Tensor Decomposition

Qiang Ning

Sep 19, 2017

1 Introduction

Matrices are two-dimensional arrays. As a generalization of matrix, tensor can be defined as follows.

Definition 1.1. An element in $\mathbb{F}^{n_1 \times n_2 \times \cdots \times n_k}$, where $\mathbb{F}$ is an arbitrary field of numbers (we used $\mathbb{R}$ throughout this note) is called a k-th order $n_1 \times n_2 \times \cdots \times n_k$ tensor.

We can see that this definition is compatible with vectors and matrices: vectors are simply 1st order tensors and matrices are 2nd order tensors. A downside of this view of tensors is the lack of geometric insight. In fact, tensors can also be defined as multilinear maps (similarly to matrices which can be defined as linear maps), but here we restrict to the multidimensional array standpoint.

Before working on tensors, we need to introduce the tensor product operator "$\otimes$", which provides a concise representation for tensors.

Definition 1.2. If $V = v_1 \otimes v_2 \otimes \cdots \otimes v_k$, where $v_i \in \mathbb{R}^{n_i}$, $\forall i = 1, \ldots, k$, then V is a k-th order $n_1 \times n_2 \times \cdots \times n_k$ tensor with its $(j_1, \ldots, j_k)$-th entry being $\prod_{i=1}^{k} v_{i,j_i}$.

Note this definition is also compatible with the outer product of two vectors. Suppose $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$; then $u \cdot v^T \in \mathbb{R}^{m \times n}$ is a matrix, which is exactly the 2nd order tensor specified by $u \otimes v$.

It is no doubt that tensors are computationally expensive to work with (due to the curse of dimensionality). In recent years, however, researchers have designed successful (and relatively efficient) learning algorithms for latent variable models based on tensor decomposition, either implicitly or explicitly [1]. One may ask: what are the unique advantages that tensors can bring us. Therefore, in this lecture note, we will first review an example from Spearman on matrix decomposition [2, 3], from which we can see why matrices are inconvenient in certain settings. We then introduce tensor decomposition as a generalization of matrix decomposition and also its unique properties that make it different to matrices. Finally, we take the mixture of Gaussians as an exemplary latent variable model and show how tensor decomposition can be used to learn the model parameters.

1.1 Matrix Decomposition

Theorem 1.3. *Singular-value decomposition (SVD): Suppose $A \in \mathbb{R}^{m\times n}$. Then there exists a decomposition of the form*

$$A = U\Sigma V^T,$$

where $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix with non-negative diagonal entries, $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$, and U and V are both orthogonal matrices (i.e., $U^{-1} = U^T$, $V^{-1} = V^T$).

Another commonly used format to write SVD is

$$A = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^T,$$

where σ_i is the i-th element on the diagonal of Σ (a.k.a. the i-th singular value), u_i and v_i are the i-th column of matrix U and V, $\forall i$, respectively.

Charles Spearman was a psychologist who postulated that there are essentially two types of human intelligence: quantitative and verbal. In his experiments, he collected the performance of a thousand students on ten types of tests and put the results into a 1000×10 matrix M. He took the best rank two approximation (in the sense of minimizing the Frobenius norm) of M and found vectors $u_1, u_2 \in \mathbb{R}^{1000}$ and $v_1, v_2 \in \mathbb{R}^{10}$ such that

$$\tilde{M} = u_1 v_1^T + u_2 v_2^T = \begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix} \begin{bmatrix} | & | \\ v_1 & v_2 \\ | & | \end{bmatrix}^T \tag{1}$$

By the Eckart-Young theorem, this low rank approximation $\tilde{M}$ is unique if the second largest singular value is not equal to the third largest one. However, the vectors u_1, u_2, v_1, and v_2 are generally not unique. We take a detailed look into this approach via the following example.

