Tensor Decompositions

Geelon So (ags2191)

July 19, 2018

Parameter Estimation

Problem: Let θ parametrize our model for the world.

- How to determine model parameter θ using empirical data?

Method of Moments

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

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.

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:

$$\widehat{\mathbb{E}}[f(X)].$$

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:

$$\widehat{\mathbb{E}}[f(X)].$$

3. Then, we solve an inverse problem—which θ satisfies:

$$\mathbb{E}_\theta[f(X)] = \widehat{\mathbb{E}}[f(X)].$$

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:

$$\widehat{\mathbb{E}}[f(X)].$$

3. Then, we solve an inverse problem—which θ satisfies:

$$\mathbb{E}_\theta[f(X)] = \widehat{\mathbb{E}}[f(X)].$$

This yields some estimate θ of the model parameter.

Concerns

1. **Identifiability:** is determining the true parameters θ possible?

Concerns

1. **Identifiability:** is determining the true parameters θ possible?
2. **Consistency:** will our estimate $\hat{\theta}$ converge to the true θ?

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)

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?

Tensor Decompositions in Parameter Estimation

High level:

- Construct $f(X)$ a tensor-valued function.
  - Tensors have 'rigid' structure, so identifiability becomes easier.

Tensor Decompositions in Parameter Estimation

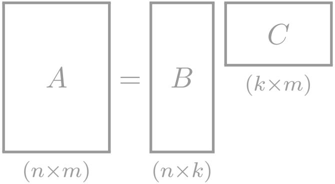
High level:

- 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.

Motivating Example I: Factor Analysis

Setup: There are n tests, k personality traits, and m students.

- each student has a linear combination of those traits
- each test is a linear function of those traits



Motivating Example I: Factor Analysis

Problem: Given A only, can we deduce k, B, and C?[1]

[1] This problem is originally due to *Spearman*, described in [M2016].

Motivating Example I: Factor Analysis

Problem: Given A only, can we deduce k, B, and C?[1]

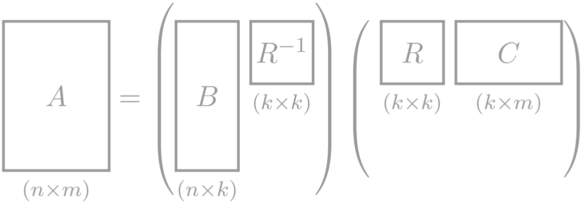
- that is, is there a unique factorization:

$$A = \sum_{i=1}^{k} B_i C_i^T$$

[1] This problem is originally due to *Spearman*, described in [M2016].

Motivating Example I: Factor Analysis

Rotation Problem: if B and C are solutions, and $R \in \mathrm{GL}(k, \mathbb{R})$:



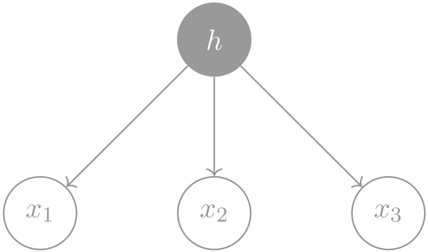
then so are BR^{-1} and RC.

- thus B and C are not unique (and so not identifiable)

Motivating Example II: Topic Modeling

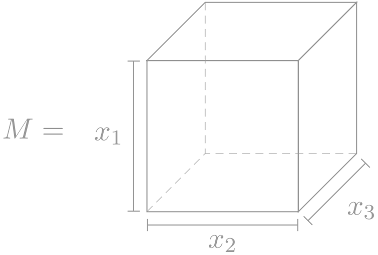
Setup: t topics, vocabulary size d, and 3-word long documents.

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



Motivating Example II: Topic Modeling

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



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

Motivating Example II: Topic Modeling

Problem: given M, can we deduce t, w_h's and P^h's?[2]

[2] This problem is presented in [H2017].

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.

Motivating Examples: Comparison

Problem II

$$M_{ijk} = \sum_{h=1}^{t} w_h P_i^h P_j^h P_k^h$$

- $[M_{ijk}]$ is an $d \times d \times d$ matrix.
- Fixing h, $[w_h P_i^h P_j^h P_k^h]$ is a $d \times d \times d$ array of 'rank' 1.

