Tensor decomposition and completion

Yuxin Chen

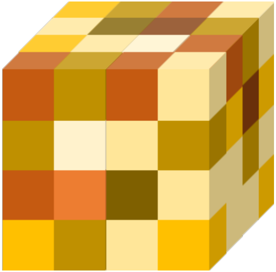
Princeton University, Fall 2018

Outline

- Tensor decomposition
- Latent variable models $\&$ tensor decomposition
- Tensor power method
- Tensor completion

Tensor decomposition

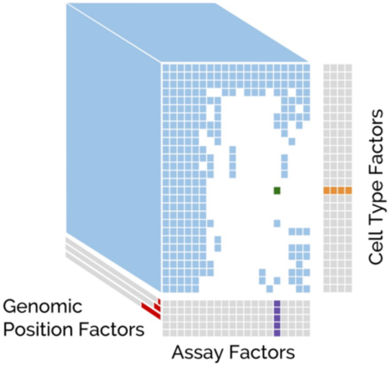
Tensor



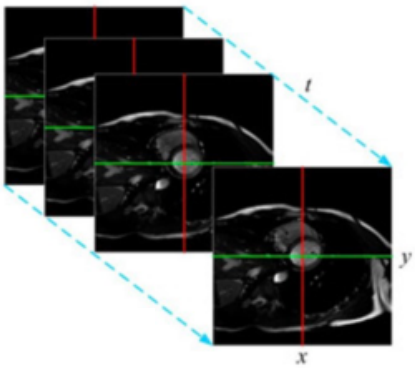
An order-d tensor $\boldsymbol{T} = [T_{i_1,\cdots,i_d}]_{1\leq i_1,\cdots,i_d\leq n}$ is a d-way array

- a matrix is a tensor of order 2

Ubiquity of high-dimensional tensor data



computational genomics
— *fig. credit: Schreiber et al. 19*



dynamic MRI
— *fig. credit: Liu et al. 17*

Basics

- **Rank-1 tensor:** $\boldsymbol{T} = \boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}$ denotes a tensor such that

$$T_{i_1,\cdots,i_d} = x_{i_1} x_{i_2} \cdots x_{i_d}$$

- the inner product of two tensors $\boldsymbol{T}$ and $\boldsymbol{A}$:

$$\langle \boldsymbol{T}, \boldsymbol{A} \rangle := \sum_{i_1,\cdots,i_d} T_{i_1,\cdots,i_d} A_{i_1,\cdots,i_d}$$

- the Frobenius norm of a tensor $\boldsymbol{T}$:

$$\|\boldsymbol{T}\|_{\mathrm{F}} := \sqrt{\sum_{i_1,\cdots,i_d} T^2_{i_1,\cdots,i_d}}$$

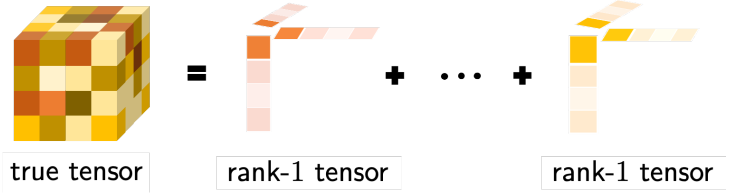
- the operator norm of an order-d tensor $\boldsymbol{T}$:

$$\|\boldsymbol{T}\| = \max_{\{\boldsymbol{u}_i\}:\|\boldsymbol{u}_i\|_2=1} \langle \boldsymbol{T}, \boldsymbol{u}_1 \otimes \cdots \otimes \boldsymbol{u}_d \rangle$$

Tensor decomposition

Suppose we observe an order-d tensor

$$\boldsymbol{T} = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \otimes \boldsymbol{u}_i \otimes \cdots \otimes \boldsymbol{u}_i$$



Question: can we recover $\{\boldsymbol{u}_i\}$ and $\{\lambda_i\}$ given $\boldsymbol{T}$?

