COMS 4995: Unsupervised Learning (Summer'18) Jun 21 & 25, 2018

Lecture 10 & 11 – Tensor Decomposition

Instructor: *Geelon So* Scribes: *Wenbo Gao, Xuefeng Hu*

1 Method of Moments

Let X be data we observe generated by model with θ.

1. $f(X)$ is a function that measures something about the data.
2. From our data, we can form an empirical estimate: $\hat{\mathbb{E}}[f(X)]$
3. Then, we solve an inverse problem — which θ satisfied: $\mathbb{E}_\theta[f(X)] = \hat{\mathbb{E}}[f(X)]$.

   This yields estimate θ of the model parameter.

2 Concerns

1. **Identifiability**: is determining the true parameters θ possible?
2. **Consistency**: will our estimate $\hat{\theta}$ converge to the true θ?
3. **Complexity**: how many samples? How much time? (for ϵ, δ)
4. **Bias**: How off is the model's best?

3 Tensor Decompositions in Parameter Estimation

High level:

- Construct $f(X)$ a tensor-valued function.
  - Tensors have 'rigid' structure, so identifiability becomes easier.
- There are efficient algorithms to decompose tensors.
  - This allows us to retrieve model parameters.

4 Motivating Example I: Factor Analysis

Problem: Given A only, can we deduce k, B, and C?
Rotation Problem: If B and C are solutions, and $R \in \mathrm{GL}(k, \mathbb{R})$: then so are BR^{-1} and RC. Thus, B and C are not unique (and so not identifiable).

5 Motivating Example II: Topic Modeling

Notation: define the 3-way array M to be:

$$M_{ijk} = \mathbb{P}[x_1 = i, x_2 = j, x_3 = k] = \sum_{h=1}^{t} w_h P_i^h P_j^h P_k^h$$

6 Motivating Examples: Comparison

Problem I

$$A_{rs} = \sum_{i=1}^{k} B_{ri} C_{is}$$

- $[A_{rs}]$ is an $n \times m$ matrix.
- Fixing i, $[B_{ri}C_{is}]$ is a $n \times m$ matrix with rank 1.

7 Outline

- Coordinate-free linear algebra
- Multilinear algebra and tensors
- SVD and low-rank approximations
- Tensor decompositions
- Latent variable models

8 Dual Vector Space

Definition 1. *Let V be a finite-dimensional vector space over $\mathbb{R}$. The dual vector space V^* is the space of all real-valued linear functions $f : V \to \mathbb{R}$.*

9 Vector Space and its Dual

How should we make sense of V and V^*?

- V is the space of *objects* or *states*
  - the dimension of V is how many degrees of freedom / ways for objects to be different

Example 2 (Traits)**.** *Let V be the space of personality traits of an individual.*

- *Perhaps, secretly, we know that there are k independent traits, so $V = span(e_1, \ldots, e_k)$*
- *We can design tests $e^1, \ldots, e^k$ that measures how much an individual has those traits:*

$$e^i(e_j) = \delta_{ij}$$

Say Alice has personality trait $v \in V$. Then, her ith trait has magnitude:

$$\alpha^i := e^i(v)$$

which is a scalar in $\mathbb{R}$.

- *Since $v = \sum_i \alpha^i e_i$, we can represent her personality in coordinates with respect to the basis e_i by a 1D array*

$$[v] = \begin{bmatrix} \alpha^1 \\ \vdots \\ \alpha^k \end{bmatrix}.$$

On the other hand, say we have a personality test $f \in V^*$.

- *The amount that f tests for the ith trait is:*

$$\beta_i := f(e^i),$$

which is a scalar.

The score Alice gets on the test f is then:

$$f(v) = \begin{bmatrix} \beta_1 \cdots \beta_k \end{bmatrix} \begin{bmatrix} \alpha^1 \\ \vdots \\ \alpha^k \end{bmatrix} = \sum_{i=1}^{k} \alpha^i \beta_i.$$

Let's introduce a machine $T : V \to V$ that takes in a person and purges them of all personality except for the first trait, e_1.

- *i.e. T projects $v \in V$ onto e_1.*

Thus, given $v \in V$ the machine T:

1. *measures the magnitude of trait e_1 using $e^1 \in V^*$*
2. *outputs $e^1(v)$ attached to $e_1 \in V$:*

$$T(v) = e_1 \otimes e^1(v)$$

where we informally use $\otimes$ to mean 'attach'.

Naturally, we say that $T = e_1 \otimes e^1$.

The matrix representation of $T = e_1 \otimes e^1$ is:

$$[T] = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & & \\ \vdots & & \ddots & \\ 0 & & & 0 \end{bmatrix}.$$

The first row of $[T]$ determines what $[Tv]_1$ is; indeed the first row is the dual vector e^1.