Outline

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

Coordinate-Free Linear Algebra

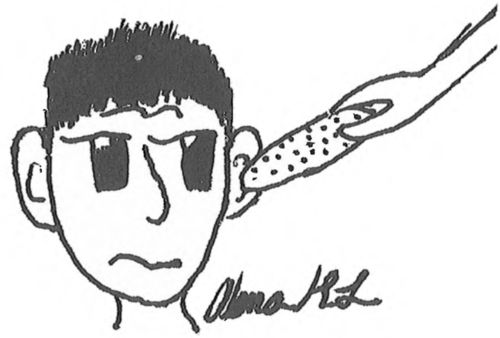


Figure 1: "Don't use coordinates unless someone holds a pickle to your head." *J. M. Landsberg* [L2012]

Dual Vector Space

Definition

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 $f : V \to \mathbb{R}$.

Dual Vector Space

Definition

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 $f : V \to \mathbb{R}$.

- *We call vectors in V^* dual vectors.*

Vector Space and its Dual

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

Vector Space and its Dual

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

- V is the space of *objects* or *states*

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

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
- V^* makes a real-valued *measurement* on an object/state

Vector Space and its Dual

Example (Traits)

Let V be the space of personality traits of an individual.

Vector Space and its Dual

Example (Traits)

Let V be the space of personality traits of an individual.

- *Perhaps, secretly, we know that there are k independent traits, so $V = \mathrm{span}(e_1, \ldots, e_k)$*

Vector Space and its Dual

Example (Traits)

Let V be the space of personality traits of an individual.

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

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

Vector Space and its Dual

Example (Traits, cont.)

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

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

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

Vector Space and its Dual

Example (Traits, cont.)

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

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

Vector Space and its Dual

Example (Traits, cont.)

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

Vector Space and its Dual

Example (Traits, cont.)

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.

Vector Space and its Dual

Example (Traits, cont.)

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.
- *It follows that the e^i's form a basis on V^*, and $f = \sum \beta_i e^i$. We can represent f in coordinates:*

  $$[f] = \begin{bmatrix} \beta_1 & \cdots & \beta_k \end{bmatrix}.$$

Vector Space and its Dual

Example (Traits, cont.)

The score Alice gets on the test f is then:

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

Vector Space and its Dual

Example (Traits, cont.)

Notice that we can define the operation $C : V^* \times V \to \mathbb{R}$

$$C(f, v) = f(v),$$

which conceptually means to 'take the measurement f *on* v*'.*

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 $\mathbb{R}^n$.

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, $[v] = \sum \alpha^i e_i$.

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

Vector Space and its Dual: payoff, prelude

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

Vector Space and its Dual: payoff, prelude

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.

Vector Space and its Dual: payoff, prelude

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.

Linear Transformations

Example (Traits, cont.)

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.

Linear Transformations

Example (Traits, cont.)

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*.*

Example (Traits, cont.)

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

Linear Transformations

Example (Traits, cont.)

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

1. *measures the magnitude of trait e_1 using $e^1 \in V^*$*

Linear Transformations

Example (Traits, cont.)

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'.*

Linear Transformations

Example (Traits, cont.)

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 $T = e_1 \otimes e^1$.

Linear Transformation

Example (Traits, cont.)

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*.*

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, $T^i : V \to \mathbb{R}e_i$.

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 $\mathbb{R}$.

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 the output to the vector e_i:

$$e_i \otimes T^i$$

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 the output to the vector e_i:

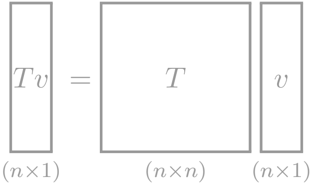
$$e_i \otimes T^i$$

- Recomposing T, we get:

$$T = \sum_{i=1}^{n} e_i \otimes T^i.$$

Linear Transformations

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.

Linear Transformations

Definition

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

Linear Transformations

Definition

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

- Objects in $V \otimes V^*$ are linear combinations of $v \otimes f$, where $v \in V$ and $f \in V^*$.

Linear Transformations

Definition

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

- Objects in $V \otimes V^*$ are linear combinations 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:

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

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:

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 $V \to V$

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 $g \mapsto g(v) \cdot f$

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

Wire Diagram

Coordinate-Free Objects

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

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

Tensors

God created the matrix.
The Devil created the tensor.

—G. Ottaviani [O2014]

Tensors: definitions

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

The Matrix: physical picture

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

Tensor Product: physical picture

Contraction: physical picture

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

Tensor Product: coordinate definition

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.

