# The Application of Symmetric Tensor Decomposition to Parameter Estimation of Latent Variable Models

Longde Wang
Advisor: Professor Jiawang Nie

**Abstract**

For certain latent variable models, such as single topic models and Gaussian mixture models, their low-order moments yield symmetric tensor structures. Thus, parameter estimation of these models can be converted to symmetric tensor decomposition. In this paper, we analyze the application of two symmetric tensor decomposition methods to parameter estimation of certain latent variable models. Our discussion consists of three main parts. We identify the symmetric tensor structures of the second and third moments of certain latent variable models, examine the method of orthogonalization proposed by Anandkumar et al. and implemented through the tensor power method, and illustrate the method of generating polynomials proposed by Nie and related to the apolarity lemma. We conclude with numerical experiments of the two methods.

## 1 Introduction

For parameter estimation of statistical models, the method of moments is usually less computationally expensive than the maximum likelihood, especially in the case of high-dimensional data, such as the latent variable models. However, the data in the latent variable models cannot be directly observed, so the efficiency of the method of moments is not guaranteed. Nevertheless, in certain types of latent variable models, such as the exchangeable single topic model and the spherical Gaussian mixtures, their low-order moments yield symmetric tensor structures, which allow for a symmetric tensor decomposition approach to estimate the parameters of the models.

Symmetric tensor decomposition can be regarded as a generalization of the singular value decomposition of symmetric matrices, and the tensor power method, analogous to the power method of symmetric matrix decomposition can be useful. The tensor power method, intended for decomposing a symmetric tensor by finding its eigenvectors, was first proposed by Lathauwer et al. [12] and further analyzed by Kolfidis and Regalia [18], but the problem is that the convergence of this method is not guaranteed. However, in this paper, we demonstrate that the convergence of this method on the orthogonally-decomposable (odeco) symmetric tensors. A variant of this method, called the shifted tensor power method, proposed by Kolda and Mayo [16], is also guaranteed to converge, but without the orthogonality condition, the obtained eigenvectors are not useful for symmetric tensor decomposition. We apply the tensor power method to the parameter estimation of latent variable models in this paper.

Another method is the method of generating polynomials proposed by Nie [5]. The linear relations of recursive patterns of symmetric tensors induces the concept of generating polynomial, which is a representation of the linear relation. Nie proposed that obtaining the symmetric tensor decomposition of the symmetric tensor depends on a set of generating polynomials represented by a generating matrix. This method is motivated by the apolarity lemma proposed by Iarrobino and Kanev [9]. We apply the method of generating polynomials to the parameter estimation of latent variable models in this paper.

In Section 2, we introduce some preliminary knowledge for this paper. In Section 3, we identify the symmetric tensor structures of low-order moments of certain latent variable models. In Section 4, we examine the method of orthogonalization and its application to parameter estimation of latent variable models. In Section 5, we analyze the method of generating polynomials and its application to parameter estimation of latent variable models. In Section 6, we present the results of some numerical experiments.

# 2 Preliminaries

In this section, we introduce some preliminary knowledge for this paper.

## 2.1 Latent Variable Models

In this subsection, we illustrate what a latent variable model is.

**Definition 1.** A latent variable model is a statistical model that contains both observable variables and latent variables, which cannot be observed but can be inferred from observable variables, and explains the observable variables based on the latent variables.

**Example 1.** One simple example of latent variable model is the congeneric test model, where the observed variable is usually a psychological test score. This model can be represented by a simple linear regression

$$Y_i = \lambda_i \xi + \epsilon_i,$$

where $Y_i$ is the observable variable, $\xi$ is a latent variable, $\lambda_i$ is the measure of association between the observable and the latent variables, and $\epsilon_i$ is assumed to be normal: $\epsilon_i \sim N(0, \sigma_i^2)$. Hence, the conditional distribution of $Y_i$ given $\xi$ is normal with mean $\lambda_i \xi$ and variance $\sigma_i^2$. Since $\xi$ cannot be observed, its prior distribution is often set to be standard normal.

In this paper, we consider the exchangeable single topic model, the spherical Gaussian mixture model, the latent Dirichlet allocation (LDA) model, the independent component analysis (ICA) model, and the multi-view model. In general, the parameters to estimate in these models are the weights of some components and the expected values conditional on the components. In the past, the popular parameter estimation method is the Expectation-Maximization (EM) algorithm, but the convergence rate is low and the local optima is not satisfactory. With the identification of the low-order cross moment of these models as symmetric tensors, we can use symmetric tensor decomposition methods to estimate the parameters. Thus, we need the concept of tensors.

## 2.2 Tensors

In this subsection, we examine the concept and properties of tensors. The concept of tensor arises from a high-dimensional generalization of matrix as a linear map. Let us first look at the formal definition of tensor, borrowed from Nie [1].

**Definition 2.** ([1]) Let $V_1, \cdots, V_m$ be some vector spaces. For a tuple $(v_1, \cdots, v_m) \in V_1 \times \cdots \times V_m$, their tensor product $v_1 \otimes \cdots \otimes v_m$ is defined as the multilinear functional acted on the Cartesian product of the dual spaces of $V_1, \cdots, V_m$, $V_1^* \times \cdots \times V_m^*$, such that

$$v_1 \otimes \cdots \otimes v_m(v_1^*, \cdots, v_m^*) = v_1^*(v_1) \cdots v_m^*(v_m),$$

$\forall\ (v_1^*, \cdots, v_m^*) \in V_1^* \times \cdots \times V_m^*$. $v_1 \otimes \cdots \otimes v_m$ is also called a rank-1 tensor. $V_1 \otimes \cdots \otimes V_m$, the span of all such rank-1 tensors, is called the tensor product of $V_1, \cdots, V_m$ and is also called a tensor product space.

For a general tensor $T \in V_1 \otimes \cdots \otimes V_m$, it can been decomposed as a linear combination of the rank-1 tensors, that is

$$T = \sum_{i=1}^{r} \lambda_i v_1^{(i)} \otimes \cdots \otimes v_m^{(i)}.$$

For a tensor $T = \sum_{k=1}^{r} v_1^{(k)} \otimes \cdots \otimes v_m^{(k)}$ and a basis$\{e_1^{(j)}, \cdots, e_{n_j}^{(j)}\}$ of the vector space $V_j$, since $v_j^{(k)}$ is a linear combination of $e_1^{(j)}, \cdots, e_{n_j}^{(j)}$,

$$T = \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} T_{i_1 \cdots i_m} e_{i_1}^{(1)} \otimes \cdots \otimes e_{i_m}^{(m)}.$$

Hence, $T$ is determined and represented by the high-dimensional array

$$(T_{i_1 \cdots i_m})_{i_1, \cdots, i_m = 1}^{n_1, \cdots n_m},$$

which is also called a hypermatrix. Sometimes a tensor is defined as a high-dimensional array.

In this paper, we use the concept of symmetric tensors, so we demonstrate the definition of symmetric tensors.

**Definition 3.** ([1]) Let $T = \sum_{i=1}^{r} v_1^{(i)} \otimes \cdots \otimes v_m^{(i)}$ be a tensor in the tensor product space $V \otimes \cdots \otimes V = V^{\otimes m}$ and $(v_{\pi(1)}, \cdots, v_{\pi(m)})$ be a permutation of $(v_1, \cdots, v_m)$. We say that $T$ is a symmetric tensor if for every permutation $(v_{\pi(1)}, \cdots, v_{\pi(m)})$ of $(v_1, \cdots, v_m)$,

$$\sum_{i=1}^{r} v_{\pi(1)}^{(i)} \otimes \cdots \otimes v_{\pi(m)}^{(i)} = \sum_{i=1}^{r} v_1^{(i)} \otimes \cdots \otimes v_m^{(i)} = T.$$

Observe that $v \otimes \cdots \otimes v = v^{\otimes m}$ is a symmetric tensor, and the symmetrization of a rank-1 tensor $v_1 \otimes \cdots \otimes v_m$ is

$$sym(v_1 \otimes \cdots \otimes v_m) = \sum_{\pi:\pi \ is \ a \ permutation \ of \ (1, \cdots, m)} v_{\pi(1)}^{(i)} \otimes \cdots \otimes v_{\pi(m)}^{(i)}.$$

Next, we introduce the tensor-vector product and matrix-tensor product. For a tensor $T \in \mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_m}$, and $u_{s+1} \in \mathbb{R}^{n_{s+1}}, \cdots, u_m \in \mathbb{R}^{n_m}$, we define the tensor-vector product as

$$T \times (u_{s+1}, \cdots, u_m) = (\sum_{i_{s+1}=1}^{n_{s+1}} \cdots \sum_{i_m=1}^{n_m} T_{i_1\cdots i_m}[u_{s+1}]_{i_{s+1}} \cdots [u_m]_{i_m})_{i_1,\cdots,i_s=1}^{n_1,\cdots,n_s}, \ s \in [m].$$

Note that the tensor-vector product results in a tensor of order $s$ for $s \in [m]$, and if $s = 0$, then the product is a scalar

$$\sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} T_{i_1\cdots i_m}[u_1]_{i_1} \cdots [u_m]_{i_m}$$

If $T$ is an $m$-th order tensor in $\mathbb{R}^n\otimes\cdots\otimes\mathbb{R}^n$ and $u \in \mathbb{R}^n$, we denote $T\times(u, \cdots, u)$ as $Tu^s$, where there are length of the tuple $(u, \cdots, u)$ is $s$. If we expand the $m$-th order tensor $T$ of dimension $(n_1, \cdots, n_m)$ linearly, i.e., write $T = \sum_{i=1}^r v_1^{(i)} \otimes \cdots \otimes v_j^{(i)} \otimes \cdots \otimes v_m^{(i)}$, and have a matrix $X_j \in \mathbb{R}^{p_j\times n_j}$, we define the matrix-tensor product

$$X_j \times T = \sum_{i=1}^{r} v_1^{(i)} \otimes \cdots \otimes (X_j v_j^{(i)}) \otimes \cdots \otimes v_m^{(i)}.$$

If we have matrices $X_1 \in \mathbb{R}^{p_1\times n_1}, \cdots, X_m \in \mathbb{R}^{p_m\times n_m}$, then we define the multilinear matrix multiplication

$$(X_1, \cdots, X_m) \times T = \sum_{i=1}^{r}(X_1 v_1^{(i)}) \otimes \cdots \otimes (X_m v_m^{(i)}).$$

Then, we explain what the rank of a tensor and symmetric rank of a symmetric tensor are. Since $\mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_m} \cong \mathbb{R}^{n_1\times\cdots\times n_m}$, we can see a tensor which is the sum of some rank-1 tensors $\sum_{i=1}^r v_1^{(i)} \otimes \cdots \otimes v_m^{(i)} = T$ in the tensor product space $\mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_m}$ as a tensor which is the sum of Segre products of some vectors $\sum_{i=1}^r v_1^{(i)} \otimes \cdots \otimes v_m^{(i)} = T$ in the space $\mathbb{R}^{n_1\times\cdots\times n_m}$, where $v_j^{(i)} \in \mathbb{R}^{n_j}$ and $j \in [m]$.

**Definition 4.** ([1]) For a tensor $T \in \mathbb{R}^{n_1\times\cdots\times n_m}$, its rank is defined as the smallest integer $r$ such that $T$ is the sum of $r$ rank-1 tensors, i.e.,

$$rank(T) = \min_r\{T = \sum_{i=1}^{r} v_1^{(i)} \otimes \cdots \otimes v_m^{(i)}\}.$$

$T = \sum_{i=1}^r v_1^{(i)} \otimes \cdots \otimes v_m^{(i)}$ is known as a CP rank decomposition.

Similarly, we can define the symmetric rank of a symmetric tensor. Let $S^m(\mathbb{R}^n)$ be a subspace of $\mathbb{R}^n \otimes \cdots \otimes \mathbb{R}^n$ such that it is the space of all $m$-th order symmetric tensors of dimension $n$.

**Definition 5.** ([1]) For a symmetric tensor $T \in S^3(\mathbb{R}^n)$, its symmetric rank is defined as the smallest number $r$ such that it is a linear combination of $r$ symmetric rank-1 tensors, i.e.,

$$srank(T) = \min_r\{T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes m}\}.$$

$T = \sum_{i=1}^r \lambda_i v_i^{\otimes m}$ is known as a symmetric decomposition.

Observe that for a symmetric tensor $T$, $rank(T) \leq srank(T)$.