- if $d = 2$ (matrix case), it is often not recoverable; what if $d \geq 3$?
- this question arises in a number of latent-variable models

Latent variable models and tensor decomposition

Notation

- probability simplex

$$\Delta_n := \{\boldsymbol{z} \in \mathbb{R}^n \mid z_i \geq 0, \forall i;\ \mathbf{1}^\top \boldsymbol{z} = 1\}$$

- any vector $\boldsymbol{w} \in \Delta_n$ represents a distribution (or probability mass function) over n objects

A simple topic model

Consider a collection of documents

- r: the number of distinct topics
- n: the number of distinct words in vocabulary

A simple topic model

Consider a collection of documents

- each time, draw 3 words as follows
  - pick $\underbrace{\text{a topic } h}_{\text{latent variable}}$ according to distribution $[w_1, \cdots, w_r] \in \Delta_r$ s.t.

$$\mathbb{P}\{h = j\} = w_j, \qquad 1 \leq j \leq r$$

  - given topic h, draw 3 independent words from this topic according to the distribution

$$\underbrace{\boldsymbol{\mu}_h}_{\text{determined only by the topic}} \in \Delta_n$$

Goal: recover $\{\boldsymbol{\mu}_i\}$ and $\{w_i\}$ from the collected samples

Moment method for the topic model

Denote the 3 words we draw as $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)} \in \mathbb{R}^n$:

$$\boldsymbol{x}^{(i)} = \boldsymbol{e}_j \qquad \text{if the } i\text{th word is } j$$

It is straightfoward to check

$$\boldsymbol{M}_2 := \mathbb{E}\big[\boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)}\big] = \sum_{i=1}^{r} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

$$\boldsymbol{M}_3 := \mathbb{E}\big[\boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)} \otimes \boldsymbol{x}^{(3)}\big] = \sum_{i=1}^{r} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

- $\boldsymbol{M}_2$, $\boldsymbol{M}_3$ can be reliably estimated when we have many samples
- recovering $\{\boldsymbol{\mu}_i\}$ and $\{w_i\}$ from $\boldsymbol{M}_2$, $\boldsymbol{M}_3$
  $\Longleftrightarrow$ tensor decomposition

Latent Dirichlet allocation (LDA)

More complicated topic models: mixed membership models, where each data might belong to multiple latent classes simultaneously

This means: the latent variable h is no longer an indicator of topics, but rather, a topic mixture $\boldsymbol{h} \in \Delta_r$

Latent Dirichlet allocation (LDA)

- n: the number of distinct words in the vocabulary
- r: the number of distinct topics
- topic i has word distribution $\boldsymbol{\mu}_i \in \Delta_n$ $(1 \leq i \leq n)$
- each time, draw 3 words as follows
  - draw $\underbrace{\text{topic mixture } \boldsymbol{h}}_{\text{latent variables}} \in \Delta_r$ according to Dirichlet distribution

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

  - draw $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)} \in \mathbb{R}^n$ independently according to the *mixed distribution* $\sum_{i=1}^r h_i \boldsymbol{\mu}_i$

Moment method for latent Dirichlet allocation

$$\boldsymbol{M}_1 := \mathbb{E}[\boldsymbol{x}^{(1)}]$$

$$\boldsymbol{M}_2 := \mathbb{E}[\boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)}] - \frac{\alpha_0}{\alpha_0+1}\boldsymbol{M}_1 \otimes \boldsymbol{M}_1 = \sum_{i=1}^{r} \frac{\alpha_i}{(\alpha_0+1)\alpha_0}\boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

$$\begin{aligned}
\boldsymbol{M}_3 := {} & \mathbb{E}[\boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)} \otimes \boldsymbol{x}^{(3)}] - \frac{\alpha_0}{\alpha_0+2} \\
& \cdot \Big(\mathbb{E}[\boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)} \otimes \boldsymbol{M}_1] + \mathbb{E}[\boldsymbol{x}^{(1)} \otimes \boldsymbol{M}_1 \otimes \boldsymbol{x}^{(2)}] + \mathbb{E}[\boldsymbol{M}_1 \otimes \boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)}]\Big) \\
& + \frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}\boldsymbol{M}_1 \otimes \boldsymbol{M}_1 \otimes \boldsymbol{M}_1 \\
= {} & \sum_{i=1}^{r} \frac{2\alpha_i}{(\alpha_0+2)(\alpha_0+1)\alpha_0}\boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i
\end{aligned}$$