10 Vector Space and its Dual: payoff, prelude

When we first learned linear algebra, we may have mentally substituted any (finite-dimensional) abstract vector space V by some $\mathbb{R}^n$.

- The price was coordinates, $[v] = \sum_i \alpha^i e_i$.
- And real-valued linear map as $1 \times n$ matrix (more numbers).

However, if we begin to work with more complicated spaces and maps, coordinates might reduce clarity.

- For now, just understand that V is a space of objects, while V^* is a space of devices that make linear measurements.
- These are dual objects, and there is a natural way we can apply two dual objects to each other.

11 Linear Transformations

More generally, let $T : V \to V$ be a linear transformation:

$$T : V \to V = \mathbb{R}e_1 \oplus \cdots \oplus \mathbb{R}e_n,$$

so we can decompose T into n maps, $T^i : V \to \mathbb{R}e_i$.

- But notice that $\mathbb{R}e_i$ is isomorphic to $\mathbb{R}$.
- So really, T^i is a *measurement* in V^* (it produces a scalar), but we've attached output to the vector e_i:
$$e_i \otimes T^i$$
- Recomposing T, we get:
$$T = \sum_{i=1}^{n} e_i \otimes T^i.$$

Relying on how we usually use matrices, the ith row of $[T]$ gives the coordinate representation of the dual vector $T^i \in V^*$ that we then attach to e_i.

Definition 3. *Let $V \otimes V^*$ be the vector space of all linear maps $T : V \to V$.*

- *Objects in $V \otimes V^*$ are linear combination of $v \otimes f$, where $v \in V$ and $f \in V^*$.*
- *The action of $(v \otimes f)$ on a vector $u \in V$ is:*

$$(v \otimes f)(u) = v \otimes f(u) = f(u) \cdot u.$$

12 Other views

Stepping back a bit, we have objects $v \in V$ and dual objects $f \in V^*$. We stuck them together producing $v \otimes f$. It is:

- a linear map $V \to V$
- a linear map $V^* \to V^*$, with $g \mapsto g(v) \cdot f$
- a linear map $V^* \times V \to \mathbb{R}$, with $(g, u) \mapsto g(v) \cdot f(u)$

13 Wire Diagram

14 Coordinate-Free Objects

Importantly, our definition of V, V^* and $V \otimes V^*$ are *coordinate-free* and do not depend on a basis. Thus, each has 'physical reality' outside of a basis:

- object
- measuring-device
- object-attached-to-measuring-device

15 Tensors

"God created the matrix. The Devil created the tensor." — G. Ottaviani [O2014]

16 Tensors: definitions

1. coordinate-free
2. coordinate
3. formal
4. multilinear

17 The Matrix: physical picture

We can describe a matrix as this object in $V \otimes V^*$:

18 Tensor Product: coordinate definition

The tensor product of $\mathbb{R}^n$ and $\mathbb{R}^m$ is the space

$$\mathbb{R}^n \otimes \mathbb{R}^m = \mathbb{R}^{n\times m}.$$

If $e_1, \ldots, e_n$ and $f_1, \ldots, f_m$ are their bases, then

$$e_i \otimes f_j$$

form a basis on $\mathbb{R}^n \otimes \mathbb{R}^m$.

We think of an element of $\mathbb{R}^n \otimes \mathbb{R}^m$ as an array of size $n \times m$. Given any $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$, their tensor product is:

$$(u \otimes v)_{ij} = u_i v_j,$$

coinciding with the usual outer product uv^T.

19 Tensor Product: formal definition

Definition 4. *Let V and W be vector spaces. The tensor product $V \otimes W$ is the vector space generated over elements of the form $v \otimes w$ modulo the equivalence:*

$$(\lambda v) \otimes w = \lambda(v \otimes w) = v \otimes (\lambda w)$$

$$(v_1 + v_2) \otimes w = v_1 \otimes w + v_2 \otimes w$$

$$v \otimes (w_1 + w_2) = v \otimes w_1 + v \otimes w_2,$$

where $\lambda \in \mathbb{R}$ and $v, v_1, v_2 \in V$ and $w, w_1, w_2 \in W$.

A general element of $V \otimes W$ is of the form (nonuniquely):

$$\sum_{i=1}^{\ell} \lambda_i v_i \otimes w_i,$$

where $\lambda_i \in \mathbb{R}$ and $v_i \in V$ and $w_i \in W$.

Definition 5 (basis)**.** *Let $v_1, \ldots, v_n \in V$ and $w_1, \ldots, w_m \in W$ be bases. Then, the elements of the form*

$$v_i \otimes w_j$$

form a basis for $V \otimes W$, where $1 \le i \le n$ and $1 \le j \le m$.