**Example 2.** ([1]) Consider the tensor $T \in S^3(\mathbb{R}^3)$ determined by the high-dimensional array $(i_1 + i_2 + i_3)_{1 \leq i_1, i_2, i_3 \leq 3}$ which takes the representation

$$\left[\begin{array}{ccc|ccc|ccc} 3 & 4 & 5 & 4 & 5 & 6 & 5 & 6 & 7 \\ 4 & 5 & 6 & 5 & 6 & 7 & 6 & 7 & 8 \\ 5 & 6 & 7 & 6 & 7 & 8 & 7 & 8 & 9 \end{array}\right].$$

One can verify that $rank(T) = 3$, and the CP rank decomposition is given by

$$T = \begin{bmatrix}1\\1\\1\end{bmatrix} \otimes \begin{bmatrix}1\\1\\1\end{bmatrix} \otimes \begin{bmatrix}1\\2\\3\end{bmatrix} + \begin{bmatrix}1\\1\\1\end{bmatrix} \otimes \begin{bmatrix}1\\2\\3\end{bmatrix} \otimes \begin{bmatrix}1\\1\\1\end{bmatrix} + \begin{bmatrix}1\\1\\1\end{bmatrix} \otimes \begin{bmatrix}1\\1\\1\end{bmatrix} \otimes \begin{bmatrix}1\\2\\3\end{bmatrix}.$$

Also, $srank(T) = 3$, and the symmetric decomposition is given by

$$T = \frac{1}{2}\begin{bmatrix}2\\3\\4\end{bmatrix}^{\otimes 3} + \frac{1}{2}\begin{bmatrix}0\\1\\2\end{bmatrix}^{\otimes 3} - \begin{bmatrix}1\\2\\3\end{bmatrix}^{\otimes 3}.$$

For this tensor $T$, the rank and the symmetric rank are equal.

# 3 Latent Variable Models

In this section, we examine the symmetric tensor structure of the second-order and third-order cross moments of certain types of latent variable models, observed by Anandkumar et al. [2]. Specifically, the structure of the cross moments yield a symmetric tensor decomposition, in which every rank-one tensor is written in terms of the model parameters. In some models, since the cross moments do not yield a symmetric tensor structure directly, the moments are required to be modified to obtain the desired structure.

## 3.1 Exchangeable Single Topic Model

In this subsection, we introduce the exchangeable single topic model, a special case of the topic model. A topic model is a type of statistical model for identifying the topics, seen as latent variables, that occur in a collection of documents. The simplest topic model, the exchangeable single topic model assumes the words in the documents, seen as random variables $x_1, \cdots, x_l$, to be exchangeable, that is,

$$f_{x_1 \cdots x_l}(x_1, \cdots, x_l) = f_{x_{\pi(1)} \cdots x_{\pi(l)}}\left(x_{\pi(1)}, \cdots, x_{\pi(l)}\right)$$

where $x_{\pi(1)}, \cdots, x_{\pi(l)}$ is a permutation of the observed variables. According to De Finetti's theorem [23], in the single topic model, the random variables (words) $x_1, \cdots, x_l$ are conditionally independent and identically distributed given a latent variable (topic) $h$, and the conditional distribution is invariant among all the nodes.

The formation of the single topic model is as follows. First, a topic $h$ is drawn from the discrete distribution $w\colon =(w_1, \cdots, w_r)$, so $P(h = j) = w_j$. Then, given $h = j$, $l$ words are drawn independently from the discrete distribution that is determined by the probability vector $\mu_j$ where $j \in [r]$. The $l$ words, represented by $x_1, \cdots, x_l \in \mathbb{R}^n$, are from the word pool consisting of $n$ words, each of which is represented by an integer $i \in [n]$. Hence, $x_1, \cdots, x_l$ satisfies that $x_t = e_i$ when the $t$-th word is $i$, where $e_1, \cdots, e_n$ is the canonical basis for $\mathbb{R}^n$.

According to the generative process, the single topic model has two properties. First, the $(i_1, \cdots .i_l)$-th entry of the tensor of moment $E[x_1 \otimes \cdots \otimes x_l]$ is simply the joint probability $P(x_1 = e_{i_1}, \cdots, x_l = e_{i_l}) = P(1st\ word = i_1, \cdots, l-th\ word = i_l)$. Second, the conditional expectation $E[x_t|h=j]$ is simply topic $j$'s probability vector $\mu_j$. Since the random variables are conditionally independent given the latent variable, the conditional moment $E[x_1 \otimes x_2 | h = j] = \mu_j \otimes \mu_j$. Thus, the following theorem follows.

**Theorem 1.** *([20]) If*

$$M := E\left[x_1 \otimes x_2\right],$$
$$T := E\left[x_1 \otimes x_2 \otimes x_3\right],$$

*then*

$$M = \sum_{i=1}^{r} w_i \mu_i^{\otimes 2}$$
$$T = \sum_{i=1}^{r} w_i \mu_i^{\otimes 3}$$

*Proof.* Note that $x_t = e_p$ if and only if the $t$-th word in the document is $p$, and the $p$-th coordinate of the probability vector $\mu_j$ is the probability that the $t$-th word in the document is $p$ given topic $h = j$. Since $x_t = \sum_{p=1}^{n} [x_t]_p e_p$,

$$E[x_1 \otimes x_2] = E[\sum_{p=1}^{n}\sum_{q=1}^{n}[x_1]_p[x_2]_q e_p \otimes e_q]$$

$$= \sum_{p=1}^{n}\sum_{q=1}^{n} E[[x_1]_p[x_2]_q] e_p \otimes e_q.$$

Note that, due to the conditional i.i.d property of $x_t$,

$$E[[x_1]_p[x_2]_q] = \sum_{i=1}^{r} P(h=i) E[[x_1]_p[x_2]_q]|h=i]$$

$$= \sum_{i=1}^{r} w_i E[[x_1]_p]|h=i] E[[x_2]_q|h=i]$$

$$= \sum_{i=1}^{r} w_i [\sum_{t=1}^{n}[\mu_i]_t e_t]_p [\sum_{t=1}^{n}[\mu_i]_t e_t]_q$$

$$= \sum_{i=1}^{r} w_i [\mu_i]_p [\mu_i]_q$$

Hence, it follows that

$$E[x_1 \otimes x_2] = \sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{i=1}^{r} w_i [\mu_i]_p [\mu_i]_q e_p \otimes e_q = \sum_{i=1}^{r} w_i \mu_i^{\otimes 2}$$

Similarly, for the three-word case,

$$E[x_1 \otimes x_2 \otimes x_3] = \sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{s=1}^{n}\sum_{i=1}^{r} w_i[\mu_i]_p[\mu_i]_q[\mu_i]_s e_p \otimes e_q \otimes e_s = \sum_{i=1}^{r} w_i \mu_i^{\otimes 3}$$

□

## 3.2 Nonsymmetric Moments

In this subsection, we introduce the models in which the moments do not have a symmetric tensor structure. In order to obtain the symmetric tensor structure, various manipulations need to be applied to the cross moments.

### 3.2.1 Spherical Gaussian Mixtures

In this subsection, spherical Gaussian mixtures is presented. Here, the probability given in a mixture of $k$ Gaussians is

$$P(x) = \sum_{i=1}^{r} w_i N(x|\mu_i, \Sigma_i),$$

where $\sum_{i=1}^{r} w_i = 1$ and the covariance matrix is spherical, that is $\Sigma_i = \sigma_i^2 I$. Similarly, $w_i$'s determines the distribution of the latent variable $h$. In spherical Gaussian mixtures, an observation is written as

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

where $z$ is conditionally independent given $h$ and $z|h \sim N(0, {\sigma_i}^2 I)$. Unlike the single topic model, in which the draws of the random variables are under the same condition $h$, the Gaussian mixture model present random variable draws related to different conditions $h$.

In this model, the second-order and third-order moments do not directly yield a symmetric tensor form, so some modifications are needed. Define the modfied expectation $V$ to be $E\left[x\left(v^\top(x - E[x])\right)^2\right]$, where $v$ is a normalized eigenvector corresponding to the smallest eigenvalue $\bar{\sigma}^2$ of the covariance matrix, which is the weighted average variance. This modified expectation $V$ and the weighted average variance $\bar{\sigma}^2$ are used to modify the moments, as shown in the following theorem.

**Theorem 2.** *([3]) Suppose $n \geq r$. The smallest eigenvalue of the coviriance matrix $E[x^{\otimes 2}] - E[x] \otimes E[x]$ is the weighted average variance $\overline{\sigma}^2 = \sum_{i=1}^{r} w_i \sigma_i^2$. Suppose $v$ is a normalized eigenvector corresponding to $\overline{\sigma}^2$. If*

$$V := E[x(v^\top(x - E[x]))^2],$$

$$M := E\left[x^{\otimes 2}\right] - \overline{\sigma}^2 I,$$

$$T := E\left[x^{\otimes 3}\right] - \sum_{i=1}^{n}(V \otimes e_i \otimes e_i + e_i \otimes V \otimes e_i + e_i \otimes e_i \otimes V),$$

*then*

$$M = \sum_{i=1}^{r} w_i \mu_i^{\otimes 2}$$

$$T = \sum_{i=1}^{r} w_i \mu_i^{\otimes 3}.$$

*Proof.* Suppose $\bar{\mu} = E[x] = E[\mu_h] = \sum_{i=1}^r w_i \mu_i$. Then the covariance matrix is

$$E[(x-\bar{\mu})^{\otimes 2}] = E[(\mu_h + z - \bar{\mu})^{\otimes 2}]$$

$$= \sum_{i=1}^r w_i((\mu_i - \bar{\mu})^{\otimes 2} + \sigma_i^2 I) = \sum_{i=1}^r w_i(\mu_i - \bar{\mu})^{\otimes 2} + \bar{\sigma}^2 I.$$

Since $\sum_{i=1}^r (\mu_i - \bar{\mu}) = 0$, the $(\mu_i - \bar{\mu})$'s are linearly dependent, and hence $\sum_{i=1}^r w_i(\mu_i - \bar{\mu})^{\otimes 2}$ has rank $\widetilde{r} < r$. Thus, the $d - \widetilde{r}$ smallest eigenvalues are $\bar{\sigma}^2$, and the other eigenvalues are strictly larger. Due to the strict separation of eigenvalues, the eigenvector $v$ corresponding to $\bar{\sigma}^2$ satisfies $v^\top(\mu_i - \bar{\mu}) = 0$. Then

$$V = E[x(v^\top(x - E[x]))^2] = E[(\mu_h + z)(v^\top(\mu_h - \bar{\mu} + z))^2] = E[(\mu_h + z)(v^\top z)^2] = E[\mu_h \sigma_h^2],$$

since $z|h \sim N(0, \sigma_h I)$. Moreover, since $E[z \otimes z] = \sum_{i=1}^r w_i \sigma_i^2 I = \bar{\sigma}^2 I$,

$$M = E\left[x^{\otimes 2}\right] - \bar{\sigma}^2 I = E[\mu_h^{\otimes 2}] + E[z^{\otimes 2}] - \bar{\sigma}^2 = E[\mu_h^{\otimes 2}] = \sum_{i=1}^r w_i \mu_i^{\otimes 2}.$$

Furthermore, since $E[\mu_h \otimes z \otimes z] = E[E[\mu_h \otimes z \otimes z|h]] = E[\sum_{i=1}^n \sum_{j=1}^n [z]_i [z]_j \mu_h \otimes e_i \otimes e_j|h]] = E[\sum_{i=1}^n \sigma_h^2 \mu_h \otimes e_i \otimes e_i] = \sum_{i=1}^n V \otimes e_i \otimes e_i$, and similarly $\sum_{i=1}^n e_i \otimes V \otimes e_i$ and $\sum_{i=1}^n e_i \otimes e_i \otimes V$ can be derived,

$$T = E\left[x^{\otimes 3}\right] - E[\mu_h \otimes z \otimes z] - E[z \otimes \mu_h \otimes z] - E[z \otimes z \otimes \mu_h] = E[\mu_h^{\otimes 3}] = \sum_{i=1}^r w_i \mu_i^{\otimes 3}.$$

□

### 3.2.2 Latent Dirichlet Allocation

In this subsection, latent Dirichlet allocation (LDA) is illustrated. LDA belongs to the family of the mixed membership models, which is defined as follows.

**Definition 6.** ([2]) A latent variable model is a mixed membership model if it satisfies the following conditions:

- Data are grouped.
- Each group is modeled with a mixture model (e.g., Gaussian mixture model).
- The mixture components are shared across groups.
- The mixture proportions vary across groups.

LDA allows different sets of observations to be explained by groups of latent variables. In LDA, a document has a mixture of topics, with distribution of topic mixtures as Dirichlet distribution $Dir(\alpha)$ with parameter $\alpha \in \mathbb{R}_{++}^r$. The probability density function of $Dir(\alpha)$ is

$$P_\alpha(h) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^r \Gamma(\alpha_i)} \prod_{i=1}^r h_i^{\alpha_i - 1}$$

where $\alpha_0 := \sum_{i=1}^r \alpha_i$.

The formation of the LDA model is as follows. The topics are, as before, determined by the probability vectors $\mu_1, \cdots, \mu_r$. First, a topic mixture $h = (h_1, \cdots, h_r)$ is drawn from $Dir(\alpha)$. Next, given $h$, $l$ words $x_1, \cdots, x_l$ are drawn independently from the discrete distribution that is specified by the probability vector $\sum_{i=1}^r h_i\mu_i$. Just like the single topic model, the $l$ words, represented by $x_1, \cdots, x_l$, satisfies that $x_t = e_i$ when the $t$-th word is $i$.

The parameter $\alpha_0$ is an indicator of how concentrated the distribution of $h$ is. As $\alpha_0$ goes to zero, the model becomes a single topic model, and as $\alpha_0$ goes to infinity, $h$ has a uniform distribution with parameter $\frac{1}{r}$. The first-order moment $V$ and the parameters $\alpha$ and $\alpha_0$ are used to symmetrize the moments, as shown in the following theorem.