- estimate $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\boldsymbol{M}_3$ from samples (assuming α_0 is known)
- recover $\{\boldsymbol{\mu}_i\}$ and $\{\alpha_i\}_{i\geq 1}$ from $\boldsymbol{M}_2$, $\boldsymbol{M}_3$ (tensor decomposition)

Gaussian mixture model

- r Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 \boldsymbol{I}_n)$ $(1 \leq i \leq r)$
- a sample $\boldsymbol{x} \in \mathbb{R}^n$ is drawn as follows
    - the latent indicator variable h is generated according to distribution $[w_1, \cdots, w_r] \in \Delta_r$ s.t.

$$\mathbb{P}(h = i) = w_i, \qquad 1 \leq i \leq r$$

    - generate $\boldsymbol{x}$ from $\mathcal{N}(\boldsymbol{\mu}_h, \sigma^2 \boldsymbol{I}_n)$

Moment method for Gaussian mixture model

$$M_2 := \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}] - \sigma^2 \boldsymbol{I} = \sum_{i=1}^{r} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

$$\begin{aligned} \boldsymbol{M}_3 &:= \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}] \\ &\qquad - \sigma^2 \sum_{i=1}^{n} \left(\mathbb{E}[\boldsymbol{x}] \otimes \boldsymbol{e}_i \otimes \boldsymbol{e}_i + \boldsymbol{e}_i \otimes \mathbb{E}[\boldsymbol{x}] \otimes \boldsymbol{e}_i + \boldsymbol{e}_i \otimes \boldsymbol{e}_i \otimes \mathbb{E}[\boldsymbol{x}] \right) \\ &= \sum_{i=1}^{r} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \end{aligned}$$

- $\boldsymbol{M}_2$, $\boldsymbol{M}_3$ and $\mathbb{E}[\boldsymbol{x}]$ can all be reliably estimated when there are many samples
- recover $\{\boldsymbol{\mu}_i\}$ and $\{w_i\}$ from $\boldsymbol{M}_2$, $\boldsymbol{M}_3$ (tensor decomposition)

Tensor power method

Main task

Given

$$\boldsymbol{M}_2 = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \otimes \boldsymbol{u}_i$$
$$\boldsymbol{M}_3 = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \otimes \boldsymbol{u}_i \otimes \boldsymbol{u}_i$$

where $\lambda_i > 0$

Question: can we recover $\{\lambda_i\}$ and $\{\boldsymbol{u}_i\}$ from $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$?

An easier case: orthogonal decomposition

Given

$$\boldsymbol{M}_2 = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \otimes \boldsymbol{u}_i$$

$$\boldsymbol{M}_3 = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \otimes \boldsymbol{u}_i \otimes \boldsymbol{u}_i$$

where $\lambda_i > 0$, $r \leq n$, and $\{\boldsymbol{u}_i\}$ are orthonormal

Question: can we recover $\{\lambda_i\}$ and $\{\boldsymbol{u}_i\}$ from $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$?

Tensor power method

Define

$$\boldsymbol{T}(\boldsymbol{I}, \boldsymbol{x}, \cdots, \boldsymbol{x}) := \sum_{i=1}^{r} \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{x})^{d-1} \boldsymbol{u}_i$$

- if $d = 2$ (matrix case): $\boldsymbol{T}(\boldsymbol{I}, \boldsymbol{x}) = \boldsymbol{T}\boldsymbol{x}$

Algorithm 5.1 Tensor power method

1: **initialize** $\boldsymbol{x}_0 \leftarrow$ random unit vector
2: **for** $t = 1, 2, \cdots$ **do**
3: $\quad \boldsymbol{x}_t = \boldsymbol{T}(\boldsymbol{I}, \boldsymbol{x}_{t-1}, \cdots, \boldsymbol{x}_{t-1})$ (power iteration)
4: $\quad \boldsymbol{x}_t \leftarrow \frac{1}{\|\boldsymbol{x}_t\|_2} \boldsymbol{x}_t$ (re-normalization)