Definition 6. *If $V_1, \ldots, V_n$ are vector spaces, then $V_1 \otimes \cdots \otimes V_n$ is the vector space generated by taking the iterated tensor product*

$$V_1 \otimes \cdots \otimes V_n := ((V_1 \otimes V_2) \otimes V_3) \otimes \cdots V_n).$$

- *We say that a tensor in this tensor product space has order n.*

20 Tensor Product: coordinate picture

We arrive back to the picture of the n-dimensional array of coordinates. For example, here $T \in U \otimes V \otimes W$ is:

$$T = \sum_{i,j,j} T_{ijk} u_i \otimes v_j \otimes w_k.$$

21 Multilinear Function

Definition 7. *Let $V_1, \ldots, V_n, W$ be vector spaces. A map $A : V_1 \times \cdots \times V_n \to W$ is multilinear if it is linear in each argument.*

- That is, for all $v_k \in V_k$ and for all i,

  $$A(v_1, \ldots, v_{i-1}, \cdot, v_{i+1}, \ldots, v_n) : V_i \to W$$

  is a linear map.

Exercise 8. *If $A : V_1 \times V_2 \times V_n \to R$ is multilinear, is it linear? What is a basis of $V_1 \times ... \times V_n$ as a vector space?*

Answer Not linear, consider the following examples,

Example 9. *Let $f : \mathcal{R} \times \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ be defined by $f(x, y, z) = xyz$*

Example 10. *Let $X : V \times V^* \to V \times V^*$ be defined by $X(v, f) = v \times f$*

The above examples give demonstration that tells the multilinear map is often not a linear map. To intuitively understand the relation between a multilinear map and a linear map, consider the following examples:

Example 11. *Let $A : V_1 \times ... \times V_n \to \mathcal{R}$ to be map, say V_1 are the individual's personality traits, ..., V_n are drugs the individual has taken, and $A(v_1, ..., v_n)$ is how well the individual performs on a test, given their characteristics $v_1, ..., v_n$.*

Therefore, if A is multilinear, we have

$$A(v_1, ..., 2v_n) = 2A(v_1, ..., v_n)$$

and if A is linear, we have

$$A(v_1, ..., 2v_n) = A(v_1, ..., v_n) + A(0, ..., v_n)$$

where a linear map suggest each of its coordinates are independent, while coordinates in a multilinear map is conceptually entangled together.

22 Tensor Product: Curring, Vector Space and Contraction

Curring describes the operation to transform a multi-variable function into the compound of a series of single-variable function chained together. For example, a 2-variable function $f(x, y)$, we can define maps

$$g : x \to f_x$$

$$f_x : y \to f(x, y)$$

and therefore

$$f(x, y) = f_x(y) = g(x)(y)$$

Therefore, consider the multilinear map $A : V_1 \times ... \times V_n \to W$, we can replace the object from V_1 with the operation from V_1^*, and therefore

$$\begin{aligned} &A : V_1 \times ... \times V_n \to W \\ \equiv &A : V_1^* \otimes V_2 \times ... \times V_n \to W \end{aligned}$$

continue the above procedure we will finally have

$$\begin{aligned} &A : V_1 \times ... \times V_n \to W \\ \equiv &A : V_1^* \otimes ... \otimes V_n^* \to W \end{aligned}$$

Therefore, tensor product can be consider as a process to turn a linear map $A : V_1 \times ... \times V_n \to W$ into a multilinear map $A : V_1^* \otimes ... \otimes V_n^* \to W$ by attaches the objects $v_1 \in V_1, ..., v_n \in V_n$ together into a single object $v_1^* \otimes ... \otimes v_n^* \in V_1^*, ..., V_n^*$, where $V_1, ..., V_n$ are vector spaces, and $V_1 \otimes ... \otimes V_n$ itself can also be considered as a vector space.

Definition 12. *Consider a type (m, n) tensor $(m \geq 1, n \geq 1)$, which is an element from vector space $V \otimes ... \otimes V \otimes V^* \otimes ... \otimes V^*$, which includes m times of V and n times of V^*. A (k, l) contraction is a linear operation that applying the natural pairing on k-th V factor and l-th V^* factor and yield a $(m-1, n-1)$ type tensor as the result.*

For example, consider a $(1, 1)$ tensor $f \otimes v \in V^* \otimes V$, the $(1, 1)$ contraction would be an linear operation $C : V^* \otimes V \to k$, where k is the field of the natural pairing result, and in most cases the natural pairing will be corresponding to the bilinear form $\langle f, v\rangle = f(v)$ and k will just be $\mathcal{R}$. Notice that $V^* \otimes V$ actually corresponding to the matrix space $V \times V$, and the contraction will be corresponding to the trace operation in this case.

23 Tensor Decomposition

Notations Let $V^{\otimes d}$ denotes the tensor space $V \otimes ... \otimes V$ (d times), and let $v^{\otimes d}$ denotes the elements from $V^{\otimes d}$

Definition 13. *A tensor $T \in V_1 \otimes ... \otimes V_n$ is decomposable or pure if there are vectors $v_1 \in V_1, ..., v_n \in V_n$ such that:*

$$T = v_1 \otimes ... \otimes v_n$$

For example, let $M \in V \otimes V^*$ is decomposable, we have $M = v \otimes f$.