**Theorem 3.** *([11]) If*

$$V: = E[x_1],$$

$$M: = E\left[x_1 \otimes x_2\right] - \frac{\alpha_0}{\alpha_0+1}V^{\otimes 2}$$

$$T: = E\left[x_1 \otimes x \otimes x_3\right] - \frac{\alpha_0}{\alpha_0+2}(E[V \otimes x_1 \otimes x_2] + E[x_1 \otimes V \otimes x_2] + E[x_1 \otimes x_2 \otimes V])$$

$$+\frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}V^{\otimes 3},$$

*then*

$$M = \sum_{i=1}^r \frac{\alpha_i}{(\alpha_0+1)\alpha_0}\mu_i^{\otimes 2},$$

$$T = \sum_{i=1}^r \frac{2\alpha_i}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\mu_i^{\otimes 3}.$$

*Proof.* Note that $V = E[x_1] = E[E[x_1|h]] = E[\sum_{i=1}^r h_i\mu_i] = \sum_{i=1}^r \frac{\alpha_i}{\alpha_0}\mu_i$. Then we have

$$E[x_1 \otimes x_2] = E[E[x_1 \otimes x_2|h]] = E[E[x_1|h] \otimes E[x_2|h]] = E[\sum_{i-1}^r \sum_{j=1}^r h_i h_j \mu_i \otimes \mu_j]$$

$$= \sum_{i=1}^r \sum_{j\neq i} \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)}\mu_i \otimes \mu_j + \sum_{i=1}^r \frac{\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)}\mu_i^{\otimes 2}.$$

Then

$$M = E[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0+1}V^{\otimes 2}$$

$$= E[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0+1}\sum_{i=1}^r \sum_{j=1}^r \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)}\mu_i \otimes \mu_j$$

$$= \sum_{i=1}^r \frac{\alpha_i}{(\alpha_0+1)\alpha_0}\mu_i^{\otimes 2}$$

Similarly,

$$E[x_1 \otimes x_2 \otimes x_3] = \sum_{i=1}^r \sum_{j\neq i} \sum_{s\neq j,i} \frac{\alpha_i\alpha_j\alpha_s}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\mu_i \otimes \mu_j \otimes \mu_s + \sum_{i=1}^r \sum_{s\neq i} \frac{\alpha_i(\alpha_i+1)\alpha_s}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\mu_i \otimes \mu_i \otimes \mu_s$$

$$+\sum_{i=1}^{r}\sum_{s\neq i}\frac{\alpha_i(\alpha_i+1)\alpha_s}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\mu_i\otimes\mu_s\otimes\mu_i+\sum_{i=1}^{r}\sum_{s\neq i}\frac{\alpha_i(\alpha_i+1)\alpha_s}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\mu_s\otimes\mu_i\otimes\mu_i$$

$$+\sum_{i=1}^{r}\frac{\alpha_i(\alpha_i+1)(\alpha_i+2)}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\mu_i^{\otimes 3}.$$

Thus, after similar manipulations, we have

$$T = E\left[x_1\otimes x\otimes x_3\right]-\frac{\alpha_0}{\alpha_0+2}(E[V\otimes x_1\otimes x_2]+E[x_1\otimes V\otimes x_2]+E[x_1\otimes x_2\otimes V])$$

$$+\frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}V^{\otimes 3},$$

$$=\sum_{i=1}^{r}\frac{2\alpha_i}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\mu_i^{\otimes 3}.$$

□

### 3.2.3 Independent Component Analysis

In this subsection, independent component analysis (ICA) is demonstrated. ICA is a computational method for separating a multivariate signal into independent subcomponents with Gaussian noises [22]. In this model, the observed random vector $x \in \mathbb{R}^n$ satisfies that

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

where $h \in \mathbb{R}^r$ is a latent random vector with independent entries, $A \in \mathbb{R}^{n\times r}$ is the matrix for mixing the signals, $z \in \mathbb{R}^n$ is the Gaussian noise, and $h$ and $z$ are independent. Let $\mu_i$ be the $i$-th column of $A$. In ICA, the modified fourth-order moment $K$ yield a symmetric tensor structure.

**Theorem 4.** *([21]) Let*

$$K = E[x^{\otimes 4}] - W,$$

*where $W \in S^4(\mathbb{R}^n)$ is such that*

$$W_{i_1i_2i_3i_4} = E[x_{i_1}x_{i_2}]E[x_{i_3}x_{i_4}] + E[x_{i_1}x_{i_3}]E[x_{i_2}x_{i_4}] + E[x_{i_1}x_{i_4}]E[x_{i_2}x_{i_3}],$$

*for $1 \le i_1, i_2, i_3, i_4 \le n$. Then*

$$K = \sum_{i=1}^{r}\kappa_i\mu_i^{\otimes 4},$$

*where $\kappa_i = E[h_i^4] - 3$.*

We can obtain the second-order tensor $M$ and the third-order tensor $T$ by taking $M = K\times(u,v) = \sum_{i=1}^{r}\kappa_i(\mu_i^\top u)(\mu_i^\top v)\mu_i^{\otimes 2}$ and $T = Kv = \sum_{i=1}^{r}\kappa_i(\mu_i^\top v)\mu_i^{\otimes 3}$, for $u, v \in \mathbb{R}^n$.

### 3.3 Multi-View Models

In this subsection, multi-view models, sometimes also called naive Bayes models, are brought into investigation. These models are similar to the single topic model in the sense that they assume the conditional independence of the observed variables given a latent variable. However, multi-view models do not require the same conditional distributions of the variables. The latent variable $h$ still satisfies that $P(h = j) = w_j$. Unlike the single topic model, the multi-view model specifies that the observations $x_1, \cdots, x_l$ are each determined by the conditional mean $\mu_{t,j} = E[x_t|h = j]$. Due to this setting, the cross moments have the following form:

$$E[x_s \otimes x_t] = \sum_{i=1}^{r} w_i \mu_{s,i} \otimes \mu_{t,i}, s, t \in [3], s \neq t,$$

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

Note that the cross moments do not yield a symmetric tensor form under the condition of different conditional distributions of the observations. To symmetrize the moments, a linear transformation that correlates $x_1$ and $x_2$ to $x_3$ is required, as specified by the following theorem.

**Theorem 5.** *([11]) Assume $\mu_{v,1}, \cdots, \mu_{v,r}$ are linearly independent for each $v \in [3]$. If*

$$\widetilde{x_1} :\ = E[x_3 \otimes x_2] E[x_1 \otimes x_2]^{-1} x_1,\ \widetilde{x_2} :\ = E[x_3 \otimes x_1] E[x_2 \otimes x_1]^{-1} x_2,$$

$$M :\ = E\left[\widetilde{x_1} \otimes \widetilde{x_2}\right],$$

$$T :\ = E\left[\widetilde{x_1} \otimes \widetilde{x_2} \otimes x_3\right],$$

*then*

$$M = \sum_{i=1}^{r} w_i \mu_{3,i}^{\otimes 2},$$

$$T = \sum_{i=1}^{r} w_i \mu_{3,i}^{\otimes 3}.$$

# 4 Method of Orthogonalization

In this section, we illustrate the orthogonal tensor decomposition, which is the symmetric tensor decomposition for orthogonally-decomposable (odeco) symmetric tensors, using the tensor power method, a higher-order generalization of the matrix power method for finding eigenvalues and eigenvectors. We introduce this decomposition obtained by tensor power method, relate it to parameter estimation of the latent variable models, and analyze the ensured convergence of the tensor power method.

Since we would like to obtain the symmetric tensor decomposition for the third-order cross moment, our discussion here is restricted to the third-order tensors. Let $S^3(\mathbb{R}^n)$ be the space of symmetric cubic tensors of dimension $n$ and $T \in S^3(\mathbb{R}^n)$ be a symmetric cubic tensor. Then its orthogonal decomposition is given by

$$T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes 3}, \tag{1}$$

where $\lambda_i \geq 0$ and $\{v_1, \cdots, v_r\}$ are orthonormal. Note that such an orthogonal decomposition is not guaranteed to exist for every symmetric tensor. In order to recover the $v_i$'s and the $\lambda_i$'s, we can aim at either first finding the eigenvectors and then identifying the eigenvalues, or vice versa. The advantage of orthogonal tensors is that a unique decomposition can be obtained given that the coefficients $\lambda_i > 0 \ \forall\ i \in [r]$. The uniqueness condition justifies the use of orthogonal tensor decomposition to estimate parameters of latent variable models.

## 4.1 Tensor Eigenvalues and Eigenvectors

In this subsection, we generalize the concept of matrix eigenvalue / eigenvector to tensor eigenvalue / eigenvector. Recall that for a matrix $X \in \mathbb{R}^{n\times n}$, if

$$Mu = \lambda u,$$

for some scalar $\lambda$ and vector $u \in \mathbb{R}^n$, then we say $\lambda$ is an eigenvalue of $X$ and $u$ is an eigenvector of $M$. Next, we examine what the eigenvalue / eigenvector of a symmetric tensor is, and relate the concepts to orthogonal tensor decomposition.

### 4.1.1 Identification of the Tensor Eigenvectors

**Definition 7.** ([4]) For $T \in S^3(\mathbb{R}^n)$ and $u \in \mathbb{R}^n$, if

$$Tu^2 = \lambda u,$$

then we say $\lambda$ is an Z-eigenvalue of $T$ and $u$ is the corresponding Z-eigenvector. For simplicity, we just call them eigenvalue / eigenvector.

Since $T$ has the structure pecified by (1), we define the map

$$f(u) = Tu^2 = (\sum_{j=1}^{n}\sum_{k=1}^{n} T_{ijl}[u]_j[u]_k)_{i=1}^n = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} T_{ijl}(e_j^\top u)(e_k^\top u)e_i. \tag{2}$$

Note that this map is not linear. To get rid of scaling issues, we assume that the eigenvectors of $T$ are unit vectors. By orthogonality, $v_i^\top v_j = 0 \ \forall\ i \neq j$, so $Tv_i^2 = \sum_{j=1}^{k} \lambda_j (v_i^\top v_j)^2 v_j = \lambda_i v_i$. Hence, each $\lambda_i$ in (1) is an eigenvalue of $T$ and each $v_i$ is an eigenvector.

As opposed to the matrix case, even if all the $\lambda_i$'s are distinct, $v_i$'s are not the only eigenvectors. For $i \neq j$, $\frac{v_i}{\lambda_i} + \frac{v_j}{\lambda_j}$ is an eigenvector, since $T(\frac{v_i}{\lambda_i} + \frac{v_j}{\lambda_j})^2 = \lambda_i(\frac{1}{\lambda_i})^2 v_i + \lambda_j(\frac{1}{\lambda_j})^2 v_j = \frac{v_i}{\lambda_i} + \frac{v_j}{\lambda_j}$. As a result, the notion of robust eigenvectors arises because of the need to specify a complete set of eigenvectors. A unit vector $v$ is said to be a robust eigenvector if $\exists\ \epsilon > 0$ such that $\forall\ \theta \in \{u \in \mathbb{R}^n | \ \|u - v\| \leq \epsilon\}$, repeated iterations of the map

$$\phi(\theta) = \frac{T\theta^2}{\|T\theta^2\|}$$

starting from $\theta$ converges to $v$. The following theorem indicates the uniqueness of the orthogonal decomposition of tensors by robust eigenvectors.

**Theorem 6.** *([2]) Let $T$ be orthogonally decomposable such that $T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes 3}$. Then the set of $\theta \in \mathbb{R}^n$ which, under repeated iterations of the map $\phi(\theta)$, do not converge to some $v_i$, has measure zero, and the robust eigenvectors are $v_i$'s for $i \in [r]$.*

For a matrix $M$, for nearly all initial points, the map $\phi(\theta) = \frac{M\theta}{\|M\theta\|}$ converges to the eigenvector $v_1$ corresponding to the eigenvalue $\lambda_1$ such that $\|\lambda_1\| = \max_{i\in[r]} \|\lambda_i\|$. On the other hand, each $v_i$ in the orthogonal decomposition of a tensor is a robust eigenvector. Moreover, the tensor order is odd, so $-v_i$ is mapped to $v_i$ under $\phi(\theta)$ and hence the signs of the robust eigenvectors are fixed.

### 4.1.2 Identification of the Tensor Eigenvalues

If, for $T \in S^3(\mathbb{R}^n)$ and $u \in \mathbb{R}^n$, $Tu^2 = \lambda u$, then $\lambda = Tu^3$, since

$$Tu^3 = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} T_{ijk}[u]_i[u]_j[u]_k = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} T_{ijk}(e_i^\top u)(e_j^\top u)(e_k^\top u).$$

Due to orthogonality, $Tv_i^3 = \sum_{j=1}^{r} \lambda_j (v_i^\top v_j)^3 = \lambda_i$, so each $\lambda_i$ in (1) is an eigenvalue of $T$ and each $v_i$ is an eigenvector, and we arrive at the same observation as from the perspective of eigenvectors. In order to recover the $\lambda_i$'s and $v_i$'s, we can solve the optimization problem $\max_{u\in\mathbb{R}^n} Tu^3$ *s.t.* $\|u\| \leq 1$, as expained by the following theorem, since the following theorem indicates that for an orthogonally decomposable tensor, $v_i$'s are the only isolated maximizers of $Tu^3$.

