Tensor Decompositions: Exploiting Structure in Observed Moments

Anima Anandkumar

U.C. Irvine

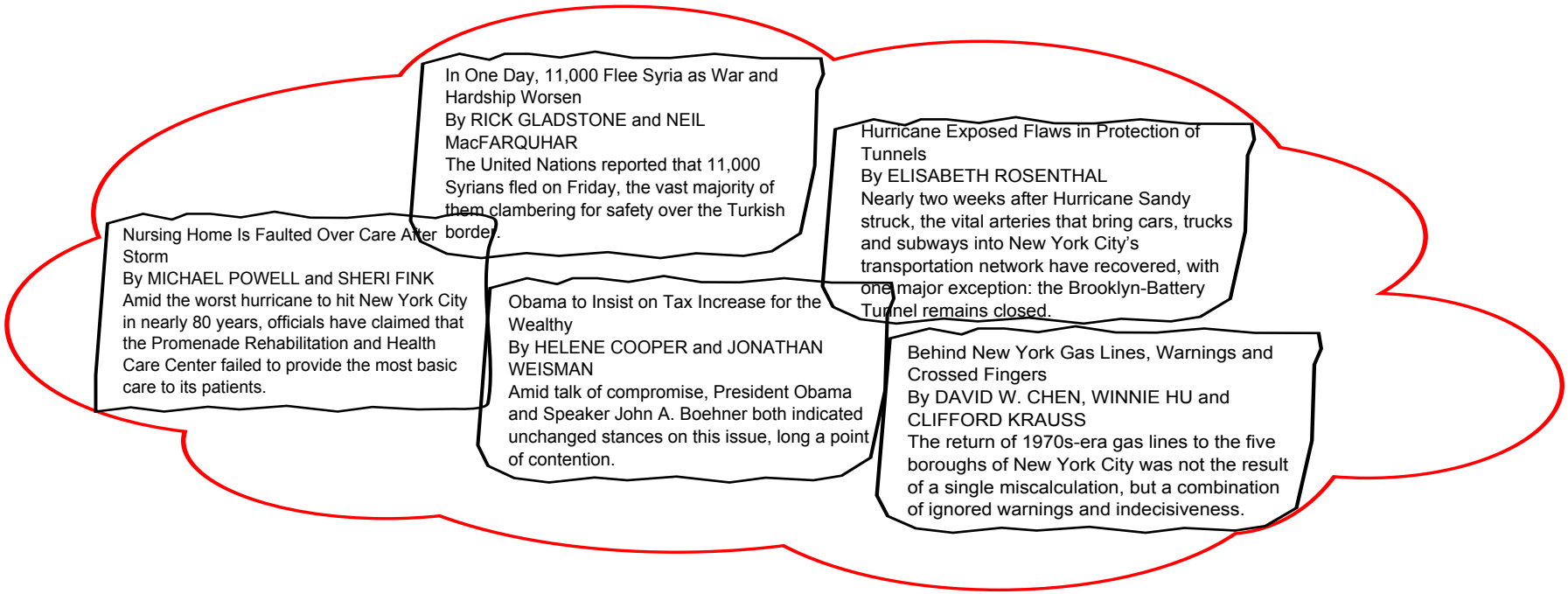
Joint work with Daniel Hsu, Rong Ge, Sham Kakade and Matus Telgarsky.

Latent Variable Modeling

Goal: Discover hidden effects from unlabeled data

Example: document modeling

- Observations: words. Hidden: topics.



Unsupervised learning of latent variable models: methods and guarantees

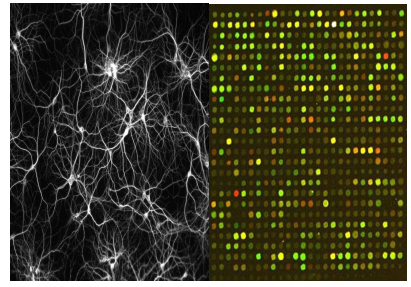
Other Applications of Latent Variable Modeling

Social Network Modeling

- Observed: social interactions.
- Hidden: communities, relationships

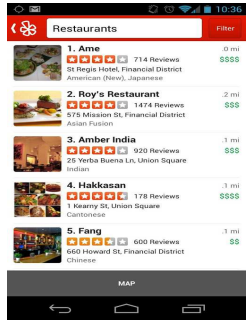
Bio-Informatics

- Observed: gene expressions or neural activity.
- Hidden: gene regulators, functional mapping.