Problem	**Complexity**
Bivariate Matrix Functions over $\mathbb{R}$, $\mathbb{C}$	Undecidable (Proposition 12.2)
Bilinear System over $\mathbb{R}$, $\mathbb{C}$	NP-hard (Theorems 2.6, 3.7, 3.8)
Eigenvalue over $\mathbb{R}$	NP-hard (Theorem 1.3)
Approximating Eigenvector over $\mathbb{R}$	NP-hard (Theorem 1.5)
Symmetric Eigenvalue over $\mathbb{R}$	NP-hard (Theorem 9.3)
Approximating Symmetric Eigenvalue over $\mathbb{R}$	NP-hard (Theorem 9.6)
Singular Value over $\mathbb{R}$, $\mathbb{C}$	NP-hard (Theorem 1.7)
Symmetric Singular Value over $\mathbb{R}$	NP-hard (Theorem 10.2)
Approximating Singular Vector over $\mathbb{R}$, $\mathbb{C}$	NP-hard (Theorem 6.3)
Spectral Norm over $\mathbb{R}$	NP-hard (Theorem 1.10)
Symmetric Spectral Norm over $\mathbb{R}$	NP-hard (Theorem 10.2)
Approximating Spectral Norm over $\mathbb{R}$	NP-hard (Theorem 1.11)
Nonnegative Definiteness	NP-hard (Theorem 11.2)
Best Rank-1 Approximation	NP-hard (Theorem 1.13)
Best Symmetric Rank-1 Approximation	NP-hard (Theorem 10.2)
Rank over $\mathbb{R}$ or $\mathbb{C}$	NP-hard (Theorem 8.2)
Enumerating Eigenvectors over $\mathbb{R}$	#P-hard (Corollary 1.16)
Combinatorial Hyperdeterminant	NP-, #P-, VNP-hard (Theorems 4.1 , 4.2, Corollary 4.3)
Geometric Hyperdeterminant	Conjectures 1.9, 13.1
Symmetric Rank	Conjecture 13.2
Bilinear Programming	Conjecture 13.4
Bilinear Least Squares	Conjecture 13.5

Note: Except for positive definiteness and the combinatorial hyperdeterminant, which apply to 4-tensors, all problems refer to the 3-tensor case.

Figure 1: "Most tensor problems are NP-hard", Hillar, Lim, [H2013]

Exercise 14. *Describe the action of $M \in V \to V$. What is its rank? What would its singular value decomposition look like?*

Physically, it is a 'machine' that is sensitive to one direction, and spits out a vector also only in one direction. Therefore the rank of M is 1. However, what if $M = \sum_i v_i \otimes f^i$? Now we can define the rank for tensors as follows,

Definition 15. *The rank of a tensor $T \in V_1 \otimes ... \otimes V_n$ is the minimum number r such that T is a sum of r decomposable tensors:*

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

The tensor rank coincides with the matrix rank. However, intuition from matrices don't carry over to tensors.

- row rank = column rank is generally false for tensors.
- rank $\leq$ minimum dimension is also false.
- For general n dimensional tensor, computing the rank of the tensor is NP-hard as shown in Figure 1.

with the help of *rank* for tensors, now we can take a look at the singular value decomposition (SVD) for tensors. Since we want to begin talking about SVD, we need a notion of inner product on our space.

24 Choice of Basis and Inner Product

Consider is a finite-dimensional vector space V, a choice of basis $e_1, ..., e_n \in V$ induces a set of basis $e^1, ..., e^n \in V^*$, and also the inner product (and norm) on V and V^*:

$$\langle u, v\rangle_V = [u]^T[v]$$

$$\langle f, g\rangle_{V^*} = [f][g]^T$$

where the $[u]^T[v]$ and $[f][g]^T$ means their coordinates with respect to the chosen standard basis.

Therefore, a choice of basis is (essentially) equivalent to a choice of inner product. In the following, we can identify V , V^*, and $\mathcal{R}^n$.

25 SVD

Theorem 16 (SVD, coordinate)**.** *Any real $m \times n$ matrix has the SVD*

$$A = U\Sigma V^\top$$

where U and $V^\top$ are orthogonal, and $\Sigma = \mathrm{Diag}(\sigma_1, \sigma_2 ...)$, with $\sigma_1 \geq \sigma_2 \geq ... > 0$

For simplicity, we'll state the version for $A \in V \otimes V^*$, where adjoints are implicit due to the identication of V with V^* (from the choice of basis).