**Theorem 7.** *([2]) Let $T$ be orthogonally decomposable such that $T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes 3}$. Then for the optimization problem*

$$\max_{u\in\mathbb{R}^n} Tu^3 \ \ s.t. \ \|u\| \leq 1,$$

*eigenvectors of $T$ are the constrained stationary points, and a constrained stationary point $u^*$ is an isolated local maximizer if and only if $u^* = v_i$ for some $i \in [r]$.*

In the matrix case, the local maximizers are the eigenvectors corresponding to the largest eigenvalues, while in the tensor case, the robust eigenvectors are the isolated local maximizers.

From the last two theorems, we see that the tensor $T = \sum_{i=1}^{r} \lambda_i v_i^{\otimes 3}$ has a unique orthogonal decomposition, since the orthonormal decomposition vectors are uniquely determined by its robust eigenvectors $v_i$, and the corresponding eigenvalues $\lambda_i$ are uniquely determined by $Tv_i^3$.

## 4.2 Some Useful Results

In this subsection, we briefly talk about some results about odeco tensors proposed by Robeva [15], which can be useful. Note that a tensor $T$ can be converted to a polynomial in the polynomial ring $\mathbb{R}[x_1, \cdots, x_n]$, i.e.,

$$T(x_1, \cdots, x_n) = \sum_{0\leq i_1,i_2,i_3\leq n} T_{i_1 i_2 i_3} x_{i_1} x_{i_2} x_{i_3}.$$

For $x \in \mathbb{C}^n$, the eigenvector definition $Tx^2 = \lambda x$ is equivalent to $\nabla T(x) = 3\lambda x$, and hence $\nabla T(x)$ and $x$ are parallel. Then the variety of eigenvectors of $T$, denoted by $\nu_T$, is the $2 \times 2$ minors of the matrix $[\nabla T(x)|x] \in \mathbb{R}^{n\times 2}$. Since the hypersurface $T(x) = 0$ has no singular points, the eigenvectors of $T(x)$ are the stationary points of the map $\nabla T(x)$ from the projective space $\mathbb{PC}^{n-1}$ that maps $[x]$ to $[\nabla T(x)]$, where $[x]$ is the projection of $x \in \mathbb{C}^n$.

**Definition 8.** ([15]) The odeco variety is the Zariski closure of the set of all odeco tensors $T \in S^3(\mathbb{R}^n)$ such that $T = \sum_{i=1}^r \lambda_i v_i^{\otimes 3}$.

For an odeco tensor $T$, its polynomial form $T(x)$ is in the $r$-th secant variety of the 3rd Veronese variety $\sigma_r(v_3(\mathbb{C}^n))$. According to Robeva [16], the dimension of the odeco variety, which is irreducible, in $S^3(\mathbb{C}^n)$ is $\binom{n+1}{2}$. Moreover, by Cartwright and Sturmfels [19], the number of equivalence classes of a tensor $T \in S^3(\mathbb{R}^n)$ with finitely many equivalence classes of eigenpairs is $2^n - 1$, with multiplicity, and by Robeva [16], that of an odeco tensor $T \in S^3(\mathbb{R}^n)$ such that $T = \sum_{i=1}^r \lambda_i v_i^{\otimes 3}$ is $2^r - 1$.

## 4.3 Application to Parameter Estimation

In this subsection, we relate parameter estimation of the latent variable model to the orthogonal tensor decomposition introduced in the last subsection by applying the method to the third-order cross moment $T$. We take the exchangeable single topic model, in which the cross moments have the form $M = \sum_{i=1}^r w_i\mu_i^{\otimes 2}$ and $T = \sum_{i=1}^r w_i\mu_i^{\otimes 3}$, with $\mu_i \in \mathbb{R}^n$, for instance. In the single topic model,the $w_i$'s in $M$ and $T$ assumed to be same, while in general, this may not be the case (LDA), and mild modifications are required in addition to the process introduced in this section. Throughout this subsection, the nondegeneracy condition, which is that the $\mu_i$'s are linearly independent and the $w_i$'s are strictly positive, is assumed to hold. Note that this condition implies that $M$ is positive semidefinite and has rank $r$.

We assume that the $\mu_i$'s are linearly independent, but they are not necessarily orthogonal, and hence the orthogonal decomposition cannot be applied. Thus, orthogonalization of the third-order tensor $T$ using the second-order tensor $M$, as shown in the following process. Let $X \in \mathbb{R}^{n\times r}$ be a matrix such that $X^\top MX = I$. Since $M \succeq 0$ ude to the non-degeneracy condition, we can perform an eigendecomposition to X. Take $X = PD^{-\frac{1}{2}}$, where $P$ is the matrix with the orthornormal eigenvectors of $M$ as columns, and $D$ is the diagonal matrix of positive eigenvalues of $M$, because $X^\top MX = (D^{-\frac{1}{2}})^\top P^\top PDP^\top P(D^{-\frac{1}{2}})^\top = I$. Let $\widetilde{\mu}_i = \sqrt{w_i}X^\top\mu_i$. Note that $\widetilde{\mu}_i \in \mathbb{R}^r$. Observe that

$$X^\top MX = \sum_{i=1}^r X^\top(\sqrt{w_i}\mu_i)(\sqrt{w_i}\mu_i)^\top X = \sum_{i=1}^r \widetilde{\mu}_i\widetilde{\mu}_i^\top = I,$$

so the $\widetilde{\mu}_i$'s are orthornormal. Let $\widetilde{T} = (X^\top, X^\top, X^\top) \times T$. Observe that

$$\widetilde{T} = \sum_{i=1}^r w_i(X^\top\mu_i)^{\otimes 3} = \sum_{i=1}^r w_i(\frac{\widetilde{\mu}_i}{\sqrt{w_i}})^{\otimes 3} = \sum_{i=1}^r \frac{1}{\sqrt{w_i}}\widetilde{\mu}_i^{\otimes 3} \tag{3}$$

is orthogonally decomposable. After these steps, the third-order cross moment $T$ is converted to an orthogonally decomposable cubic tensor $\widetilde{T}$. The following theorem states that the orthogonal decomposition of $\widetilde{T}$ can be obtained using its robust eigenvectors, and then $w_i$ and $\mu_i$ can be identified.

**Theorem 8.** *([2]) Suppose that the $\mu_i$'s are linearly independent and the $w_i > 0$, and $\widetilde{T} = (X^\top, X^\top, X^\top) \times T = \sum_{i=1}^k \frac{1}{\sqrt{w_i}} \widetilde{\mu}_i^{\otimes 3}$. Then, the robust eigenvectors of $\widetilde{T}$ are $\{\widetilde{\mu}_1, \cdots, \widetilde{\mu}_r\}$, and the corresponding eigenvalues are $\frac{1}{\sqrt{w_i}}$'s, Furthermore, let $(X^\top)^\dagger$ denote the Moore-Penrose inverse of $X^\top$, and $(\lambda, v)$ be a robust eigenvalue and eigenvector pair of $\widetilde{T}$. Then $\lambda (X^\top)^\dagger v = \mu_i$ for some $i \in [r]$.*

## 4.4 Convergence Analysis of Tensor Power Method

In this subsection, we ensure the convergence of the tensor power method for orthogonally decomposable tensors. Notice that if the orthogonally decomposable tensor $\widetilde{T}$ is approximated by $\widehat{T}$, an orthogonal decomposition for $\widehat{T}$ may not exist. For the latent variable model, this is exactly the case, since the cross moments are empirical ones estimated by the method of moments. Thus, the tensor power method for obtaining the approximate decomposition is needed. For an orthogonally decomposable cubic tensor, as characterized in (1), the tensor power method is the repeated iteration of the map

$$\phi(\theta) = \frac{T\theta^2}{\|T\theta^2\|}. \tag{4}$$

The following theorem claims this map converges at a quadratic rate. The initial point determines the convergent point, which is a robust eigenvector of $T$.

**Lemma 1.** ([2]) Let $T \in S^3(\mathbb{R}^n)$ be orthogonally decomposable, that is, $T = \sum_{i=1}^r \lambda_i v_i^{\otimes 3}$. Assume that the largest value of $|\lambda_1 v_1^\top \theta_0|, \cdots, |\lambda_1 v_k^\top \theta_0|$ is unique for $\theta_0 \in \mathbb{R}^n$. Let $|\lambda_1 v_k^\top \theta_0|$ be the largest value and $\lambda_2 v_k^\top \theta_0$ be the second largest value. For $t = 1, 2, \cdots$, let

$$\theta_t = \frac{T\theta_{t-1}^2}{\|T\theta_{t-1}^2\|}.$$

Then, the repeated iteration starting from $\theta_0$ converges to $v_1$ at a quadratic rate. Precisely,

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

*Proof.* Define $\widetilde{\theta}_0 = \theta_0$ and $\widetilde{\theta}_t = T\theta_{t-1}^2$. Let $c_i = v_i^\top \theta_0$. Note that $\theta_t = \frac{\widetilde{\theta}_t}{\|\widetilde{\theta}_t\|}$. Also, observe that $\widetilde{\theta}_t = \sum_{i=1}^r \lambda_i^{2^t-1} c_i^{2^t} v_i$ through iterative computation

$$\widetilde{\theta}_t = \sum_{i=1}^r \lambda_i (v_i^\top \theta_{t-1})^2 v_i = \sum_{i=1}^r \lambda_i (v_i^\top \sum_{j=1}^k \lambda_j (v_j^\top \theta_{t-2})^2 v_j)^2 v_i$$

$$= \sum_{i=1}^k \lambda_i (\lambda_i (v_i^\top \theta_{t-2})^2)^2 v_i = \sum_{i=1}^r \lambda_i (\lambda_i (v_i^\top \sum_{j=1}^r \lambda_j (v_j^\top \theta_{t-3})^2 v_j)^2)^2 v_i$$

$$= \sum_{i=1}^r \lambda_i (\lambda_i (\lambda_i (v_i^\top \theta_{t-3})^2)^2)^2 v_i = \cdots = \sum_{i=1}^r \lambda_i^{2^t-1} c_i^{2^t} v_i.$$

Then

$$1 - (v_1^\top \theta_t)^2 = 1 - \frac{(v_1^\top \widetilde{\theta}_t)^2}{\|\widetilde{\theta}_t\|^2} = 1 - \frac{\lambda_1^{2^{t+1}-2} c_1^{2^{t+1}}}{\sum_{i=1}^r \lambda_i^{2^{t+1}-2} c_i^{2^{t+1}}} \leq \frac{\sum_{i=2}^r \lambda_i^{2^{t+1}-2} c_i^{2^{t+1}}}{\sum_{i=1}^r \lambda_i^{2^{t+1}-2} c_i^{2^{t+1}}} \leq \lambda_1^2 \sum_{i=2}^r \frac{1}{\lambda_i^2} |\frac{\lambda_2 c_2}{\lambda_1 c_1}|^{2^{t+1}}$$

Since $\lambda_1 > 0$, $v_1^\top \theta_t > 0$. Then

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

□

Applying the tensor power method to the deflated tensor $T - \sum_j \lambda_j v_j^{\otimes 3}$ after getting the first few robust eigenvalue / eigenvector pairs, all robust eigenvectors can be obtained. Note that the convergence is ensured only if the symmetric tensor is orthorgonally decomposable (odeco). For a general symmetric tensor, the power iteration does not converge as shown by Lathauwer et al. [12], since the eigenvectors are no longer decomposition vectors when the tensor is not odeco.

## 4.5 Algorithm for Parameter Estimation

In this subsection, we propose an algorithm for parameter estimation of the latent variable models using the techniques and methods introduced in Section 4. Suppose that the empirical second-order cross moment $\widehat{M}$ and the empirical third-order cross moment $\widehat{T}$ are available. Also, for simplicity, assume the length of decomposition $r$ is given. Algorithm 1 is provided to give a the estimated parameters.