Example 1.4. *Assume we have three students: Alice, Bob, and Carol; three tests: classics, math, and music. We put their performance on each test into the following matrix.*

	Classics	Math	Music
Alice	19	26	17
Bob	8	17	9
Carol	7	12	7

By writing the score matrix as in Eq. (1)*, we can see that there are multiple options:*

$$\begin{bmatrix} 19 & 26 & 17 \\ 8 & 17 & 9 \\ 7 & 12 & 7 \end{bmatrix} = \begin{bmatrix} 4 & 3 \\ 3 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 5 & 2 \\ 5 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 5 & 2 \\ 6 & 7 & 5 \end{bmatrix}$$

Although the uniqueness of low rank approximation can be guaranteed under certain conditions per the Eckart-Young theorem, how to decompose the low rank approximation is generally not unique. One reason for this is the "rotation problem". Consider $M = UV^T$ as a decomposition of M. Then for any orthogonal matrix O, we have that

$$M = UV^T = UOO^TV^T$$

is also a valid decomposition of M. Another reason for this ambiguity comes from the fact that the singular vectors are not uniquely determined if there exist multiplet singular values.

To overcome this ambiguity, one can collect the performance of students during day and night, respectively, assuming that the time of the day plays a role in students' performance.

Example 1.5. *Assume the performance matrix in Example 1.4 is obtained in the day. In the night, we obtain another score matrix as following.*

	Classics	Math	Music
Alice	23	46	25
Bob	11	32	15
Carol	9	22	11

By stacking this score matrix onto the matrix in Example 1.4, we get a tensor T in $\mathbb{R}^{3\times3\times2}$, where the third dimension represents "day" or "night".

We can get a decomposition of T is

$$\begin{bmatrix} 4 \\ 3 \\ 2 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 5 \\ 2 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 5 \\ 2 \\ 3 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{2}$$

By a little abuse of the notation "$\otimes$"[1], we write it more compactly:

$$\begin{bmatrix} 4 & 3 \\ 3 & 1 \\ 2 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 5 \\ 5 & 2 \\ 2 & 3 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}.$$

We can see the first two matrices are the first decomposition in Example 1.4. If we put the second decomposition in Example 1.4 as the first two matrices, we will find that there exists no valid third matrix that decomposes T. This is by no coincidence because the uniqueness of tensor decomposition is guaranteed under certain conditions, which we will discuss next.

2 Tensor Decomposition

Definition 2.1. A rank one tensor is a tensor of the form $T = x \otimes y \otimes w$, where x, y and w are non-zero vectors.

[1] It is an abuse because the tensor product of two $\mathbb{R}^{3\times2}$ matrices is in $\mathbb{R}^{9\times4}$

Definition 2.2. The rank of a tensor T is the minimum r such that we can write T as the sum of r rank one tensors, i.e.,

$$T = \sum_{i=1}^{r} x_i \otimes y_i \otimes w_i.$$

And we call the identification of $\{x_i, y_i, w_i\}_{i=1}^{r}$ the problem of "tensor decomposition". The following theorem gives a sufficient condition under which there exists a unique decomposition of tensor T.

Theorem 2.3. *[2] If there exists a decomposition of tensor T such that*

$$T = \sum_{i=1}^{r} x_i \otimes y_i \otimes w_i,$$

where $\{x_i\}$ is linearly independent, $\{y_i\}$ is linearly independent, and no pairs of $\{w_i\}$ are colinear, then the above decomposition is unique up to scaling.

Using this theorem, we can see that the decomposition in Eq. (2) is indeed unique. The implication of this theorem is that if we manage to find one decomposition of tensor T, then we only need to check the vectors and make sure this is the unique decomposition. However, an algorithm is yet to be invented to obtain at least one decomposition.

2.1 Jenrich's Algorithm

Albeit the advantages brought by tensor methods (as seen in Example 1.5), there are obstacles as well when working on tensors. As a result, classic decomposition algorithms for matrices do not hold for tensors. For example, for a matrix A, subtracting the best rank one approximation $\tilde{A}$ from A will lead to a lower rank in $A - \tilde{A}$, so that we can remove the best rank one approximation step-by-step from A. But subtracting the best rank one approximation from a tensor T can actually increase the rank of T.

In addition to the issue above, a more worrisome issue in fact is that the "border rank" (see definition below) of a tensor T is not necessarily equal to its rank.

