109–3112. IEEE (1991)
7. 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)
8. Cattell, R.B.: Parallel proportional profiles and other principles for determining the choice of factors by rotation. Psychometrika **9**(4), 267–283 (1944)
9. Chang, J.T.: Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. Mathematical Biosciences **137**, 51–73 (1996)
10. Comon, P.: Independent component analysis, a new concept? Signal Processing **36**(3), 287–314 (1994)
11. Comon, P., Golub, G., Lim, L.-H., Mourrain, B.: Symmetric tensors and symmetric tensor rank. SIAM Journal on Matrix Analysis Appl. **30**(3), 1254–1279 (2008)
12. Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press, Elsevier (2010)
13. Delfosse, N., Loubaton, P.: Adaptive blind separation of independent sources: a deflation approach. Signal Processing **45**(1), 59–83 (1995)
14. Alan, M., Frieze, M.J., Kannan, R.: Learning linear transformations. In: Thirty-Seventh Annual Symposium on Foundations of Computer Science, pp. 359–368 (1996)
15. Golub, G.H., van Loan, C.F.: Matrix Computations. Johns Hopkins University Press (1996)
16. Hsu, D., Kakade, S.M.: Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In: Fourth Innovations in Theoretical Computer Science (2013)
17. Hsu, D., Kakade, S.M., Liang, P.: Identifiability and unmixing of latent parse trees. In: Advances in Neural Information Processing Systems 25 (2012)
18. Hsu, D., Kakade, S.M., Zhang, T.: A spectral algorithm for learning hidden Markov models. Journal of Computer and System Sciences **78**(5), 1460–1480 (2012)
19. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks **13**(4–5), 411–430 (2000)
20. Kolda, T.G., Mayo, J.R.: Shifted power method for computing tensor eigenpairs. SIAM Journal on Matrix Analysis and Applications **32**(4), 1095–1124 (2011)
21. De Lathauwer, L., De Moor, B., Vandewalle, J.: On the best rank-1 and rank-$(R_1, R_2, ., R_n)$ approximation and applications of higher-order tensors. SIAM J. Matrix Anal. Appl. **21**(4), 1324–1342 (2000)
22. Le Cam, L.: Asymptotic Methods in Statistical Decision Theory. Springer (1986)

23. Lim, L.-H.: Singular values and eigenvalues of tensors: a variational approach. In: Proceedings of the IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing vol. 1, pp. 129–132 (2005)
24. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press (1967)
25. McCullagh, P.: Tensor Methods in Statistics. Chapman and Hall (1987)
26. 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)
27. Mossel, E., Roch, S.: Learning nonsingular phylogenies and hidden Markov models. Annals of Applied Probability **16**(2), 583–614 (2006)
28. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, 1999
29. Pearson, K.: Contributions to the mathematical theory of evolution. In: Philosophical Transactions of the Royal Society, London, A., p. 71 (1894)
30. Qi, L.: Eigenvalues of a real supersymmetric tensor. Journal of Symbolic Computation **40**(6), 1302–1324 (2005)
31. Stegeman, A., Comon, P.: Subtracting a best rank-1 approximation may increase tensor rank. Linear Algebra and Its Applications **433**, 1276–1300 (2010)
32. Wedin, P.: Perturbation bounds in connection with singular value decomposition. BIT Numerical Mathematics **12**(1), 99–111 (1972)
33. Zhang, T., Golub, G.: Rank-one approximation to high order tensors. SIAM Journal on Matrix Analysis and Applications **23**, 534–550 (2001)