**Algorithm 1**

**Input** symmetric tensor $\widehat{M} \in \mathbb{R}^{n \times n}$, symmetric tensor $\widehat{T} \in \mathbb{R}^{n \times n \times n}$, length of decomposition $r$
**Output** the estimated eigenvalue / eigenvector pair, the deflated tensor

Obtain the eigendecomposition for $\widehat{M}$, that is $M = PDP^\top$, and let $X = PD^{-\frac{1}{2}}$
Let $\widetilde{T} = (X^\top, X^\top, X^\top) \times \widehat{T}$
**for** $b = 1$ to $B$ **do**
    draw $\theta_0^{(b)}$ uniformly at random from the unit sphere in $\mathbb{R}^n$
    **for** $t = 1$ to $T$ **do**
        Compute $\theta_t^{(b)} = \frac{\widetilde{T}\theta_{t-1}^{(b)2}}{\|\widetilde{T}\theta_{t-1}^{(b)2}\|}$
    **end for**
**end for**
Let $b^* = \arg\max_{b \in [B]} \{\widetilde{T}\theta_T^{(b)3}\}$ and $\widehat{\theta}_0 = \theta_T^{(b^*)}$
**for** $t = 1$ to $T$ **do**
    Compute $\widehat{\theta}_t = \frac{T\widehat{\theta}_{t-1}^2}{\|T\widehat{\theta}_{t-1}^2\|}$
**end for**
Let $\widehat{\theta} = \widehat{\theta}_T$ and set $\widehat{\lambda} = \widehat{T}\widehat{\theta}^3$
Let $\widetilde{\lambda} = (\frac{1}{\widehat{\lambda}})^2$ and $\widetilde{\theta} = \widehat{\lambda}(X^\top)^+\widehat{\theta}$
**return** $(\widetilde{\lambda}, \widetilde{\theta})$, $\widehat{T} - \widehat{\lambda}\widehat{\theta}^{\otimes 3}$

# 5 Method of Generating Polynomials

In Section 4, we employ the technique of orthogonalization of the cross moment tensor and orthogonal tensor decomposition. The requirement of orthogonality is intended for ensuring the uniqueness of the decomposition. Otherwise, the tensor to decompose will have multiple decompositions, and the problem of choosing the right one will arise. Also, the orthornomral robust eigenvectors obtained by the tensor power method can be used as decomposition vectors in the orthogonal decomposition. Without this setting, we can still find the eigenvectors of a symmetric tensor using some variant of the power method that is guaranteed to converge, proposed by Kolda and Mayo [16], but the eigenvectors cannot be used as the vectors in the symmetric decomposition of the tensor. Furthermore, this setting implies that the tensor rank is smaller than $n$, where the tensor $T \in S^3(\mathbb{R}^n)$, and the existence of tractable decomposition methods.

However, if the requirement of orthogonality is omitted, under some condition, we can still get the unique decomposition of the cross moment tensor, seen as a general symmetric tensor. In this section, we explore another symmetric tensor decomposition method, proposed by Nie [5], that uses generating polynomials, which describe the linear relation of recursive patterns of the symmetric tensor entries. Due to the construction of the method, we let a complex third-order tensor $T$ be in $S^3(\mathbb{C}^{n+1})$. For $T$, this method gives the decomposition

$$T = \sum_{i=1}^{r} u_i^{\otimes 3}. \tag{5}$$

Notice that the dimension used in this method is $n + 1$ instead of $n$. Since we want to apply the method to cross moment tensors of latent variable models, in which the conditional expectation vectors are rarely dependent and the cross moment tensors have low-rank structures, it is reasonable to assume that the decomposition vectors are linearly independent and $r < n+1$. Our discussion in this section is also restricted to third-order tensors.

## 5.1 Uniqueness of Symmetric Tensor Decomposition

In this subsection, we introduce the uniqueness condition of the symmetric tensor decomposition. Recent research about symmetric tensor decompositions show that the uniqueness of the decomposition of a symmetric tensor can be satisfied under some conditions. In particular, the setting of linear independence of the decomposition vectors implies that the symmetric tensor decomposition is unique.

Before stating the result, we need the concept of generic rank of tensors.

**Definition 9.** ([1]) Let $S^3(\mathbb{C}^{n+1})$ be the space of symmetric cubic tensors of dimension $n+1$ and $S^3_r(\mathbb{C}^{n+1})$ be the set of symmetric third-order tensors with dimension $n+1$ and symmetric rank $r$. The generic rank of $S^3(\mathbb{C}^{n+1})$ is defined as the smallest $r$ such that the closure of $S^3_r(\mathbb{C}^{n+1})$ is the entire symmetric tensor space $S^3(\mathbb{C}^{n+1})$, i.e.,

$$\overline{S^3_r(\mathbb{C}^{n+1})} = S^3(\mathbb{C}^{n+1}).$$

The result about uniqueness of symmetric tensor decomposition is related to this concept.

The result is given by Chiantini, Ottaviani and Vannieuwenhoven [6]. According to Alexander and Hirschowitz [7], for a generic third order symmetric tensor of dimension

$n+1$ over the complex field $\mathbb{C}$, the symmetric ranks is given by the formula

$$SR(3, n+1) = \lceil \frac{1}{n+1} \binom{n+3}{3} \rceil, \tag{6}$$

except for $n = 4$, and for this case, the rank should be increased by one. The uniqueness condition is related to this formula. A generic third-order symmetric tensor of dimension $n+1$ over the complex field has a unique symmetric decomposition if the symmetric tensor has a subgeneric rank $r$, i.e.,

$$r < \lceil \frac{1}{n+1} \binom{n+3}{3} \rceil,$$

except for the case that $n = 5$ and $r = 9$, in which there are two symmetric decompositions. We can use this result to check if the third-order moment tensor has a unique decomposition. Oeding, Ottaviani and Vandewalle [13] also proposed conditions that imply the uniqueness of decomposition.

For parameter estimation of latent variable models, it is reasonable to assume that the nondegeneracy condition, which is that the $\mu_i$'s are linearly independent and the $w_i$'s are strictly positive, holds. By this condition, the cross moments tensors can be written as

$$T = \sum_{i=1}^{r} (\sqrt[3]{w_i}\mu_i)^{\otimes 3} = \sum_{i=1}^{r} u_i^{\otimes 3},$$

where the decomposition vectors $u_i$ are linearly independent. Note that if the $u_i$'s are linearly independent, then the tensor $T$ has a unique symmetric tensor decomposition.

## 5.2 Some Useful Results

In this subsection, we introduce some useful results that are related to determining the symmetric rank of a tensor. Let $S^3(\mathbb{C}^{n+1})$ be the space of symmetric cubic tensors of dimension $n+1$ and $\mathbb{P}S^3(\mathbb{C}^{n+1})$ be the projective space. Note that $S^3(\mathbb{C}^{n+1})$ has dimension $\binom{n+3}{3}$ and $\mathbb{P}S^3(\mathbb{C}^{n+1})$ has dimension $\binom{n+3}{3} - 1$ [8].

**Definition 10.** ([5]) We define the variety of symmetric tensors $\sigma_r$ to be the Zariski closure [14] of equivalent classes of symmetric tensors with decomposition $\sum_{i=1}^{r} u_i^{\otimes 3}$ in $\mathbb{P}S^3(\mathbb{C}^{n+1})$.

Note that $\sigma_r$ is also defined as the $r$-th secant variety of the 3rd Veronese variety of $n+1$ variables, just like in the case of odeco tensors. In order to get a decomposition as in (6), determining the length of decomposition $r$ is important. The best case is that $r$ equals to the symmetric rank of the tensor, in which the decomposition is called a symmetric rank decomposition. For a general third order symmetric tensor of dimension $n+1$, its symmetric rank is given by $SR(3, n+1)$. If $r \leq SR(3, n+1)$, then the dimension of $\sigma_r$ is given by

$$dim(\sigma_r) = min\{r(n+1) - 1, \binom{n+3}{3} - 1\},$$

except for $(n, r) = (4, 7)$, for which $dim(\sigma_r) = \binom{n+3}{3} - 2$ [8].

**Definition 11.** ([1]) Denote the symmetric rank of a tensor $T$ by $srank(T)$. For a tensor $T \in \sigma_r$, when $srank(T) > r$, i.e., the length of decomposition is smaller than the symmetric rank, the symmetric border rank, denoted by $sbrank(T)$, is always used. The symmetric border ranks is defined as

$$sbrank(T) = \min_r \{T = \lim_{k \to \infty} T_k, srank(T_k) \leq r\} = \min_r \{T \in \sigma_r\}.$$

In this case, instead of giving the exact border rank, (7) gives the upper bound of the border rank. There is a useful result demonstrating the relations between these ranks.

**Lemma 2.** ([5]) For a symmetric tensor $T \in S^3(\mathbb{C}^{n+1})$,

$$rank(Cat(T)) \leq sbrank(T) \leq srank(T), \tag{7}$$

where the definition of the catalecticant matrix of $T$, denoted by $Cat(T)$, is given by Definition 12.

This lemma is useful for choosing the length of decomposition and determining the symmetric rank. Since $rank(Cat(T))$ gives a lower bound of the symmetric rank, if a length-$t$ decomposition is for $T$ is available, then the range of symmetric rank of $T$ can be obtained, i.e, $rank(Cat(T)) \leq srank(T) \leq t$.

## 5.3 Monomial Indexing and Catalecticant Matrix

In this subsection, we introduce the concepts monomial indexing and catalecticant matrix. For a third-order symmetric tensor $T$ of dimension $n + 1$, it can be represented by an array indexed by a tuple $i = (i_1, i_2, i_3)$, i.e.,

$$T = (T)_{i:0 \leq i_1, i_2, i_3 \leq n}.$$

Since the tensor is symmetric, all the entries indexed by the permutations of $(i_1, i_2, i_3)$ are the same as the one indexed by $(i_1, i_2, i_3)$, so $T$ is uniquely determined by its upper triangular part $(T)_{i:0 \leq i_1 \leq i_2 \leq i_3 \leq n+1}$.

Each entry of the upper triangular part can also be employed a monomial indexing. Let $\theta = (\theta_0, \cdots, \theta_n)$ be a tuple. We can index the upper triangular entries by $\theta$ such that $|\theta| = \theta_0 + \cdots \theta_n = 3$, i.e.,$T_\theta = T_i$ if

$$x_{i_1} x_{i_2} x_{i_3} = x_0^{\theta_0} \cdots x_n^{\theta_n}.$$

Here, we index $T$ using monomials of degree 3, and each $\theta_k$ denotes the number of repeats of number $k$ in index $i$. For example, for a third-order symmetric tensor of dimension 5, the index $(1, 1, 1)$ is the same as the monomial index $(3, 0, 0, 0, 0)$, and the index $(2, 4, 5)$ is the same as $(0, 1, 0, 1, 1)$. Since the monomial indexing and the indexing of upper triangular part of a symmetric tensor $T$ has a one-to-one correspondence, $T$ is also uniquely determined by $T_{\theta:|\theta|=3}$.

Highly related to the monomial indexing is the concept of catalecticant matrix [5]. It is the representing matrix of a linear mapping from the space of homogeneous polynomials of some degree $k$ to the space of symmetric tensor with order $3 - k$.

**Definition 12.** ([5]) The catalecticant matrix of a third order symmetric tensor $T \in S^3(\mathbb{C}^{n+1})$ is a matrix indexed by monomials $\gamma$ and $\theta$, defined as

$$Cat^{3-k,k}(T) = (T_{\gamma+\theta})_{|\gamma|=3-k,|\theta|=k}.$$

Since $k$ can be any nonnegative integers smaller than or equal to 3, there are 4 catalecticant matrices for $T$. In practice, we usually choose $k = 2$, and denote the catalecticant matrix of $T$ as

$$Cat(T) = (T_{\gamma+\theta})_{|\gamma|=1,|\theta|=2}. \tag{8}$$

Note that $T$ can be equivalently indexed by $\alpha = (\alpha_1, \cdots, \alpha_n)$ such that $|\alpha| = \alpha_1 + \cdots + \alpha_n \leq 3$,if we let $x_0 = 1$. Note that $\alpha$ is a monomial of degree less than or equal to 3, and $(\alpha_1, \cdots, \alpha_n) = (\theta_1, \cdots, \theta_n)$. Similarly, we have $T_\alpha = T_i$ if

$$x_{i_1}x_{i_2}x_{i_3} = x_1^{\alpha_1} \cdots x_n^{\alpha_n}.$$

Also, the catalecticant matrix can be indexed by monomials $\beta$ and $\alpha$, i.e.,

$$Cat(T) = (T_{\beta+\alpha})_{|\beta|\leq 1,|\alpha|\leq 2}. \tag{9}$$

**Example 3.** ([1]) Take the tensor $T =$

$$\left[\begin{array}{ccc|ccc|ccc} 7 & -3 & 9 & -3 & 13 & 20 & 9 & 20 & 19 \\ -3 & 13 & 20 & 13 & -27 & 6 & 20 & 6 & 6 \\ 9 & 20 & 19 & 20 & 6 & 6 & 19 & 6 & 45 \end{array}\right]$$

for example. By the definition of catalecticant matrix, $Cat(T) = (T_{\beta+\alpha})_{|\beta|\leq 1,|\alpha|\leq 2} =$

$$\begin{bmatrix} 7 & -3 & 9 & 13 & 20 & 19 \\ -3 & 13 & 20 & -27 & 6 & 6 \\ 9 & 20 & 19 & 6 & 6 & 45 \end{bmatrix},$$

where the index of the three rows are $(0,0), (1,0), (0,1)$ or $1, x_1, x_2$, and the index of the six columns are $(0,0), (1,0), (0,1), (2,0), (1,1), (0,2)$ or $1, x_1, x_2, x_1^2, x_1x_2, x_2^2$.

## 5.4 Generating Polynomials

In this subsection, we illustrate the most important concept for this decomposition method, generating polynomials. The concept of generating polynomials arises from the fact that the entries of a symmetric tensor has recursive patterns, demonstrated by linear relations and described by a set of polynomials. A powerful result obtained by Nie [5] is that the common roots of the polynomials are the decomposition vectors, that is, the $u_i$'s in (5). With this result, we can convert the problem of symmetric tensor decomposition to the problem of finding common zeros of a set of polynomials. Next, we examine the constructions and properties of the generating polynomials.