Definition 2.4. The border rank of a tensor. Let $T \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k}$. The border rank of T is the minimum r such that for arbitrarily small $\epsilon > 0$, there exists $T' = \sum_{i=1}^{r} v_1^{(i)} \otimes \cdots \otimes v_k^{(i)}$ and $\|T - T'\| < \epsilon$, where $v_j^{(i)} \in \mathbb{R}^{n_j}$, $\forall i, j$.

When $k = 2$ (i.e., for matrices), the border rank is equal to the rank. But for general tensors, this property does not hold any more (no matter what norm is used above).

The Algorithm 1 is the Jenrich's algorithm for tensor decomposition under "mild" conditions [2].

Algorithm 1: Jenrich's Tensor Decomposition Algorithm

```
  Input: Tensor $T \in \mathbb{R}^{m \times n \times p}$ satisfying the conditions in Theorem 2.3. In
         addition, its rank $r \leq \min(m, n)$.
1 Randomly choose unit-length vectors $a, b \in \mathbb{R}^p$
2 $T_a \leftarrow T(*, *, a) = \sum_{j=1}^{p} a_j T_{\cdot,\cdot,j} \in \mathbb{R}^{m \times n}$
3 $T_b \leftarrow T(*, *, b) = \sum_{j=1}^{p} b_j T_{\cdot,\cdot,j} \in \mathbb{R}^{m \times n}$
4 Perform the eigendecomposition of $T_a T_b^{\dagger} = \sum_{i=1}^{r} \lambda_i x_i x_i^T$ and
   $T_b T_a^{\dagger} = \sum_{i=1}^{r} \alpha_i y_i y_i^T$
5 Pair up $x_i$ and $y_j$ if $\lambda_i = 1/\alpha_j$
6 Solve for $\{w_i\}$ in the linear equation system $T = \sum_{i=1}^{r} x_i \otimes y_i \otimes w_i$
7 return $\{x_i, y_i, w_i\}_{i=1}^{r}$
```

Note 2.5. *In Algorithm 1, $T_a = \sum_{i=1}^{r} (w_i^T a) x_i y_i^T$, $T_b = \sum_{i=1}^{r} (w_i^T b) x_i y_i^T$. Let X and Y be the matrices formed by $\{x_i\}$ and $\{y_i\}$. Let $D_a = diag(w_1^T a, w_2^T a, \ldots, w_r^T a)$ and $D_b = diag(w_1^T b, w_2^T b, \ldots, w_r^T b)$. Then we can represent T_a and T_b compactly:*

$$T_a = X D_a Y^T, \ T_b = X D_b Y^T.$$

Then it is straightforward to see that $T_a T_b^{\dagger} = X D_a D_b^{\dagger} X^T$ and $T_b T_a^{\dagger} = X D_b D_a^{\dagger} X^T$. This fact explains why on Line 4 we know $T_a T_b^{\dagger}$ and $T_b T_a^{\dagger}$ are diagonalizable and have the same rank r, and also explains why on Line 5 it is guaranteed that each x_i has a y_j to pair up.

Since we know T satisfies the conditions in Theorem 2.3, there exists at least one solution to the equation on Line 6. It is yet to see if this solution is unique. This requires the introduction of Khatri-Rao product.

Definition 2.6. The Khatri-Rao product $\otimes_{KR}$ between two matrices $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ is in $\mathbb{R}^{mn \times r}$, where its i-th column is

$$(U \otimes_{KR} V)_i = u_i \otimes v_i.$$

Using the Khatri-Rao product, we can rewrite the linear equation system on Line 6 as $T = (X \otimes_{KR} Y)W$, where $W = [w_1^T \ldots w_r^T]^T$. So the problem now is whether $X \otimes_{KR} Y$ is full rank. The following theorem provides such a guarantee.

Theorem 2.7. *If rank(U)=rank(V)=$r \leq m + n - 1$, then rank($U \otimes_{KR} V$)=r.*

Since X and Y are both formed by eigenvectors and thus have full column rank, by the above theorem it is obvious that Line 6 yields only one unique solution W.