Theorem 17 (SVD, coordinate-free)**.** *Let $A \in V \otimes V^*$. Then there is a decomposition (SVD)*

$$A = \sum_{i=1}^{k} \sigma_i \left(v_i \otimes f^i\right)$$

where $\sigma_1 \geq \sigma_2 \geq ... > 0$ such that the v_i's are unit vectors and pairwise orthogonal, and similarly for the f_i's.

Similar to what we have in PCA, SVD has a geometric intuition.

Theorem 18 (SVD, geometric)**.** *Let $A \in \mathbb{R}^{m\times n}$, and let $U\Sigma V^\top$ be its SVD, where $\Sigma = \Sigma_1 + ... + \Sigma_k$ (with $\sigma_1 \geq \sigma_2 \geq ... \geq 0$), then $U\Sigma_1 V^\top$ is the best rank-1 approximation of A:*

$$\|A - U\Sigma_1 V^\top\|_F \leq \|A - X\|_F$$

where X is any rank-1 matrix in $\mathbb{R}^{m\times n}$.

Therefore, we can iteratively generate $U\Sigma_{i+1}V^\top$ by finding the best rank-1 approximation of A after being deflated of its first i singular values:

$$A - \sum_{j=1}^{i} U\Sigma_j V^\top$$

However, a key problem in this process is how do you determine whether the rank of the tensor is less than k? We first take a look at several ways to determine the rank of matrices,

- Determinants of $k \times k$ minors.
- The determinant is a polynomial equation over the $e_i \otimes f^j$'s.
- The subset of $m \times n$ matrices:

$$\mathcal{M}_k = \{m \times n \text{ matrices of rank } \leq k\}$$

  is the zero set of some set of polynomial equations.

Note that the $\mathcal{M}_k$'s contain each other:

$$0 = \mathcal{M}_0 \subset \mathcal{M}_1 \subset ... \subset \mathcal{M}_{\min(m,n)} = \mathbb{R}^{m \times n}$$

a following result gives relation between SVD and ranks,

Theorem 19 (Eckart-Young)**.** *Let $A = U\Sigma V^\top$ be the SVD and $1 \leq\leq rank(A)$. Then, All critical points of the distance function from A to the (smooth) variety $\mathcal{M}_r \backslash \mathcal{M}_{r-1}$ are given by:*

$$U\left(\Sigma_{i1} + ... + \Sigma_{ir}\right) V^\top$$

where $1 \leq i_p \leq rank(A)$. If the nonzero singular values of A are distinct, then the number of critical points is $\binom{rank(A)}{k}$.

For tensors, now we see use a tensor style notation to see the SVD of matrix $A \in \mathbb{R}^{m \times n}$

$$A = \dot{\Sigma}(U, V)$$

where $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal, and $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are unitary. Similarly, we have the *Tucker decomposition* for $A \in \mathbb{R}^{n_1 \times ... \times n_p}$:

$$A = \dot{\Sigma}(U_1, ..., U_p)$$

where $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal, $U_1 \in \mathbb{R}^{n_1 \times n_1}$,..., $U_p \in \mathbb{R}^{n_p \times n_p}$ are unitary. However, several problems occurs when we want to extend the best rank r approximation from matrices to tensors:

- The set of rank k tensors $\mathcal{M}_k$ may not be a closed set, so minimizer might not exist.
- The best rank-1 tensor may have nothing to do with the best rank-k tensor.
- Deflating by the best rank-1 tensor may increase the rank.

To get rid of those problems, a border Rank definition is suggested:

Definition 20. *The border rank $\underline{R}(T)$ of a tensor T is the minimum r such that T is the limit of tensors of rank r. If $\underline{R}(T) = R(T)$, we say that T is an open boundary tensor (OBT).*

While no direct analog of SVD theorem is possible on tensors, there are a few generalizations. We can relax Tucker's criteria:

- Higher-order SVD: Σ no longer has to be diagonal.
- CP decomposition: U, V, W no longer need to be orthonormal.

26 Symmetric and Odeco Tensors

Now we may try to find which kind of tensors in $V^{\otimes d}$ have a 'eigendecomposition':

$$\lambda_1 v_1^{\otimes d} + ... + \lambda_k v_k^{\otimes d}$$

where v_i's form a Inspired by the Spectral Theorem for matrices,

Definition 21. *Let P_d defines a group of permutation on d objects, if $\sigma \in P_d$, it acts elements in $V^{\otimes d}$ by*

$$\sigma(v_1 \otimes ... \otimes v_d) \to v_{\sigma(1)} \otimes ... \otimes v_{\sigma(d)}$$

Definition 22. *Symmetric Subspace $S^d V$ of symmetric tensors in $V^{\otimes d}$ is the collection of tensors invariant to permutations $\sigma \in G_d$:*

$$S^d V := \{T \in V^{\otimes d} : \sigma(T) = T\}$$

Then, we define what is orthogonally decomposable for tensors:

Definition 23. *A symmetric tensor $T \in S^d V$ is orthogonally decomposable (ODECO) if it can be written as:*

$$T = \sum_{i=1}^{k} \lambda_i v_i^{\otimes d}$$

where the $v_i \in V$ form an orthonormal basis of V.