Convergence analysis

Theorem 5.1 (Convergence of tensor power method)

Suppose $\{\boldsymbol{u}_i\}$ *are orthonormal,* $\lambda_i > 0$ *(*$1 \leq i \leq r$*),* $r \leq n$*, and* $d = 3$*. Then for any* $1 \leq i \leq r$*,*

$$1 - \frac{\left(\boldsymbol{u}_i^\top \boldsymbol{x}_t\right)^2}{\|\boldsymbol{x}_t\|_2^2} \leq \lambda_i^2 \sum_{j:j\neq i} \lambda_j^{-2} \Big(\frac{\lambda_j \boldsymbol{u}_j^\top \boldsymbol{x}_0}{\lambda_i \boldsymbol{u}_i^\top \boldsymbol{x}_0} \Big)^{2^{t+1}}$$

- tensor power method converges quadratically to some $\boldsymbol{u}_i$
- it converges to a point $\boldsymbol{u}_i$ associated with the largest $\lambda_i \boldsymbol{u}_i^\top \boldsymbol{x}_0$
  - both the eigenvalue and the initial point matter!

Proof of Theorem 5.1

Note that removing "re-normalization" steps does not affect $\frac{(\boldsymbol{u}_i^\top \boldsymbol{x}_t)^2}{\|\boldsymbol{x}_t\|_2^2}$ at all. For simplicity, we assume

$$\boldsymbol{x}_t = \boldsymbol{T}(\boldsymbol{I}, \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-1}) = \sum_{i=1}^{r} \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{x}_{t-1})^2 \boldsymbol{u}_i$$

Observe that

- since $\boldsymbol{x}_1 = \sum_{i=1}^{r} \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{x}_0)^2 \boldsymbol{u}_i$, we have

$$\left(\boldsymbol{u}_i^\top \boldsymbol{x}_1\right)^2 = \lambda_i^2 (\boldsymbol{u}_i^\top \boldsymbol{x}_0)^4$$

- since $\boldsymbol{x}_2 = \sum_{i=1}^{r} \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{x}_1)^2 \boldsymbol{u}_i$, we have

$$\left(\boldsymbol{u}_i^\top \boldsymbol{x}_2\right)^2 = \lambda_i^2 (\boldsymbol{u}_i^\top \boldsymbol{x}_1)^4 = \lambda_i^6 (\boldsymbol{u}_i^\top \boldsymbol{x}_0)^8$$

- since $\boldsymbol{x}_3 = \sum_{i=1}^{r} \lambda_i (\boldsymbol{u}_i^\top \boldsymbol{x}_2)^2 \boldsymbol{u}_i$, we have

$$\left(\boldsymbol{u}_i^\top \boldsymbol{x}_3\right)^2 = \lambda_i^2 (\boldsymbol{u}_i^\top \boldsymbol{x}_2)^4 = \lambda_i^{14} (\boldsymbol{u}_i^\top \boldsymbol{x}_0)^{16}$$

Proof of Theorem 5.1 (cont.)

By induction, one has

$$\big(\boldsymbol{u}_i^\top \boldsymbol{x}_t\big)^2 = \lambda_i^{2^{t+1}-2} (\boldsymbol{u}_i^\top \boldsymbol{x}_0)^{2^{t+1}}, \qquad 1 \le i \le r$$