### 5.4.1 Identification of the Generating Polynomials

There are two ways to identify the generating polynomials. The first one is closely related to the apolarity lemma. Let a polynomial $\widetilde{f} \in \mathbb{C}[x_0, \cdots, x_n]$ be $\widetilde{f}(x_0, \cdots, x_n) = \sum_\theta \widetilde{f}_\theta x_0^{\theta_0} \cdots x_n^{\theta_n}$. A tensor $T$ can be converted to the polynomial form, i.e,

$$T(x_0, \cdots, x_n) = \sum_{i:0\leq i_1,i_2,i_3\leq n} T_i x_{i_1}x_{i_2}x_{i_3} = \sum_{|\theta|=3} T_\theta \frac{3!}{\theta_0! \cdots \theta_n!} x_0^{\theta_0} \cdots x_n^{\theta_n}.$$

**Definition 13.** ([5]) For $\widetilde{f} \in \mathbb{C}[x_0, \cdots, x_n]$ be $\widetilde{f}(x_0, \cdots, x_n) = \sum_\theta \widetilde{f}_\theta x_0^{\theta_0} \cdots x_n^{\theta_n}$, it is said to be apolar to $T$ if

$$\widetilde{f} \circ T = \sum_\theta \widetilde{f}_\theta \frac{\partial^{\theta_0 + \cdots + \theta_n}}{\partial x_0^{\theta_0} \cdots \partial x_n^{\theta_n}} T(x_0, \cdots, x_n) = 0$$

If a polynomial is apolar to $T$, it is said to belong to the aploar ideal of $T$, $Ann(T)$.

The following lemma is the so-called apolarity lemma.

**Lemma 3.** ([9]) A symmetric tensor $T \in S^3(\mathbb{C}^{n+1})$ can be decomposed as in (5), i.e.,

$$T = \sum_{i=1}^r u_i^{\otimes 3},$$

with the $u_i$'s pairwisely linearly independent, if and only if the set of polynomials whose common zeros are the $u_i$'s are apolar to $T$.

**Example 4.** ([1]) Take the tensor $T = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}^{\otimes 3} + \begin{bmatrix} 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}^{\otimes 3}$ as an example. First, we convert $T$ to its polynomial form $T(x_0, \cdots, x_n) = (x_0 + x_1 + \cdots + x_n)^3 + (x_0 - x_1 + \cdots - x_n)^3$. Then, we observe that $p_i(x_0, \cdots, x_n) = x_0^2 - x_i^2$ for $i \in [n]$ are all apolar to $T$, since

$$\frac{\partial^2}{\partial x_0^2} T(x_0, \cdots, x_n) = 6(x_0 + x_1 + \cdots + x_n) + 6(x_0 - x_1 - \cdots - x_n) = \frac{\partial^2}{\partial x_i^2} T(x_0, \cdots, x_n),$$

for $i \in [n]$, which is precisely the definition of apolarity.

With the apolarity lemma and Proposition 2.2 of [5], the generating polynomials can be identified in the following way. Let $\mathbb{C}[x_1, \cdots, x_n]_k$ be the space of polynomials with degree $k$. A polynomial $f(x_1, \cdots, x_n) \in \mathbb{C}[x_1, \cdots, x_n]_k$ with $k \leq m$, the dehomogenization of $\widetilde{f}(x_0, \cdots, x_n)$, is a generating polynomial if and only if $\widetilde{f}(x_0, \cdots, x_n)$ is apolar to $T$, i.e.,

$$\widetilde{f}(x_0, \cdots, x_n) = x_0^k f(\frac{x_1}{x_0}, \cdots, \frac{x_n}{x_0}) \in Ann(T).$$

Another approach corresponds to the formal definition of generating polynomials, and is more useful in practice. We say that $f \in \mathbb{C}[x_1, \cdots, x_n]_k$ is a generating polynomial of $T$ if the vector of coefficients of $f$ belongs to the null space of the catalecticant matrix of $T$, i.e,

$$vec(f) \in Ker(Cat(T)), \tag{10}$$

where $vec(f)$ is the vector of coefficients of $f$. Note that this condition is equivalent to that for every $\beta = (\beta_1, \cdots, \beta_n)$ such that $|\beta| \leq 3 - deg(f)$,

$$\sum_{\alpha:|\alpha| \leq 3} (f(x_1, \cdots, x_n) x_1^{\beta_1} \cdots x_n^{\beta_n})_\alpha T_\alpha = 0,$$

because the catalecticant matrix is given by $Cat(T) = (T_{\beta+\alpha})_{|\beta| \leq 1, |\alpha| \leq 2}$.

### 5.4.2 Construction of the Generating Polynomials

Next, we describe how to construct these polynomials.

**Definition 14.** ([5]) Let $\mathbb{B}_0$ be the set of first $r$ monomials in $\{1, x_1, \cdots, x_n, x_1^2, x_1x_2, \cdots\}$ and $\mathbb{B}_1 = (\mathbb{B}_0 \cup x_1\mathbb{B}_0 \cup \cdots \cup x_n\mathbb{B}_0) \setminus \mathbb{B}_0$. For simplicity, we write $\beta \in \mathbb{B}_0$ if $x_1^{\beta_1} \cdots x_n^{\beta_n} \in \mathbb{B}_0$ and $\alpha \in \mathbb{B}_1$ if $x_1^{\alpha_1} \cdots x_n^{\alpha_n} \in \mathbb{B}_1$. Let $F \in \mathbb{C}^{\mathbb{B}_0 \times \mathbb{B}_1}$ be a matrix indexed by $(\beta, \alpha)$. For any $\alpha \in \mathbb{B}_1$, consider the polynomial

$$f[F, \alpha] = \sum_{\beta \in \mathbb{B}_0} F_{\beta,\alpha} x_1^{\beta_1} \cdots x_n^{\beta_n} - x_1^{\alpha_1} \cdots x_n^{\alpha_n}. \tag{11}$$

The set of $|\mathbb{B}_1|$ polynomials are the generating polynomials, $F$ is called the generating matrix, and the entries of $F$ are the coefficients of $f[F, \alpha]$.

Now we illustrate how to obtain the coefficients. Since we have to make sure that the common zeros of the generating polynomials can be used to construct the decomposition as in (5), we have to first ensure that there are $r$ common roots of the system of polynomials, where $r$ is the length of decomposition. Consider the linear map from the quotient space $\mathbb{C}[x_1, \cdots, x_n]/\langle (f[F, \alpha])_{\alpha \in \mathbb{B}_1} \rangle$, where $\langle (f[F, \alpha])_{\alpha \in \mathbb{B}_1} \rangle$ is the ideal generated by the set of polynomials, to itself, represented by a matrix $N_{x_i}$, that maps a polynomial $p$ to $x_i p$. The matrices $N_{x_i}(F)$ for $i \in [n]$ are called the companion matrices of $F$.

**Definition 15.** ([5]) Let $\mathbb{B}_0$ and $\mathbb{B}_1$ be defined as above, and $\gamma \in \mathbb{B}_0$ and $\kappa \in \mathbb{B}_1$.Then the companion matrices are $N_{x_i}(F)$ of the set of generating polynomials $f[F, \alpha])_{\alpha \in \mathbb{B}_1}$ are defined by

$$N_{x_i}(F)_{\gamma,\kappa} = \begin{cases} 0 & \text{if } x_i x_1^{\kappa_1} \cdots x_n^{\kappa_n} \in \mathbb{B}_0 \text{ and } \gamma = \kappa + e_i \\ 1 & \text{if } x_i x_1^{\kappa_1} \cdots x_n^{\kappa_n} \in \mathbb{B}_0 \text{ and } \gamma \neq \kappa + e_i \\ F_{\gamma,\kappa+e_i} & \text{if } x_i x_1^{\kappa_1} \cdots x_n^{\kappa_n} \in \mathbb{B}_1 \end{cases}$$

**Example 5.** ([1]) Take the tensor $T =$

$$\left[\begin{array}{ccc|ccc|ccc} 7 & -3 & 9 & -3 & 13 & 20 & 9 & 20 & 19 \\ -3 & 13 & 20 & 13 & -27 & 6 & 20 & 6 & 6 \\ 9 & 20 & 19 & 20 & 6 & 6 & 19 & 6 & 45 \end{array}\right]$$

for example. By the definition of catalecticant matrix, $Cat(T) = (T_{\beta+\alpha})_{|\beta| \leq 1, |\alpha| \leq 2} =$

$$\begin{bmatrix} 7 & -3 & 9 & 13 & 20 & 19 \\ -3 & 13 & 20 & -27 & 6 & 6 \\ 9 & 20 & 19 & 6 & 6 & 45 \end{bmatrix}.$$

We have $\mathbb{B}_0 = \{1, x_1, x_2\}$ and $\mathbb{B}_1 = \{x_1^2, x_1x_2, x_2^2\}$. Then the set of generating polynomials is

$$\begin{cases} F_{(0,0),(2,0)} + F_{(1,0),(2,0)}x_1 + F_{(0,1),(2,0)}x_2 - x_1^2 \\ F_{(0,0),(1,1)} + F_{(1,0),(1,1)}x_1 + F_{(0,1),(1,1)}x_2 - x_1x_2 \\ F_{(0,0),(0,2)} + F_{(1,0),(0,2)}x_1 + F_{(0,1),(0,2)}x_2 - x_2^2 \end{cases}.$$

If we compute the kernels of the three corresponding submatrices of the catalecticant matrix and scale them, we get the system of generating polynomials

$$\begin{cases} \frac{14}{5} - \frac{1}{5}x_1 - \frac{4}{5}x_2 - x_1^2 \\ \frac{4}{5} - \frac{6}{5}x_1 + \frac{6}{5}x_2 - x_1x_2 \\ \frac{14}{5} + \frac{4}{5}x_1 + \frac{1}{5}x_2 - x_2^2 \end{cases}.$$

Furthermore, the companion matrices are

$$N_{x_1} = \begin{bmatrix} 0 & 14/5 & 4/5 \\ 1 & -1/5 & -6/5 \\ 0 & -4/5 & 6/5 \end{bmatrix}, N_{x_2} = \begin{bmatrix} 0 & 4/5 & 14/5 \\ 0 & -6/5 & 4/5 \\ 1 & 6/5 & 1/5 \end{bmatrix}.$$

The necessary condition for $(f[F,\alpha])_{\alpha\in\mathbb{B}_1}$ to have $r$ complex roots is that the companion matrices commute, i.e.,

$$N_{x_i}(F)N_{x_j}(F) = N_{x_j}(F)N_{x_i}(F), i < j, i, j \in [n]. \tag{12}$$

If the set of generating polynomials has $r$ solutions, then they are called consistent generating polynomials. In addition to ensuring that there are $r$ common roots of the system of polynomials, we have to make sure that the $r$ roots are distinct so that they can be used in symmetric tensor decomposition. By proposition 2.4 of [5], this is true if and only if the companion matrices $N_{x_i}(F)$ are simultaneously diagonalizable. In this case, the generating polynomials are called nondefective.

### 5.4.3 Properties of the Generating Polynomials

For a tensor $T \in S^3(\mathbb{C}^{n+1})$, the symmetric decomposition is $T = \sum_{i=1}^r u_i^{\otimes 3}$ for $u_i \in \mathbb{C}^{n+1}$. If $[u_i]_0 \neq 0$, which is usually the case, then the decomposition is equivalent to

$$T = \sum_{i=1}^{r} \lambda_i \begin{bmatrix} 1 \\ v_i \end{bmatrix}^{\otimes 3}, \tag{13}$$

where $\lambda_i = ([u_i]_0)^3$ and $v_i = \begin{bmatrix} [u_i]_1/[u_i]_0 \\ \vdots \\ [u_i]_n/[u_i]_0 \end{bmatrix}$. Next, we introduce some useful results about the properties of the generating polynomials proposed by Nie [5]. The first result is important since it illustrates the relationship between the generating polynomials and the symmetric tensor decomposition.