Note in Algorithm 1, we have required that the rank of tensor T be no larger than m and n. This is because $T_a T_b^{\dagger}$ and $T_b T_a^{\dagger}$ are both $\mathbb{R}^{m \times n}$ matrices and we need to perform eigendecomposition on them. There exist extensions that handle the case when r is larger than any of the dimensions of its factors [2].

3 Mixture of Gaussians: A Tensor Approach

The method of moments, as a classical parameter estimation technique, sees the difficulty of learning latent variable models (especially high-dimensional ones) since it may involve solving computationally intractable polynomial equations. By far, the most popular heuristic method is the Expectation-Maximization (EM) algorithm. However, EM may suffer from slow convergence and poor quality of local optima [1]. In this section, we talk about how to use the tensor decomposition introduced above to approach the learning problem of latent variable models, using the mixture of Gaussians as an example. For other latent variable models, similar tensor structures can also be applied. See [1, 2].

Consider a mixture of k Gaussian distributions. Let $w_i \in (0,1)$ be the probability of choosing the i-th Gaussian distribution and the mean of it is $\mu_i \in \mathbb{R}^d$, $i = 1, 2, \dots, k$. Then

$$x := \mu_h + z, \text{ w.p. } w_h, \text{for } h = 1, 2, \dots, k,$$

where "w.p." stands for "with probability" and $z|h \sim \mathcal{N}(0, \sigma_h^2 I_d)$, where I_d is the d-th order identity matrix. Obviously, x is d dimensional as well. The following theorem gives a way to learning $\{\mu_h, \sigma_h, w_h\}$.

Theorem 3.1. *[4] Assume $w_i > 0$, $\forall i$ and the matrix $A = [\mu_1|\mu_2|\dots|\mu_k]$ has full column rank (this also implies that $d \geq k$). Denote the average variance by $\bar{\sigma}^2 = \sum_{i=1}^k w_i \sigma_i^2$. Let $v \in \mathbb{R}^d$ be any unit length eigenvector corresponding to the eigenvalue of $\bar{\sigma}^2$. Define*

$$\begin{aligned} M_1 &= \mathbb{E}\left[x(v^T(x - \mathbb{E}[x]))^2\right] \in \mathbb{R}^d, \\ M_2 &= \mathbb{E}[x \otimes x] - \bar{\sigma}^2 I_d \in \mathbb{R}^{d \times d}, \\ M_3 &= \mathbb{E}[x \otimes x \otimes x] - \sum_{i=1}^d (M_1 \otimes e_i \otimes e_i + e_i \otimes M_1 \otimes e_i + e_i \otimes e_i \otimes M_1), \end{aligned}$$

where $e_i \in \mathbb{R}^d$ is the basis vector with only the i-th entry being 1 and others being 0. Then

$$M_1 = \sum_{i=1}^k w_i \sigma_i^2 \mu_i, M_2 = \sum_{i=1}^k w_i \mu_i \otimes \mu_i, M_3 = \sum_{i=1}^k w_i \mu_i \otimes \mu_i \otimes \mu_i.$$

Proof. First, $\bar{\sigma}^2$ is the an eigenvalue of the covariance matrix

$$\mathbb{E}\left[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T\right],$$

and $v^T(\mu_i - \bar{\mu}) = 0$ for all i. To see this, let $\bar{\mu} = \mathbb{E}[x] = \sum_i w_i \mu_i$. Then the covariance matrix is

$$\mathbb{E}[(x - \bar{\mu}) \otimes (x - \bar{\mu})] = \sum_{i=1}^k w_i (\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu}) + \bar{\sigma}^2 I_d.$$

Noticing that the vectors $\mu_i - \bar{\mu}$ are linearly dependent because $\sum_i w_i(\mu_i - \bar{\mu}) = 0$, we know that the semidefinite matrix $\sum_{i=1}^k w_i(\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu})$ is not full rank (its smallest eigenvalue is thus zero). Because of the existence of $\bar{\sigma}^2 I_d$, the smallest eigenvalue of $\mathbb{E}\left[(x - \bar{\mu}) \otimes (x - \bar{\mu})\right]$ is $\bar{\sigma}^2$. Its eigenvectors associated with $\bar{\sigma}^2$ (i.e., v) are also the eigenvectors of $\sum_{i=1}^k w_i(\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu})$ associated with 0. Since $\sum_{i=1}^k w_i(\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu})$ is obviously non-zero (otherwise, x is simply a constant), all of its eigenvectors associated with non-zero eigenvalues must be orthogonal to v. Hence, v is in the null space of $\sum_{i=1}^k w_i(\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu})$ and $v^T(\mu_i - \bar{\mu}) = 0$.