Since $S^d V$ is just set of symmetric matrices when $d = 2$, then by spectral theorem all $S^2 V$ are odeco. For $d \geq 2$, we have the following theorem

Theorem 24 (Alexander-Hirschowitz)**.** *For $d > 2$, the generic symmetric rank $\overline{R}_S$ of a tensor in $S^d \mathbb{C}^n$ is equal to:*

$$\overline{R}_S \left[\frac{1}{n} \binom{n+d-1}{d} \right],$$

except when $(d, n) \in \{(3, 5), (4, 3), (4, 4), (4, 5)\}$,where it should be increased by 1. From the theorem, we can note that the rank of a tensor over $\mathbb{C}$ lower bounds the rank of a tensor over $\mathbb{R}$.

While Odeco tensors must have rank n implies that not all of $S^d V$ are Odeco, in fact:

Lemma 25. *The dimension of the odeco variety in $S^d \mathbb{C}^n$ is $\binom{n+1}{2}$, and The dimension of $S^d \mathbb{C}^n$ is $\binom{n+d+1}{d}$*

Again, in general finding a symmetric decomposition of a symmetric tensor is NP-hard. However, the lucky news is that it is computationally efficient for odeco tensors with the power method.

27 Power Method

Definition 26. *Let $T \in S^d V$. A unit vector $v \in V$ is an eigenvector of T with eigenvalue $\lambda \in \mathbb{R}$ if:*

$$T \cdot v^{\otimes d-1} = \lambda v$$

for example, if $T = e_1^{\otimes d}$. Its eigenvectors will be those $v \in V$ such that :

$$\begin{aligned} T \cdot v^{\otimes d-1} &= (e_1 \otimes ... \otimes e_1) \cdot (v \otimes ... \otimes v) \\ &= (e_1 \cdot v)^{d-1} \otimes e_1 \\ &= e^1(v)^{d-1} e_1 \\ &= \lambda e_1 \end{aligned}$$

which implies the only eigenvector for T is e_1. Notice when $d = 2$, it becomes

$$T \cdot v = \lambda v$$

which coincide the definition of eigenvectors for matrices.

Since we can always normalize the eigenvectors by adjusting the corresponding eigenvalue, we now require the eigenvector v's to have unit length.

Definition 27. *Let $T \in S^d V$. A unit vector $v \in V$ is a robust eigenvector of T if there is a closed ball B of radius $\epsilon > 0$ centered at v such that for all $v_0 \in B$, the repeated iteration of the map:*

$$\phi := u \to \frac{T \cdot u^{\otimes d-1}}{\|T \cdot u^{\otimes d-1}\|}$$

converges to v, which implies an alternative definition of the robust eigenvectors: the attracting fixed points of ϕ.

With robust eigenvectors, we have the following results:

Theorem 28. *Suppose $T \in S^3\mathbb{R}^n$ is odeco, $T = \sum_{i=0}^{d} \lambda_i v_i^{\otimes 3}$*

- *The set of $u \in \mathbb{R}^n$ that do not converge to some v_i under repeated iteration of ϕ has measure zero.*
- *The set of robust eigenvectors of T is equal to $\{v_1, ..., v_k\}$.*

which implies the following corollary,

Corollary 29. *Suppose $T \in S^3\mathbb{R}^n$ is odeco, its decomposition is unique.*

Specifically, the robust eigenvectors of matrices $M \in S^2\mathbb{R}^n$ would just be the (normalized) eigenvectors.

By the above results, now we have the power method to estimate the robust eigenvectors.

Suppose that $u \in \mathbb{R}^n$ satisfies

$$|\lambda \langle v_1, u \rangle| > |\lambda \langle v_2, u \rangle| \geq ...$$

Denote by $\phi^{(t)}(u)$ the output of t repeated iteration of ϕ on u. We should have

$$\|v_1 - \phi^{(t)}(u)\|^2 \leq O\left(|\frac{\lambda_2 \langle v_2, u \rangle}{\lambda_1 \langle v_2, u \rangle}|^{2^t} \right)$$

which means that u converges to v_1 as a quadratic rate. (an interesting fact for $d = 2$ is the rate is linearly upper bounded by $\frac{\lambda_1}{\lambda_2}$.)