**Theorem 9.** *([5]) If $b_1, \cdots, b_r$ are the distinct zeros of the generating polynomials $(f[F,\alpha])_{\alpha\in\mathbb{B}_1}$, then $(b_1, \cdots, b_r)$ is a permutation of $(v_1, \cdots, v_r)$ in (14), i.e.,*

$$T = \sum_{i=1}^{r} \lambda_{\pi(i)} \begin{bmatrix} 1 \\ b_i \end{bmatrix}^{\otimes 3},$$

*where the $\lambda_{\pi(i)}$'s are the the $\lambda_i$'s permuted in the same way.*

*Proof.* Given $\alpha \in \mathbb{B}_1$, denote the homogenization of $f[F,\alpha](x_1, \cdots, x_n)$ as $\widetilde{f}[F,\alpha](x_0, \cdots, x_n)$. By proposition 2.2 of [5], $\widetilde{f}[F,\alpha](x_0, \cdots, x_n)$ is apolar to $T$. Since the $v_i$'s are distinct zeros of $(f[F,\alpha])_{\alpha\in\mathbb{B}_1}$, they are pairwisely linearly independent. Since the $\begin{bmatrix} 1 \\ v_i \end{bmatrix}$'s are the pairwisely linearly independent common zeros of $(\widetilde{f}[F,\alpha])_{\alpha\in\mathbb{B}_1}$, we can find the $\lambda_i$'s to construct the symmetric tensor decomposition as in (14) by the apolarity lemma. □

This result proves that the common zeros of the generating polynomials can be used to construct the symmetric decomposition of the tensor. Recall that $\sigma_r$ is the Zariski closure of equivalent classes of $\sum_{i=1}^{r} u_i^{\otimes 3}$ in $\mathbb{P}S^3(\mathbb{C}^{n+1})$ and let $d = r(n+1) - 1 - dim(\sigma_r)$ be the dimension gap. The second result, by proposition 3.8 of [5], is that for a general tensor $T \in \sigma_r$, the dimension of the space of consistent generating polynomials is $d$, and if $d > 0$, the consistent generating polynomials in a subspace of $\mathbb{C}^{\mathbb{B}_0 \times \mathbb{B}_1}$ with codimension $d$ form a finite set. This result specifies the condition of finitness of the consistent generating matrices. The third result, by proposition 3.10 of [5], is concerned with the nondefectiveness of the generating polynomials, that is, the generating polynomials are defective, i.e., have a repeated root only if the discriminants of the companion matrices are all 0, i.e.,

$$dis(N_{x_1}(F)) = \cdots = dis(N_{x_n}(F)) = 0$$

where $dis(N_{x_i}(F)) = \prod_{i<j}(\lambda_i - \lambda_j)^2$, with the $\lambda_i$'s as the eigenvalues of $N_{x_i}(F)$. If one of the discriminants is not 0, then the generating polynomials are nondefective.

## 5.5 Implementation of Symmetric Tensor Decomposition

In this subsection, we examine the implementation methods used in the symmetric tensor decomposition. We consider how to accomplish the two main goals: finding a generating matrix $F$ such that the set of generating polynomials $(f[F,\alpha])_{\alpha \in \mathbb{B}_1}$ is nondefective, and obtaining the common zeros $v_1, \cdots, v_r$ of $(f[F,\alpha])_{\alpha \in \mathbb{B}_1}$ and the corresponding coefficients $\lambda_1, \cdots, \lambda_r$. We explore the methods proposed by Nie [5].

### 5.5.1 Finding the Generating Matrix

Let us first see how to get the generating matrix $F$, which provides the coefficients of the generating polynomials $(f[F,\alpha])_{\alpha \in \mathbb{B}_1}$. Let $T \in S^3(\mathbb{C}^{n+1})$ be a symmetric cubic tensor. For $\alpha \in \mathbb{B}_1$, define

$$A[T,\alpha]_{\gamma,\beta} = T_{\gamma+\beta}, |\gamma| \leq 3 - |\alpha|, \beta \in \mathbb{B}_0,$$

$$b[T,\alpha]_{\gamma} = T_{\gamma+\alpha}, |\gamma| \leq 3 - |\alpha|.$$

The generating polynomials $(f[F,\alpha])_{\alpha \in \mathbb{B}_1}$ are consistent if the columns of the generating matrix $F$ satisfies that

$$A[T,\alpha]F_{:,\alpha} = b[T,\alpha].$$

This system of equations may have no solution if the symmetric rank of $T$ is greater than $r$, in which case the length of decomposition $r$ needs to be increased.

**Example 6.** ([5]) Take the tensor $T =$

$$\left[\begin{array}{ccc|ccc|ccc} 7 & -3 & 9 & -3 & 13 & 20 & 9 & 20 & 19 \\ -3 & 13 & 20 & 13 & -27 & 6 & 20 & 6 & 6 \\ 9 & 20 & 19 & 20 & 6 & 6 & 19 & 6 & 45 \end{array}\right]$$

for example. We have $\mathbb{B}_0 = \{1, x_1, x_2\}$ and $\mathbb{B}_1 = \{x_1^2, x_1x_2, x_2^2\}$. Then the set of generating polynomials is

$$\begin{cases} F_{(0,0),(2,0)} + F_{(1,0),(2,0)}x_1 + F_{(0,1),(2,0)}x_2 - x_1^2 \\ F_{(0,0),(1,1)} + F_{(1,0),(1,1)}x_1 + F_{(0,1),(1,1)}x_2 - x_1x_2 \\ F_{(0,0),(0,2)} + F_{(1,0),(0,2)}x_1 + F_{(0,1),(0,2)}x_2 - x_2^2 \end{cases}.$$

Then $A[T,\alpha]$ for $\alpha \in \mathbb{B}_1$ are

$$A[T,(2,0)] = A[T,(1,1)] = A[T,(0,2)] = \begin{bmatrix} T_{(0,0)} & T_{(1,0)} & T_{(0,1)} \\ T_{(1,0)} & T_{(2,0)} & T_{(1,1)} \\ T_{(0,1)} & T_{(1,1)} & T_{(0,2)} \end{bmatrix},$$

and $b[T,\alpha]$ for $\alpha \in \mathbb{B}_1$ are

$$b[T,(2,0)] = \begin{bmatrix} T_{(2,0)} \\ T_{(3,0)} \\ T_{(2,1)} \end{bmatrix}, b[T,(1,1)] = \begin{bmatrix} T_{(1,1)} \\ T_{(2,1)} \\ T_{(1,2)} \end{bmatrix}, b[T,(0,2)] = \begin{bmatrix} T_{(0,2)} \\ T_{(1,2)} \\ T_{(0,3)} \end{bmatrix}.$$

If the system of equations is consistent, then we can solve for the generating matrix $\widehat{F}$. In the context of cross moment tensors, since the symmetric rank $r$ is assumed to be smaller than the dimension $n$, the consistent system, according to the construction of $\mathbb{B}_0$ and $\mathbb{B}_1$, has a unique solution $\widehat{F}$. Due to the noise in the cross moment tensors, the companion matrices $N_{x_i}(\widehat{F})$ commute approximately and are simultaneously diagonalizable approximately.

### 5.5.2 Finding the Roots of the Generating Polynomials

Next, we examine the method to compute the common roots of the system of generating polynomials $(f[\widehat{F},\alpha])_{\alpha\in\mathbb{B}_1}$. According to the implication of Stickelberger's theorem [24], the set of common zeros to the generating polynomials is the set of eigenvalues corresponding to the common eigenvector of the companion matrices $N_{x_i}(\widehat{F})$ for $i \in [n]$. Hence, the goal is to find the common eigenvectors of the companion matrices.

Corless, Gianni and Trager [10] proposed a more practical method than finding the common eigenvectors of $N_{x_i}(F(\widehat{\eta}))$. Here, we assume the set of polynomials is nondefective. Pick a tuple of positive scalars $\tau = (\tau_1, \cdots, \tau_n)$ such that $\sum_{i=1}^n \tau_i = 1$. Let $N(\tau)$ be a linear combination of $N_{x_i}(F(\widehat{\eta}))$,

$$N(\tau) = \sum_{i=1}^{n} \tau_i N_{x_i}(\widehat{F}).$$

Then, obtain the Schur decomposition $N(\tau) = QUQ^*$, where $Q \in \mathbb{C}^{r\times r}$ is unitary and $U \in \mathbb{C}^{r\times r}$ is an upper triangular matrix

$$U = \begin{bmatrix} U_{11} & U_{12} & \dots & U_{1r} \\ 0 & U_{22} & \dots & U_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & U_{rr} \end{bmatrix}.$$

Next we compute $\widetilde{N}_{x_i}(\widehat{F}) = Q^* N_{x_i}(\widehat{F}) Q$, so that

$$\widetilde{N}_{x_i}(F(\eta)) = \begin{bmatrix} N_{11}^{(i)} & N_{12}^{(i)} & \dots & N_{1r}^{(i)} \\ 0 & N_{22}^{(i)} & \dots & N_{2r}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & N_{rr}^{(i)} \end{bmatrix}.$$

The vectors $v_j = \begin{bmatrix} N_{jj}^{(1)} \\ \vdots \\ N_{jj}^{(n)} \end{bmatrix}$ for $j \in [r]$ are the $r$ roots of the set of generating polynomials $(f[\widehat{F}, \alpha])_{\alpha \in \mathbb{B}_1}$.

## 5.6 Application to Parameter Estimation

In this subsection, we relate parameter estimation of the latent variable model to the symmetric tensor decomposition introduced in the previous subsections by applying the method to the third-order cross moment $T$. We look at the exchangeable single topic model, in which the cross moments have the form $M = \sum_{i=1}^{r} w_i \mu_i^{\otimes 2}$ and

$$T = \sum_{i=1}^{r} w_i \mu_i^{\otimes 3},$$

with $\mu_i \in \mathbb{R}^{n+1}$. Applying the previous method to the the cross moment tensor $T$, we get the decomposition

$$T = \sum_{i=1}^{r} u_i^{\otimes 3}.$$

Now, the task is to recover the weights $w_i$'s and the conditional expectations $\mu_i$'s from the decomposition vectors $u_i$'s. Note that the obtained decomposition can be written as

$$T = \sum_{i=1}^{r} w_i \left(\frac{u_i}{\sqrt[3]{w_i}}\right)^{\otimes 3}, \tag{14}$$

where $\mu_i = \frac{u_i}{\sqrt[3]{w_i}}$.

The next step is to determine the weights $w_i$ so that we can also get the conditional expectations $\mu_i = \frac{u_i}{\sqrt[3]{w_i}}$. The easiest way is to use the first-order moment $E[x_1]$. In the exchangeable single topic model,

$$E[x_1] = \sum_{i=1}^{r} P(h = i) E[x_1 | h = i] = \sum_{i=1}^{r} w_i \mu_i = \sum_{i=1}^{r} w_i \frac{u_i}{\sqrt[3]{w_i}} = \sum_{i=1}^{r} w_i^{2/3} u_i.$$

Then, we can form a system of linear equations to solve for $w_i^{2/3}$, i.e.,

$$Uw = V, \tag{15}$$

where

$$U = [u_1 | \cdots | u_r],\ w = \begin{bmatrix} w_1^{2/3} \\ \cdots \\ w_r^{2/3} \end{bmatrix},\ V = E[x_1].$$

Here, $U \in \mathbb{R}^{(n+1)\times r}$, $w \in \mathbb{R}^r$ and $V \in \mathbb{R}^{n+1}$. Since $r < n+1$, this system is overdetermined and has a unique solution. This ensures that we can identify the weights $w_i$ from the solution $w$.

**Theorem 10.** *Suppose that a symmetric tensor $T \in S^3(\mathbb{C}^{n+1})$ has a symmetric decomposition with $T = \sum_{i=1}^{r} u_i^{\otimes 3}$ where $r < n + 1$ and the $u_i$'s are linearly independent. Then the symmetric decomposition is unique. Also, in order to recover the $w_i$'s and $\mu_i$'s in $T = \sum_{i=1}^{r} w_i (\frac{u_i}{\sqrt[3]{w_i}})^{\otimes 3}$, we can solve the system of linear equations $Uw = V$ as in (15), which has a unique solution.*

### 5.7 Algorithm for Parameter Estimation

In this subsection, we propose an algorithm for parameter estimation of the latent variable models using the techniques and methods introduced in Section 5. Suppose that the empirical first-order cross moment $\widehat{V}$ and the empirical third-order cross moment $\widehat{T}$ are available. Also, for simplicity, assume the length of decomposition $r$ is given. Algorithm 2 is provided to give the estimated parameters.