Then using this fact, we have

$$\begin{aligned} M_1 &= \mathbb{E}\left[x(v^T(x - \mathbb{E}[x]))^2\right] = \mathbb{E}\left[(\mu_h + z)(v^T(\mu_h - \bar{\mu} + z))^2\right] \\ &= \mathbb{E}\left[(\mu_h + z)(v^T z)^2\right] = \mathbb{E}\left[\mathbb{E}\left[(\mu_h + z)(v^T z)^2 | h\right]\right] = \mathbb{E}\left[\mu_h \sigma_h^2\right] \\ &= \sum_{i=1}^k w_i \sigma_i^2 \mu_i. \end{aligned}$$

Next, since $\mathbb{E}[z \otimes z] = \sum_{i=1}^k w_i \sigma_i^2 I_d = \bar{\sigma}^2 I_d$, we have

$$\begin{aligned} M_2 &= \mathbb{E}[x \otimes x] - \bar{\sigma}^2 I_d = \mathbb{E}[\mu_h \otimes \mu_h] + \mathbb{E}[z \otimes z] - \bar{\sigma}^2 I_d = \mathbb{E}[\mu_h \otimes \mu_h] \\ &= \sum_{i=1}^k w_i \mu_i \otimes \mu_i. \end{aligned}$$

Finally for M_3, we observe that

$$\begin{aligned} \mathbb{E}[x \otimes x \otimes x] &= \mathbb{E}[\mu_h \otimes \mu_h \otimes \mu_h] + \mathbb{E}[\mu_h \otimes z \otimes z] \\ &\quad + \mathbb{E}[z \otimes \mu_h \otimes z] + \mathbb{E}[z \otimes z \otimes \mu_h], \end{aligned}$$

where terms like $\mathbb{E}[\mu_h \otimes \mu_h \otimes z]$ and $\mathbb{E}[z \otimes z \otimes z]$ vanish because $z|h \sim \mathcal{N}(0, \sigma_h^2 I_d)$. We also observe that

$$\begin{aligned} \mathbb{E}[\mu_h \otimes z \otimes z] &= \mathbb{E}[\mathbb{E}[\mu_h \otimes z \otimes z | h]] \\ &= \mathbb{E}\left[\mathbb{E}\left[\sum_{i,j=1}^d z_i z_j \mu_h \otimes e_i \otimes e_j | h\right]\right] \\ &= \mathbb{E}\left[\sum_{i=1}^d \sigma_h^2 \mu_h \otimes e_i \otimes e_j\right] \\ &= \sum_{i=1}^d M_1 \otimes e_i \otimes e_j. \end{aligned}$$

Then

$$M_3 = \mathbb{E}[\mu_h \otimes \mu_h \otimes \mu_h]$$

obviously holds. □

Theorem 3.1 gives rise to a simple estimator that derives the desired parameters from the observable moments (see Theorem 2 in [4]).

References

[1] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2773–2832, 2014. Available at `https://arxiv.org/pdf/1210.7559.pdf`.

[2] A. Moitra, *Lecture notes*, ch. 3. 2014. Available at `http://people.csail.mit.edu/moitra/docs/bookex.pdf`.

[3] R. Ge, "Tensor methods in machine learning." blog. Available at `http://www.offconvex.org/2015/12/17/tensor-decompositions/` (retrieved on Sep 29, 2017).

[4] D. Hsu and S. M. Kakade, "Learning mixtures of spherical gaussians: moment methods and spectral decompositions," in *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pp. 11–20, ACM, 2013. Available at `https://arxiv.org/pdf/1206.5766.pdf`.