Tensor Product: formal definition

Definition

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 $w, w_1, w_2 \in W$.

Tensor Product: formal definition

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 $w_i \in W$.

Tensor Product: 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 $1 \leq i \leq n$ and $1 \leq j \leq m$.

Tensor Product: formal definition

Definition

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*[3]

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

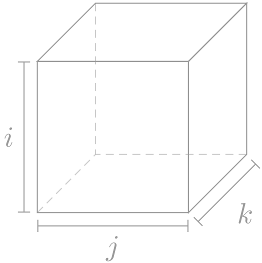
- We say that a tensor in this tensor product space has *order* n.

[3]We drop parentheses and say that $\otimes$ is associative because we can take canonical identifications between the different orders of tensor product operations (not order of the vector spaces themselves; it is not commutative).

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:

$[T] =$



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

Multilinear Function

Definition

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.

Multilinear Function

Exercise

If $A : V_1 \times \cdots \times V_n \to \mathbb{R}$ *is multilinear, is it linear? What is a basis of* $V_1 \times \cdots \times V_n$ *as a vector space?*

Multilinear Function

Example

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

Example

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

Multilinear Function: intuition

Let $A : V_1 \times \cdots \times V_n \to \mathbb{R}$ be multilinear.

Multilinear Function: intuition

Let $A : V_1 \times \cdots \times V_n \to \mathbb{R}$ be multilinear.

- Say V_1 are the individual's personality traits

  $\vdots$

  V_n are drugs the individual has taken

Multilinear Function: intuition

Let $A : V_1 \times \cdots \times V_n \to \mathbb{R}$ be multilinear.

- Say V_1 are the individual's personality traits

  $\vdots$

  V_n are drugs the individual has taken
- $A(v_1, \ldots, v_n)$ is how well the individual performs on a test, given their characteristics $(v_1, \ldots, v_n)$.

Multilinear Function: intuition

Multilinearity implies:

$$A(v_1, \ldots, 2v_n) = 2A(v_1, \ldots, v_n),$$

meaning that if Alice are on twice as many drugs, she perform twice as well/poorly.

Multilinear Function: intuition

On the other hand, if A is merely linear:

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

Here, each coordinate $v_1, \ldots, v_n$ is independent from each other.

Multilinear Function: intuition

Conceptually, a multilinear function *entangles* each of the coordinates together.

Multilinear Function: intuition

Conceptually, a multilinear function *entangles* each of the coordinates together.

- The linear function treats each coordinate independently.

Tensor Product: multilinear

Let $V_1, \ldots, V_n$ be vector spaces. The tensor product attaches the objects $(v_1, \ldots, v_n)$ together into the single:

$$v_1 \otimes \cdots \otimes v_n \in V_1 \otimes \cdots \otimes V_n$$

in such a way that any multilinear map $A : V_1 \times \cdots V_n \to W$ becomes linear $A : V_1 \otimes \cdots \otimes V_n \to W$.

Tensor Space as Vector Space

Contraction

Tensor Product: currying

Notation

Let $V^{\otimes d}$ denote the tensor space $V \otimes \overset{d \text{ times}}{\cdots} \otimes V$.

- Let $v^{\otimes d} = v \otimes \overset{d \text{ times}}{\cdots} \otimes v$ for $v \in V$.

Decomposable/Pure Tensor

Definition

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

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

Decomposable Matrix

Let $M \in V \otimes V^*$ is decomposable, so $M = v \otimes f$.

Exercise

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

Decomposable Matrix

Let $M \in V \otimes V^*$ is decomposable, so $M = v \otimes f$.

Exercise

Describe the action of $M : V \to V$. 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.

Decomposable Matrix

Let $M \in V \otimes V^*$ is decomposable, so $M = v \otimes f$.

Exercise

Describe the action of $M : V \to V$. 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.
- What if $M = \sum_i v_i \otimes f^i$?

Rank

Definition

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

$$\begin{aligned} T &= \sum_{i=1}^{r} T_i \\ &= \sum_{i=1}^{r} v_1^{(i)} \otimes \cdots \otimes v_n^{(i)}. \end{aligned}$$

Rank of Matrix

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

Rank of Matrix

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

In fact, computing the rank of a tensor is NP-hard.

Computational Complexity

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 2: "Most tensor problems are NP-hard", Hillar & Lim, [H2013]

Why do we care about rank?