Algorithm 1: Tensor Power Method

Input: $T \in S^d\mathbb{R}^n$ an odeco tensor with $d > 2$

Set $E \leftarrow \{\}$;
repeat;
Choose random $u \in \mathbb{R}^n$;
Iterate $u \leftarrow \phi(u)$ until convergence;
Compute λ using $Tu^{\otimes d-1} = \lambda u$;
$T \leftarrow T - \lambda u^{\otimes d}$;
$E \leftarrow E \cup \{(\lambda, u)\}$;
until $T = 0$;
return E;

Algorithm 2 Robust Tensor Power Method (RTPM)

input tensor $\hat{T} \in S^3\mathbb{R}^k$, iterations L and N

for $\tau = 1$ to L **do**
    Draw u_τ uniformly at random from unit sphere S^{k-1}
    Set $u_\tau \leftarrow \phi^{(N)}(u_\tau)$.
end for
Let u_τ^* be the maximizer of $\hat{T} \cdot u_\tau^{\otimes 3}$
$\hat{u} \leftarrow \phi^N(u_\tau^*)$, $\hat{\lambda} \leftarrow \hat{T} \cdot \hat{u}^{\otimes 3}$.
return $(\hat{u}, \hat{\lambda})$ and deflated tensor $\hat{T} - \hat{\lambda}\hat{u}^{\otimes 3}$.

However, In estimating an odeco tensor T, we might produce a tensor T that is not odeco, and therefore we might need an power method to estimate the robust eigenvectors of T Where the following facts follows:

- $\hat{T} = T + E \in S^3\mathbb{R}^k$ symmetric; $T = \sum_{i=0}^k \lambda_i v_i^{\otimes 3}$ odeco.
- λ_{min} and λ_{max} the min/max λ_i's.
- $\|E\|_{op} \leq \epsilon$

and we have the theorem:

Theorem 30 (Thm. 5.1, A2014)**.** *Let $\delta \in (0,1)$, if $\epsilon = O\left(\frac{\lambda_{min}}{k}\right)$, $N = \Omega\left(\log k + \log\log\left(\frac{\lambda_{max}}{\epsilon}\right)\right)$ and $L = ploy(k)\log\left(\frac{1}{\epsilon}\right)$, running RTPM^k will yield, w.p. $1-\delta$,*

$$\|v_i, \hat{v}_i\| = O\left(\frac{\epsilon}{\lambda_i}\right)$$

$$|\lambda_i, \hat{\lambda}_i| = O(\epsilon)$$

$$\|T - \sum_{i=0}^k \hat{\lambda}_j v_j^{\otimes 3}\| \leq O(\epsilon)$$

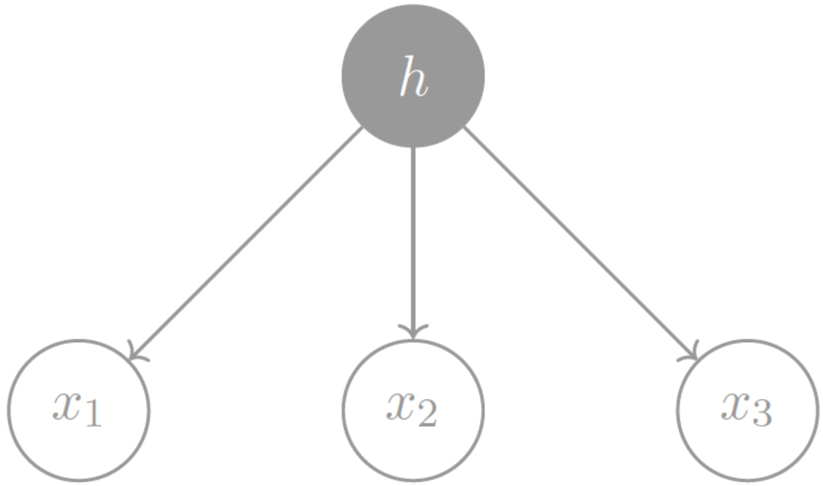

Figure 2: Topic Model

28 Back to Topic Model

Now consider back the topic model setting, where we have t topics, d-sized vocabulary and some 3 words long documents.

- topic h is chosen with probability w_h
- words x_i's are conditionally independent on topic h, according to probability distribution $P^h \in \Delta^{d-1}$, as shown in figure 2.

Therefore, we can use tensor to represent the data. From d words, $e_1, ..., e_d$ generates the vector space of all "words object"

$$V = \mathbb{R}e_1 \oplus ... \oplus \mathbb{R}e_d = \mathbb{R}^d$$

We interpret $x \in V$ as a probability vector, where the weight on the ith coordinate is the probability the word is e_i. Then, the 3 words documents space can be defined as $V^{\otimes 3}$, where

- Since we assume that the choice of 3 words in a single document is conditionally independent, this means that expectation is multilinear.
- In particular, let x_1, x_2, x_3 be the random variable for the words in a document:

$$\mathbb{E}[x_1 \otimes x_2 | h = j] = \mathbb{E}[x_1 | h = j] \otimes \mathbb{E}[x_2 | h = j] = \mu_i \otimes \mu_j$$

and we have the following result by [A2012],

Theorem 31. *If* $M_2 := \mathbb{E}[x_1 \otimes x_2]$,$M_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3]$, *then*

$$M_2 = \sum_{i=0}^{k} w_i \mu_i^{\otimes 2}$$

$$M_3 = \sum_{i=0}^{k} w_i \mu_i^{\otimes 3}$$

29 Whitening

We are almost at a point where we can use the Robust Tensor Power Method to deduce the probabilities i (i.e. the robust eigenvectors) and the weights w_i (i.e. the eigenvalues). However, we need to make sure the μ_i's are orthonormal. We can take advantage of M_2, which is just an invertible matrix, conditioned upon:

- the vectors $\mu_1, ..., \mu_k \in \mathbb{R}^d$ are linearly independent,
- the weights $w_1, ..., w_k > 0$ are strictly positive.