This implies

$$\frac{\big(\boldsymbol{u}_i^\top \boldsymbol{x}_t\big)^2}{\|\boldsymbol{x}_t\|_2^2} = \frac{\big(\boldsymbol{u}_i^\top \boldsymbol{x}_t\big)^2}{\sum_{j=1}^r \big(\boldsymbol{u}_j^\top \boldsymbol{x}_t\big)^2} = \frac{(\lambda_i \boldsymbol{u}_i^\top \boldsymbol{x}_0)^{2^{t+1}}}{\sum_{j=1}^r \big(\frac{\lambda_i}{\lambda_j}\big)^2 (\lambda_j \boldsymbol{u}_j^\top \boldsymbol{x}_0)^{2^{t+1}}}$$

and hence

$$\begin{aligned}
1 - \frac{\big(\boldsymbol{u}_i^\top \boldsymbol{x}_t\big)^2}{\|\boldsymbol{x}_t\|_2^2} &= \frac{\sum_{j:j\neq i} \big(\frac{\lambda_i}{\lambda_j}\big)^2 (\lambda_j \boldsymbol{u}_j^\top \boldsymbol{x}_0)^{2^{t+1}}}{\sum_j \big(\frac{\lambda_i}{\lambda_j}\big)^2 (\lambda_j \boldsymbol{u}_j^\top \boldsymbol{x}_0)^{2^{t+1}}} \\
&\le \frac{\sum_{j:j\neq i} \big(\frac{\lambda_i}{\lambda_j}\big)^2 (\lambda_j \boldsymbol{u}_j^\top \boldsymbol{x}_0)^{2^{t+1}}}{(\lambda_i \boldsymbol{u}_i^\top \boldsymbol{x}_0)^{2^{t+1}}} \\
&= \lambda_i^2 \sum_{j:j\neq i} \lambda_j^{-2} \Big(\frac{\lambda_j \boldsymbol{u}_j^\top \boldsymbol{x}_0}{\lambda_i \boldsymbol{u}_i^\top \boldsymbol{x}_0}\Big)^{2^{t+1}}
\end{aligned}$$

General case: reduction to orthogonally decomposable tensors

Suppose $r \leq n$, but $\{\boldsymbol{u}_i\}$ are not orthonormal

Key idea: use $\boldsymbol{M}_2$ to find a "whitening matrix" that allows us to orthogonalize $\{\boldsymbol{u}_i\}$

General case: reduction to orthogonally decomposable tensor

Let $\boldsymbol{W}$ be a whitening matrix (e.g. $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}$) obeying

$$\boldsymbol{W}^\top \boldsymbol{M}_2 \boldsymbol{W} = \boldsymbol{I} \tag{5.1}$$

Then

$$\begin{aligned} \boldsymbol{M}_3(\boldsymbol{W}, \boldsymbol{W}, \boldsymbol{W}) &= \sum_{i=1}^{r} \lambda_i (\boldsymbol{W}^\top \boldsymbol{u}_i) \otimes (\boldsymbol{W}^\top \boldsymbol{u}_i) \otimes (\boldsymbol{W}^\top \boldsymbol{u}_i) \\ &= \sum_{i=1}^{r} \lambda_i \tilde{\boldsymbol{u}}_i \otimes \tilde{\boldsymbol{u}}_i \otimes \tilde{\boldsymbol{u}}_i \end{aligned}$$

where $\{\tilde{\boldsymbol{u}}_i\}$ become orthonormal vectors

- use the tensor power method to recover $\{\tilde{\boldsymbol{u}}_i\}$

Reference

- "*Tensor decompositions for learning latent variable models*," A. Anandkumar, R. Ge, D. Hsu, S. Kakade, *Journal of machine learning research*, 2014.
- "*Tensor decompositions and applications*," T. Kolda, B. Bader, *SIAM review*, 2009.
- "*Orthogonal tensor decompositions*," T. Kolda, *SIAM journal on matrix analysis and applications*, 2001.
- "*Spectral learning on matrices and tensors*," M. Janzamin, R. Ge, J. Kossaifi and A. Anandkumar, *Foundations and Trends in Machine Learning*, 2019.
- "*On the best rank-1 and rank-*$(R_1, R_2, \cdots, R_n)$ *approximation of higher-order tensors*," L. De Lathauwer, B. De Moor, J. Vandewalle, *SIAM journal on matrix analysis and applications*, 2000.

Reference

- "*Spectral algorithms for tensor completion*," A. Montanari, N. Sun, *Communications on pure and applied mathematics*, 2018.
- "*Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees*," C. Cai, G. Li, Y. Chi, H. V. Poor, Y. Chen, *Annals of Statistics*, 2020.