We'll take a hint from singular value decomposition (SVD) for matrices.

- Since we want to begin talking about SVD, we need a notion of inner product on our space.

Choice of Basis

Remark

If V is a finite-dimensional vector space, then a choice of basis $e_1, \ldots, e_k \in V$ induces a dual basis $e^1, \ldots, e^k \in V^*$ and an inner product/norm on V and V^*:

$$\langle u, v\rangle_V := [u]^T[v] \qquad \langle f, g\rangle_{V^*} := [f][g]^T,$$

where $[u]^T[v]$ and $[f][g]^T$, we mean the standard dot product on coordinates.

Choice of Basis

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

Singular Value Decomposition

Theorem (SVD, coordinate)

Any real $m \times n$ matrix has the SVD

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

where U and V^T are orthogonal, and $\Sigma = \mathrm{Diag}(\sigma_1, \sigma_2, \dots)$, with $\sigma_1 \geq \sigma_2 \geq \cdots 0$.[4]

[4] Theorem statement from [O2015].

Singular Value Decomposition: physical version

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

Theorem (SVD, coordinate-free)

Let $A \in V \otimes V^*$. Then there is a decomposition (SVD)

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

where $\sigma_1 \geq \cdots \geq \sigma_k > 0$ such that the v_i's are unit vectors and pairwise orthogonal, and similarly for the f^i's.

Singular Value Decomposition: physical picture

Singular Value Decomposition: geometric version

Theorem (SVD, geometric)

Let $A \in \mathbb{R}^{m \times n}$, and let $U\Sigma V^t$ be its SVD, where $\Sigma = \Sigma_1 + \cdots + \Sigma_k$ (again, we assume $\sigma_1 \geq \cdots \geq \sigma_k$). Then, $U\Sigma_1 V^T$ is the best rank-1 approximation of A:

$$\left\| A - U\Sigma_1 V^T \right\|_F \leq \| A - X \|_F$$

for all matrices X of rank 1.[5]

[5] Theorem statement from [O2015].

Singular Value Decomposition: geometric version

Theorem (SVD, geometric)

Let $A \in \mathbb{R}^{m \times n}$, and let $U\Sigma V^t$ be its SVD, where $\Sigma = \Sigma_1 + \cdots + \Sigma_k$ (again, we assume $\sigma_1 \geq \cdots \geq \sigma_k$). Then, $U\Sigma_1 V^T$ is the best rank-1 approximation of A:

$$\left\| A - U\Sigma_1 V^T \right\|_F \leq \| A - X \|_F$$

for all matrices X of rank 1.[5]

[5] Theorem statement from [O2015].

Singular Value Decomposition: geometric version

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

$$A - \big(U\Sigma_1 V^T + \cdots + U\Sigma_i V^T\big).$$

Singular Value Decomposition: geometric picture

Singular Value Decomposition: geometric version

Question: How do you determine whether the rank of a matrix is less than k?

Singular Value Decomposition: geometric version

Question: How do you determine whether the rank of a matrix is less than k?

- Determinants of $k \times k$ minors.

Singular Value Decomposition: geometric version

Question: How do you determine whether the rank of a matrix is less than k?

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

Singular Value Decomposition: geometric version

Question: How do you determine whether the rank of a matrix is less than k?

- 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.

Singular Value Decomposition: geometric version

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

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

Singular Value Decomposition: geometric version

Let $A = U\Sigma V^T$ be the SVD and $1 \leq r \leq \text{rank}(A)$.

Theorem (Eckart-Young)

All critical points of the distance function from A to the (smooth) variety $\mathcal{M}_r \setminus \mathcal{M}_{r-1}$ are given by:

$$U\big(\Sigma_{i_1} + \cdots + \Sigma_{i_r}\big)V^T,$$

where $1 \leq i_p \leq \text{rank}(A)$. If the nonzero singular values of A are distinct, then the number of critical points is $\binom{\text{rank}(A)}{r}$.[6]

[6] Theorem statement from [O2015].

Singular Value Decomposition: tensor notation

Notice that SVD states that any matrix $A \in \mathbb{R}^{m \times n}$ may be decomposed into:

$$A = \Sigma \cdot (U, V),$$

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. (Keep the physical picture in mind!)

SVD for Tensors?