If the condition is satisfied, then there exists W such that:

$$M_2 \cdot (W, W) = I$$

so that setting $\overline{\mu}_i = \sqrt{w_i} W^\top \mu_i$ forms a set of orthonormal vectors. It then follows that:

$$M \cdot (W, W, W) = \sum_{i=0}^{k} \frac{1}{\sqrt{w_i}} \overline{\mu}_i^{\otimes 3}$$

Get back to the LDA model in lecture 11, define the following

- $M_1 := \mathbb{E}[x_1]$
- $M_2 := \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0+1} M_1 \otimes M_1$
- $M_3 := \mathbb{E}[x_1] - \frac{\alpha_0}{\alpha_0+2} (\mathbb{E}[x_1 \otimes x_2 \otimes M1] + ... + \mathbb{E}[M1 \otimes x_1 \otimes x_2]) + \frac{2\alpha_0^2}{(\alpha_0+1)(\alpha_0+2)} M_1^{\otimes 3}$

Therefore, by [A2012], we have

Theorem 32. *Let M_1, M_2, M_3 as above, Then:*

$$M_2 = \sum_{i=0}^{k} \frac{\alpha_i}{(\alpha_0+1)\alpha_0} \mu_i^{\otimes 2}$$

$$M_3 = \sum_{i=0}^{k} \frac{2\alpha_i}{(\alpha_0+2)(\alpha_0+1)\alpha_0} \mu_i^{\otimes 3}$$

References

[A2014] nandkumar, Animashree, et al. "Tensor decompositions for learning latent variable models." The Journal of Machine Learning Research 15.1 (2014): 2773-2832.

[C2008] omon, Pierre, et al. "Symmetric tensors and symmetric tensor rank." SIAM Journal on Matrix Analysis and Applications 30.3 (2008): 1254-1279.

[C2014] omon, Pierre. "Tensors: a brief introduction." IEEE Signal Processing Magazine 31.3 (2014): 44-53.

[D1997] el Corso, Gianna M. "Estimating an eigenvector by the power method with a random start." SIAM Journal on Matrix Analysis and Applications 18.4 (1997): 913-937.

[D2018] raisma, Jan, Giorgio Ottaviani, and Alicia Tocino. "Best rank-k approximations for tensors: generalizing Eckart,Young." Research in the Mathematical Sciences 5.2 (2018): 27.

[H2013] illar, Christopher J., and Lek-Heng Lim. "Most tensor problems are NP-hard." Journal of the ACM (JACM) 60.6 (2013): 45.

[H2017] su, Daniel. "Tensor Decompositions for Learning Latent Variable Models I , II." YouTube, uploaded by Simons Institute, 27 January 2017, link-1 link-2

[L2012] andsberg, J. M. Tensors: Geometry and Applications. American Mathematical Society, 2012.

[M1987] cCullagh, Peter. Tensor methods in statistics. Vol. 161. London: Chapman and Hall, 1987.

[M2016] oitra, Ankur. "Tensor Decompositions and their Applications." YouTube, uploaded by Centre International de Rencontres Mathematiques, 16 February 2016, link

[O2014] ttaviani, Giorgio. "Tensors: a geometric view." Simons Institute Open Lecture (2014). Video.

[O2015] ttaviani, Giorgio, and Raaella Paoletti. "A geometric perspective on the singular value decomposition." arXiv preprint arXiv:1503.07054 (2015).

[R2016] obeva, Elina. "Orthogonal decomposition of symmetric tensors." SIAM Journal on Matrix Analysis and Applications 37.1 (2016): 86-102.

[S2017] idiropoulos, Nicholas D., et al. "Tensor decomposition for signal processing and machine learning." IEEE Transactions on Signal Processing 65.13 (2017): 3551-3582.

[V2014] annieuwenhoven, Nick, et al. "On generic nonexistence of the Schmidt,Eckart,Young decomposition for complex tensors." SIAM Journal on Matrix Analysis and Applications 35.3 (2014): 886-903.

[Z2001] hang, Tong, and Gene H. Golub. "Rank-one approximation to high order tensors." SIAM Journal on Matrix Analysis and Applications 23.2 (2001): 534-550.