Recommendation Systems

- Observed: recommendations: e.g. yelp reviews.
- Hidden: User and business attributes



Learning latent variable models: efficient methods and guarantees

Challenges in Learning Latent Variable Models

Challenges in Identifiability

- When can latent variables be identified?
- Conditions on the model parameter, e.g. on topic-word matrix and on topic proportions distributions?
- Does identifiability also lead to tractable algorithms?

Challenges in Learning Latent Variable Models

Challenges in Identifiability

- When can latent variables be identified?
- Conditions on the model parameter, e.g. on topic-word matrix and on topic proportions distributions?
- Does identifiability also lead to tractable algorithms?

Challenges in Design of Learning Algorithms

- Maximum likelihood learning of topic models NP-hard (Arora et. al.)
- In practice, methods such as Gibbs sampling, variational Bayes etc. but no guarantees
- Guaranteed learning with minimal assumptions? Efficient methods? Low sample and computational complexities?

Challenges in Learning Latent Variable Models

Challenges in Identifiability

- When can latent variables be identified?
- Conditions on the model parameter, e.g. on topic-word matrix and on topic proportions distributions?
- Does identifiability also lead to tractable algorithms?

Challenges in Design of Learning Algorithms

- Maximum likelihood learning of topic models NP-hard (Arora et. al.)
- In practice, methods such as Gibbs sampling, variational Bayes etc. but no guarantees
- Guaranteed learning with minimal assumptions? Efficient methods? Low sample and computational complexities?

Moment-based approach: learning using low order observed moments

Inverse Moment Method

Two step approach:

1. Under *modeling assumptions*, what moment forms arise?
   topic models, HMMs, LDA, mixture of Gaussians models, parsing (e.g. PCFGs), Bayesian networks
2. Can we "invert"/reverse engineer the model from these moments?

This Tutorial

How to utilize observed moments?

- part 1: the moment structure and its inversion
  - When are low order moments sufficient for learning?
  - generalizations of simple (linear algebra) approach
  - aren't these problems hard/non-convex?

- part 2: Use tensor decomposition techniques for finding overlapping communities
  - Introduce mixed membership community model
  - Derive graph moment tensor forms
  - Contrast with state-of-art community detection methods

- part 3: overcomplete models
  - Latent dimensionality $\gg$ observed dimensionality
  - Exploit sparsity conditions
  - ℓ_1 optimization

Two Extremes

- Single hidden state active
  - mixture of Gaussians, single topic per document
- Independent Component Analysis
  - Blind source separation
    audio signal has different speakers talking
  - independent factors

What about the middle ground?

2. Define the models

Mixture Models

(spherical) Mixture of Gaussian:

- k means: $\mu_1, \ldots \mu_k$
- sample cluster $H = i$ with prob. w_i
- observe x, with spherical noise,

$$x = \mu_i + \eta, \quad \eta \sim \mathcal{N}(0, \sigma_i^2 I)$$

(single) Topic Models

- k topics: $\mu_1, \ldots \mu_k$
- sample topic $H = i$ with prob. w_i
- observe m (exchangeable) words

$$x_1, x_2, \ldots x_m \text{ sampled i.i.d. from}$$

- dataset: multiple points / m-word documents
- how to learn the params? $\mu_1, \ldots \mu_k$, $w_1, \ldots w_k$ (and σ_i's)

vector notation!

- k clusters, d dimensions/words, $d \geq k$
- for MOGs:
  - the conditional expectations are:

$$\mathbb{E}[x|\text{cluster i}] = \mu_i$$

- topic models:
  - binary word encoding: $x_1 = [0, 1, 0, \ldots]^\top$
  - the μ_i's are probability vectors
  - for each word, the conditional probabilities are:

$$\Pr[x_1|\text{topic i}] = \mathbb{E}[x_1|\text{topic i}] = \mu_i$$

ICA

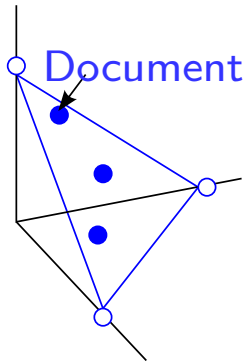
- k mixing directions: $\mu_1, \ldots \mu_k$
- each hidden (scalar) factor, $H_1, H_2, \ldots H_k$, is independently distributed
- observe mixture x, with Gaussian noise,

$$x = \sum_i \mu_i H_i + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2)$$