Let $A \in \mathbb{R}^{n_1 \times \cdots \times n_p}$ be an order-p tensor. The *Tucker decomposition* of A is:

$$A = \Sigma \cdot (U_1, \cdots, U_p),$$

where Σ is diagonal, and the U_i's are orthonormal.

Extension to Tensors

Unfortunately, the best rank-k approximation problem is *ill-posed*:

- The set of rank k tensors $\mathcal{M}_k$ may not be a closed set, so *minimizer* might not exist.[7]

[7] For example, see [V2014].

Extension to Tensors

Unfortunately, the best rank-k approximation problem is *ill-posed*:

- The set of rank k tensors $\mathcal{M}_k$ may not be a closed set, so *minimizer* might not exist.[7]
- The best rank-1 tensor may have nothing to do with the best rank-k tensor

[7] For example, see [V2014].

Extension to Tensors

Unfortunately, the best rank-k approximation problem is *ill-posed*:

- The set of rank k tensors $\mathcal{M}_k$ may not be a closed set, so *minimizer* might not exist.[7]
- 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

[7] For example, see [V2014].

Border Rank

Definition

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 $R(T) \neq \underline{R}(T)$, we say that T is an open boundary tensor (OBT).

Tensor Decompositions

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[8]

[8] CP stands either for *Canonical Polyadic* or *Candecomp/Parafac*.

What about Spectral Theorem for Symmetric Tensors?

Problem: Which tensors in $V^{\otimes d}$ have a 'eigendecomposition':

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

where the v_i's form an orthonormal basis?

Action by Symmetric Group

Definition

Let $\mathfrak{S}_d$ denote the group of permutations on d elements. If $\sigma \in \mathfrak{S}$, it acts on elements of $V^{\otimes d}$ by:

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

Symmetric Tensors

Definition

The subspace S^dV of symmetric tensors in $V^{\otimes d}$ is the collection of tensors invariant to permutations $\sigma \in \mathfrak{S}$:

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

Odeco Tensor

Definition

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*.*

Odeco Tensors: $d = 2$

If $d = 2$, then $S^d V$ are just the symmetric matrices:

- the spectral theorem says that all of $S^d V$ are odeco.

Odeco Tensors: $d > 2$

Theorem (Alexander-Hirschowitz)

For $d > 2$, the generic symmetric rank $\overline{R}_S$ of a tensor in $S^d\mathbb{C}^n$ is equal to:

$$\overline{R}_S = \left\lceil \frac{1}{n}\binom{n+d-1}{d} \right\rceil,$$

except when $(d, n) \in \{(3,5),(4,3),(4,4),(4,5)\}$, where it should be increased by 1.[9]

[9] Theorem statement from [C2008].

Odeco Tensors: $d > 2$

Theorem (Alexander-Hirschowitz)

For $d > 2$, the generic symmetric rank $\overline{R}_S$ of a tensor in $S^d\mathbb{C}^n$ is equal to:

$$\overline{R}_S = \left\lceil \frac{1}{n}\binom{n+d-1}{d} \right\rceil,$$

except when $(d, n) \in \{(3,5), (4,3), (4,4), (4,5)\}$, where it should be increased by 1.[9]

- Note that the rank of a tensor over $\mathbb{C}$ lower bounds the rank of a tensor over $\mathbb{R}$.

[9] Theorem statement from [C2008].

Odeco Tensors: $d > 2$

Rank of odeco tensor is $n \implies$ not all of S^dV are odeco. In fact...

Odeco Tensors: $d > 2$

Lemma

The dimension of the odeco variety in $S^d\mathbb{C}^n$ is $\binom{n+1}{2}$.[10]

[10] Lemma statement from [R2016].

Odeco Tensors: $d > 2$

Lemma

The dimension of the odeco variety in $S^d\mathbb{C}^n$ is $\binom{n+1}{2}$.[10]

- In contrast, the dimension of $S^d\mathbb{C}^n$ is $\binom{n+d-1}{d}$.

[10] Lemma statement from [R2016].

Symmetric Decomposition: computational complexity

Generally, finding a symmetric decomposition of a symmetric tensor is NP-hard, it is computationally efficient for odeco tensors.

Symmetric Decomposition: computational complexity

Generally, finding a symmetric decomposition of a symmetric tensor is NP-hard, it is computationally efficient for odeco tensors.

- We'll now show the *tensor power method*.