**Algorithm 2**

**Input** vector $\widehat{V} \in \mathbb{R}^{n+1}$, symmetric tensor $\widehat{T} \in S^3(\mathbb{R}^{n+1})$, length of decomposition $r$

**Output** the estimated parameters $\widehat{w}_i$, $\widehat{\mu}_i$

1: Solve the system of equations $A[\widehat{T}, \alpha]\widehat{F}_{:,\alpha} = b[\widehat{T}, \alpha]$ to get the generating matrix $\widehat{F}$

2: Obtain the companion matrices $N_{x_i}(\widehat{F})$, generate a tuple $\tau = (\tau_1, \cdots, \tau_n)$, and compute $N(\tau) = \sum_{i=1}^{n} \tau_i N_{x_i}(\widehat{F})$

3: Obtain the Schur decomposition $N(\tau) = QUQ^*$, compute $\widetilde{N}_{x_i}(\widehat{F}) = Q^* N_{x_i}(\widehat{F})Q$, and get $v_j \in \mathbb{R}^n$ with the $i$-th entry equal to the $jj$-th entry of $\widetilde{N}_{x_i}(\widehat{F})$, for $i \in [n]$ and $j \in [r]$

4: Solve $\widehat{T} = \sum_{i=1}^{r} \lambda_i \begin{bmatrix} 1 \\ v_i \end{bmatrix}^{\otimes 3}$ and obtain $u_i = \sqrt[3]{\lambda_i} \begin{bmatrix} 1 \\ v_i \end{bmatrix}$

5: Solve $U\widehat{w} = \widehat{V}$, where the entries of $w$ are $\widehat{w}_i^{2/3}$

6: **return** $\widehat{w}_i$ and $\widehat{\mu}_i = \frac{u_i}{\sqrt[3]{\widehat{w}_i}}$

# 6 Numerical Experiments

### 6.1 Tensor Power Method

In this subsection, we assess the tensor power method using the average Euclidean norms of the differences of the actual eigenvectors and the estimated eigenvectors. The numerical experiment is done in MATLAB. First, we create $10^7$ samples of $x_1, x_2, x_3$ of the exchangeable single topic model randomly. Next, we estimate the second-order and third-order cross moments using the method of moments. Then, we orthogonalize the third-order cross moment as described in Section 4.3, and implement the tensor power method on the orthogonalized third-order moment. The results are presented in the following tables. We use the length of decomposition $r = 2, 3, 4$ and dimension $n = 5, 10, 20, 30, 40, 50$.

| Dimension $n$ | Error |
|---|---|
| 5 | $4.60 \times 10^{-4}$ |
| 10 | $5.71 \times 10^{-4}$ |
| 20 | 0.0012 |
| 30 | $7.19 \times 10^{-4}$ |
| 40 | $9.83 \times 10^{-4}$ |
| 50 | $6.88 \times 10^{-4}$ |

Table 1: Tensor Power Method Error Measured by Average Euclidean Norm for $r = 2$

| Dimension $n$ | Error |
|---|---|
| 5 | $8.14 \times 10^{-4}$ |
| 10 | 0.0018 |
| 20 | 0.0020 |
| 30 | 0.015 |
| 40 | 0.0014 |
| 50 | 0.0013 |

Table 2: Tensor Power Method Error Measured by Average Euclidean Norm for $r = 3$

| Dimension $n$ | Error |
|---|---|
| 5 | 0.0040 |
| 10 | 0.0026 |
| 20 | 0.0029 |
| 30 | 0.0023 |
| 40 | 0.0022 |
| 50 | 0.0022 |

Table 3: Tensor Power Method Error Measured by Average Euclidean Norm for $r = 4$

As we can see, for $r = 2$, the error (the average Euclidean norms of the differences of the actual eigenvectors and the estimated eigenvectors) is small except for $n = 20$. For $r = 3, 4$ the errors are larger. For $r = 3$, we have a relatively small error for $n = 5$ and a relatively large error for $n = 30$. For $r = 4$, the errors for different dimensions are similar. Also, it is clear that as $r$ becomes larger, the error becomes larger, which may be a defect of the tensor power method.

We also record the time used for the tensor power method, which is the time elapsed until the end after the estimation of the cross moments. As we can see in the following tables, it takes approximately 3 to 21 seconds for the power iteration to give the estimated robust eigenvectors. As $n$ and $r$ grows, it takes longer times for this method to compute the results. For $n = 50$, it takes approximately 5 more seconds than the case of $n = 5$, and for $r = 4$, it takes approximately 13 more seconds than the case of $r = 2$. Hence, we doubt if the tensor power method could work efficiently for larger values of $n$ and $r$.

| Dimension $n$ | Time(s) |
|---|---|
| 5 | 3.076 |
| 10 | 3.137 |
| 20 | 3.778 |
| 30 | 3.503 |
| 40 | 4.10 |
| 50 | 8.98 |

Table 4: Tensor Power Method Elapsed Time for $r = 2$

| Dimension $n$ | Time(s) |
|---|---|
| 5 | 9.005 |
| 10 | 9.415 |
| 20 | 9.582 |
| 30 | 9.748 |
| 40 | 8.860 |
| 50 | 13.252 |

Table 5: Tensor Power Method Elapsed Time for $r = 3$

| Dimension $n$ | Time(s) |
|---|---|
| 5 | 16.980 |
| 10 | 16.312 |
| 20 | 16.958 |
| 30 | 17.094 |
| 40 | 17.932 |
| 50 | 21.298 |

Table 6: Tensor Power Method Elapsed Time for $r = 4$

## 6.2 Generating Polynomials Method

In this subsection, we test symmetric tensor decomposition by the method of generating polynomials using the average Euclidean norms of the differences of the actual decomposition vectors and the estimated decomposition vectors. The numerical experiment is done in MATLAB. In order to compare this method to the tensor power method, we also use the length of decomposition $r = 2, 3, 4$ and dimension $n + 1 = 5, 10, 20, 30, 40, 50$. First, we create $r$ linearly independent vectors of dimension $n + 1$ randomly, and create a third-order symmetric tensor using the sum of the tensor product of the generated vectors. Next, we obtain the catalecticant matrix and the generating polynomials as in Section 5.3 and 5.4. Then, we obtain the companion matrices and compute the common roots of the generating polynomials as described in Section 5.4 and 5.5. Let us examine five random examples,

| Dimension $n + 1$ | Error |
|---|---|
| 5 | $1.06 \times 10^{-15}$ |
| 10 | $3.54 \times 10^{-16}$ |
| 20 | $3.71 \times 10^{-15}$ |
| 30 | $5.50 \times 10^{-16}$ |
| 40 | $4.88 \times 10^{-14}$ |
| 50 | $8.79 \times 10^{-16}$ |

Table 7: Generating Polynomials Method Error Measured by Average Euclidean Norm for $r = 2$

| Dimension $n+1$ | Error |
|---|---|
| 5 | $2.46 \times 10^{-16}$ |
| 10 | $4.33 \times 10^{-15}$ |
| 20 | $1.40 \times 10^{-14}$ |
| 30 | $3.12 \times 10^{-15}$ |
| 40 | $2.70 \times 10^{-13}$ |
| 50 | $3.77 \times 10^{-12}$ |

Table 8: Generating Polynomials Method Error Measured by Average Euclidean Norm for $r = 3$

| Dimension $n+1$ | Error |
|---|---|
| 5 | $5.10 \times 10^{-16}$ |
| 10 | $1.87 \times 10^{-14}$ |
| 20 | $2.10 \times 10^{-15}$ |
| 30 | $1.74 \times 10^{-14}$ |
| 40 | $1.45 \times 10^{-14}$ |
| 50 | $9.06 \times 10^{-15}$ |

Table 9: Generating Polynomials Method Error Measured by Average Euclidean Norm for $r = 4$

In general, the method of generating polynomials behaves much better than the tensor power method, since the error (the average Euclidean norms of the differences of the actual eigenvectors and the estimated vectors) is significantly smaller for $r = 2, 3, 4$. In fact, the errors for all cases are smaller than $4 \times 10^{-12}$, which is about $10^8$ to $10^9$ times smaller than the error given by power iteration, and the errors for most cases are smaller than $2\times10^{-14}$, which is about $10^{10}$ to $10^{11}$ times smaller than the error given by power iteration. Thus, we can claim that the generating polynomials method does a much better job in recovering the linearly independent decomposition vectors of the symmetric tensors.

With respect to the time of computation, the generating polynomials method still performs much better. We record the time used for the tensor power method, which is the time elapsed until the end after the construction of the third-order tensor. As we can see in the following tables, it takes approximately 0.16 to 0.29 seconds for the generating polynomials method to give the estimated decomposition vectors, which is approximately 19 to 73 times faster than the power iteration. As $n$ and $r$ grows, the time od computation barely changes. For $n = 50$, it takes 0.255 seconds in the case of $r = 2$, 0.261 seconds in the case of $r = 3$, and 0.279 seconds in the case of $r = 4$. Thus, we may expect that the method of generating polynomials would work efficiently for larger values of $n$ and $r$, as opposed to the tensor power method.

| Dimension $n+1$ | Time(s) |
|---|---|
| 5 | 0.169 |
| 10 | 0.212 |
| 20 | 0.205 |
| 30 | 0.239 |
| 40 | 0.238 |
| 50 | 0.255 |

Table 10: Generating Polynomials Method Elapsed Time for $r = 2$

| Dimension $n+1$ | Time(s) |
|---|---|
| 5 | 0.195 |
| 10 | 0.225 |
| 20 | 0.240 |
| 30 | 0.269 |
| 40 | 0.251 |
| 50 | 0.261 |

Table 11: Generating Polynomials Method Elapsed Time for $r = 3$

| Dimension $n+1$ | Time(s) |
|---|---|
| 5 | 0.192 |
| 10 | 0.260 |
| 20 | 0.245 |
| 30 | 0.255 |
| 40 | 0.296 |
| 50 | 0.279 |

Table 12: Generating Polynomials Method Elapsed Time for $r = 4$

## 7 Conclusion

In this paper, we analyze the symmetric tensor structures of the cross moments of latent variable models like the single topic model and the multi-view model, and assess two methods for symmetric tensor decomposition, based on orthogonal tensor decomposition and generating polynomials respectively. As a consequence, the parameter estimation of the latent variable models mentioned in this paper can be done by symmetric tensor decomposition using the two methods.

The method of generating polynomials can be applied if the symmetric tensor is not orthogonally decomposable (odeco), since the eigenvectors of the tensor are not decomposition vectors anymore, and hence the tensor power method will not converge. Moreover, the method of generating polynomials gives significantly smaller errors, and cost significantly less computation time, so we may claim that the method of generating polynomials is more applicable and efficient than the power iteration.

## Acknowledgment

I would like to thank Professor Jiawang Nie for the supervision and guidance, and also appreciate the opportunity to enroll in the honors thesis program with Professor Jiawang Nie as my advisor. I am also grateful to Bingni Guo and Zi Yang for helpful comments.

## References

[1] J. Nie. *Tensor Computation Handout for MATH 277A*. UC San Diego, Handout for Graduate Mathematics, 2018.

[2] A. Anandkumar, R. Ge, D.Hsu, S. M. Kakade, and M. Telgarsky. *Tensor Decompositions for Learning Latent Variable Models*. J. Mach. Learn. Res., 15 (2014), pp. 2773-2832.

[3] D. Hsu and S. M. Kakade. *Learning mixtures of spherical Gaussians: moment methods and spectral decompositions*. Fourth Innovations in Theoretical Computer Science, 2013.

[4] L. Qi. *Eigenvalues of a real supersymmetric tensor*. Journal of Symbolic Computation, 40 (6):1302-1324, 2005.

[5] J. Nie. *Generating polynomials and symmetric tensor decompositions*. Found. Comput. Math. 17 (2017), no. 2, 423–465.

[6] L. Chiantini, G. Ottaviani, and N. Vannieuwenhoven. *On generic identifiability of symmetric tensors of subgeneric rank*. Trans. Amer. Math. Soc., 369 (2017), 4021-4042.

[7] J. Alexander and A. Hirschowitz. *Polynomial interpolation in several variables*. J. Algebraic Geom. 4 (1995), pp. 201-22.

[8] J. M. Landsberg. *Tensors: geometry and applications*. Graduate Studies in Mathematics, 128. American Mathematical Society, Providence, RI, 2012.

[9] A. Iarrobino and V. Kanev. *Power Sums, Gorenstein algebras, and determinantal varieties*. Lecture Notes in Mathematics 1721, Springer, 1999.

[10] R. M. Corless, P. M. Gianni and B. M. Trager. *A reordered Schur factorization method for zero-dimensional polynomial systems with multiple roots*. Proc. ACM Int. Symp. Symbolic and Algebraic Computation, 133-140, Maui, Hawaii, 1997.

[11] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y.-K. Liu. *A spectral algorithm for latent Dirichlet allocation*. dvances in Neural Information Processing Systems 25, 2012a.

[12] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. *On the best rank-1 and rank-$(R_1, R_2, \cdots, R_n)$ approximation and applications of higher-order tensors*. SIAM J. Matrix Anal. Appl., 21 (4):1324-1342, 2000.

[13] L. Oeding and G. Ottaviani. *Eigenvectors of tensors and algorithms for Waring decomposition*. J. Symbolic Comput. 54, 9-35, 2013.

[14] D. Cox, J. Little and D. O'Shea. *deals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer, 2007.

[15] E. Robeva. *Orthogonal Decomposition of Symmetric Tensors*. SIAM J. Matrix Anal. Appl., 37(1), 86–102, 2016.

[16] T. G. Kolda, J. R. Mayo. *Shifted Power Method for Computing Tensor Eigenpairs* SIAM J. Matrix Anal. Appl., 32(4), 1095–1124, 2011.

[17] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999.

[18] E. Kofidis and P. A. Regalia. *On the best rank-1 approximation of higher-order supersymmetric tensors*. SIAM Journal on Matrix Analysis and Applications, 23 (2002), pp. 863-884.

[19] D. Cartwright and B. Sturmfels. *The Number of Eigenvalues of a Tensor*. Linear Algebra and its Applications, 432:2 (2013) 942-952.

[20] A. Anandkumar, D. Hsu, and S. M. Kakade. *A method of moments for mixture models and hidden Markov models*. In Twenty-Fifth Annual Conference on Learning Theory, volume 23, pages 33.1-33.34, 2012c.

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

[22] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. *An introduction to independent component analysis*. J. Chemometrics, Special Issue Multi-way Analysis, 14 (2000), pp. 123–149.

[23] T. Austin. *On exchangeable random variables and the statistics of large graphs and hypergraphs*. Probab. Survey, 5:80–145, 2008.

[24] B. Sturmfels. *Solving systems of polynomial equations*. CBMS Regional Conference Series in Mathematics, 97. American Mathematical Society, Providence, RI, 2002.