Eigenvectors of Symmetric Tensors

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

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

Eigenvectors of Symmetric Tensors

Example

Let $T = e_1^{\otimes d}$. Its eigenvectors are those $v \in V$ such that:

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

Thus, the only eigenvector of T is e_1.

Eigenvectors of Symmetric Tensors

Note that by definition, an eigenvector v must be of unit length.

Exercise

Equivalently, we could remove that restriction, and say that two eigenpairs (λ, v) *and* (λ', v') *are equivalent if there exists some* $t \neq 0$ *such that:*

$$v = tv' \qquad \lambda = t^{d-2}\lambda'.$$

Explain why.

Eigenvectors of Symmetric Tensors: $d = 2$

Remark

When $d = 2$*, then* $S^d\mathbb{R}^n$ *are just the symmetric matrices. Convince yourself that the definition of eigenvectors here coincide with the usual one.*

Robust Eigenvectors

Definition

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 $u_0 \in B$, the repeated iteration of the map:

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

converges to v.[11]

[11] Definition statement from [R2016].

Robust Eigenvectors

Definition

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 $u_0 \in B$, the repeated iteration of the map:

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

converges to v.[11]

- i.e. robust eigenvectors are *attracting fixed points* of ϕ.

[11] Definition statement from [R2016].

Convergence to Robust Eigenvectors

Theorem

Suppose $T \in S^3\mathbb{R}^n$ is odeco,[12]

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

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

[12] Theorem statement from [A2014].

Uniqueness of Decomposition

Corollary

If $T \in S^3\mathbb{R}^n$ is odeco, its decomposition is unique.

Comparison to $S^2\mathbb{R}^n$

Exercise

Let $M \in S^2\mathbb{R}^n$ be a symmetric matrix, with eigenvalues

$$\lambda_1 > \cdots > \lambda_n > 0.$$

What is the set of robust eigenvectors of M?

Tensor Power Method

Algorithm 1 Tensor Power Method

input $T \in S^d\mathbb{R}^n$ an odeco tensor, $d > 2$

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

Tensor Power Method: Analysis

Lemma (Convergence to eigenvector)

Let T as before. Suppose that $u \in \mathbb{R}^n$ satisfies

$$|\lambda_1 \langle v_1, u \rangle| \gneq |\lambda_2 \langle v_2, u \rangle| \geq \cdots .$$

Denote by $\phi^{(t)}(u)$ the output of t repeated iterations of ϕ on u. Then,

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

That is, u converges to v_1 at a quadratic rate.[13]

---

[13] Lemma 5.1, [A2014].

Matrix Power Method

Remark

In contrast, for symmetric positive definite matrices, the rate of convergence is at upper bounded linearly in λ_1/λ_2.[14]

- Prove as exercise. Why is the convergence for $T \in S^3\mathbb{R}^n$ quadratic?

[14] See also, [D1999]

Perturbation of Odeco Tensor

In estimating an odeco tensor T, we might produce a tensor $\hat{T}$ that is not odeco.

Perturbation of Odeco Tensor

In estimating an odeco tensor T, we might produce a tensor $\hat{T}$ that is not odeco.

- [A2014] designed an algorithm to iteratively estimate the robust eigenvectors of T.

Robust Tensor Power Method

Algorithm 2 Robust Tensor Power Method (RTPM)

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

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

Analysis of Algorithm

In the following:

- $\hat{T} = T + E \in S^3\mathbb{R}^k$ symmetric; $T = \sum_{i=1}^k \lambda_i v_i^{\otimes 3}$ odeco
- $\lambda_{\min}$ and $\lambda_{\max}$ the min/max λ_i's
- $\|E\|_{\text{op}} \leq \epsilon$

Theorem (Thm. 5.1, [A2014])

Let $\delta \in (0,1)$. *If* $\epsilon = O(\frac{\lambda_{\min}}{k})$, $N = \Omega(\log k + \log\log(\frac{\lambda_{\max}}{\epsilon}))$, *and* $L = \text{poly}(k)\log(\frac{1}{\delta})$, *running* RTPM^k *will yield, w.p.* $1 - \delta$,

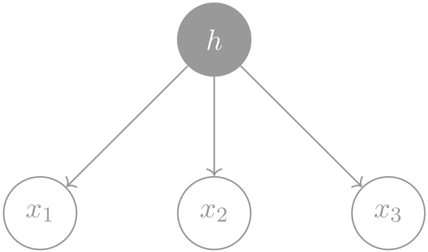
$$\|v_i - \hat{v}_i\| = O\left(\frac{\epsilon}{\lambda_i}\right) \qquad \left|\lambda_i - \hat{\lambda}_i\right| = O(\epsilon)$$

$$\left\|T - \sum_{j=1}^{k} \hat{\lambda}_j \hat{v}_j^{\otimes 3}\right\| \leq O(\epsilon).$$

Return to Topic Modeling

Setup: t topics, vocabulary size d, and 3-word long documents.

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



Using Tensors

From the d possible words, $e_1, \ldots, e_d$, generate the vector space of all 'words objects':

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

Using Tensors

From the d possible words, $e_1, \ldots, e_d$, generate the vector space of all 'words objects':

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

We interpret $x \in V$ as a probability vector, where the weight on the ith coordinate is the probability the word is e_i.

Using Tensors

Now, we want to create the space of all possible three-word documents: $V^{\otimes 3}$.

Using Tensors

Now, we want to create the space of all possible three-word documents: $V^{\otimes 3}$.

- Since we assume that the choice of 3 words in a single document is *conditionally independent*, this means that *expectation is multilinear*.

Using Tensors

Now, we want to create the space of all possible three-word documents: $V^{\otimes 3}$.

- 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:

$$\begin{aligned}\mathbb{E}[x_1 \otimes x_2 | h = j] &= \mathbb{E}[x_1 | h = j] \otimes \mathbb{E}[x_2 | h = j] \\ &= \mu_j \otimes \mu_j.\end{aligned}$$

Using Tensors

Theorem (A2012)

If $M_2 := \mathbb{E}[x_1 \otimes x_2]$ *and* $M_2 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3]$, *then:*

$$M_2 = \sum_{i=1}^{k} w_i \mu_i^{\otimes 2}$$

$$M_3 = \sum_{i=1}^{k} w_i \mu_i^{\otimes 3}$$

Whitening

We are almost at a point where we can use the Robust Tensor Power Method to deduce the probabilities μ_i (i.e. the robust eigenvectors) and the weights w_i (i.e. the eigenvalues).

Whitening

We are almost at a point where we can use the Robust Tensor Power Method to deduce the probabilities μ_i (i.e. the robust eigenvectors) and the weights w_i (i.e. the eigenvalues).

- But we need to make sure the μ_i's are orthonormal.

Whitening

We can take advantage of M_2, which is just an invertible matrix, *conditioned upon*:

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

Whitening

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

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

so that setting $\bar{\mu}_i = \sqrt{w_i} W^T \mu_i$ forms a set of orthonormal vectors.

Whitening

It then follows that:

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

Tensor Decomposition for LDA

In the LDA model, define the following:

$$
\begin{aligned}
M_1 &:= \mathbb{E}[x_1] \\
M_2 &:= \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0 + 1} M_1 \otimes M_1 \\
M_3 &:= \mathbb{E}[x_1 \otimes x_2 \otimes x_3] \\
&\quad - \frac{\alpha_0}{\alpha_0 + 2} \left(\mathbb{E}[x_1 \otimes x_2 \otimes M_1] + \cdots + \mathbb{E}[M_1 \otimes x_1 \otimes x_2] \right) \\
&\quad + \frac{2\alpha_0^2}{(\alpha_0 + 2)(\alpha_0 + 1)} M_1^{\otimes 3}
\end{aligned}
$$

Tensor Decomposition for LDA

Theorem (A2012)

Let M_1, M_2, M_3 as above. Then:

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

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

References

[A2012] Anandkumar, Anima, et al. "A spectral algorithm for latent dirichlet allocation." *Advances in Neural Information Processing Systems*. 2012.

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

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

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

[D1997] Del 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] Draisma, 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] Hillar, Christopher J., and Lek-Heng Lim. "Most tensor problems are NP-hard." *Journal of the ACM (JACM)* 60.6 (2013): 45.

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

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

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

[M2016] Moitra, Ankur. "Tensor Decompositions and their Applications." *YouTube*, uploaded by Centre International de Rencontres Mathématiques, 16 February 2016, `link`

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

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

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

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

[V2014] Vannieuwenhoven, 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] Zhang, Tong, and Gene H. Golub. "Rank-one approximation to high order tensors." *SIAM Journal on Matrix Analysis and Applications* 23.2 (2001): 534-550.