- in MOG's, only one $H_i = 1$
- how to learn the params? $\mu_1, \ldots \mu_k$

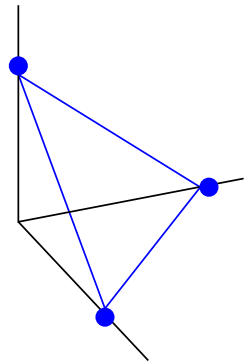
Geometric Picture for Topic Models

Topic proportions vector (H)



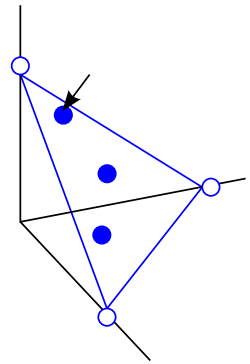
Geometric Picture for Topic Models

Single topic (H)



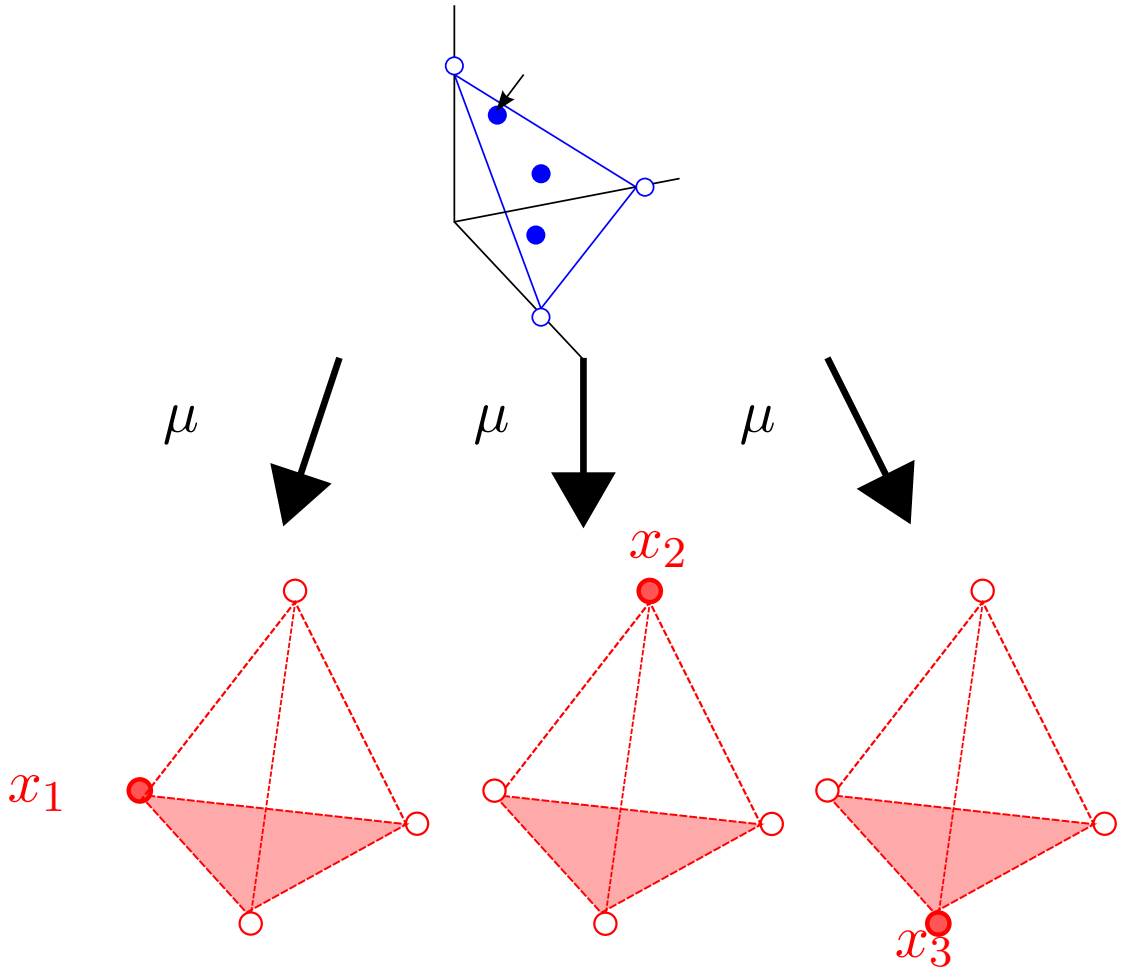
Geometric Picture for Topic Models

Topic proportions vector (H)



Geometric Picture for Topic Models

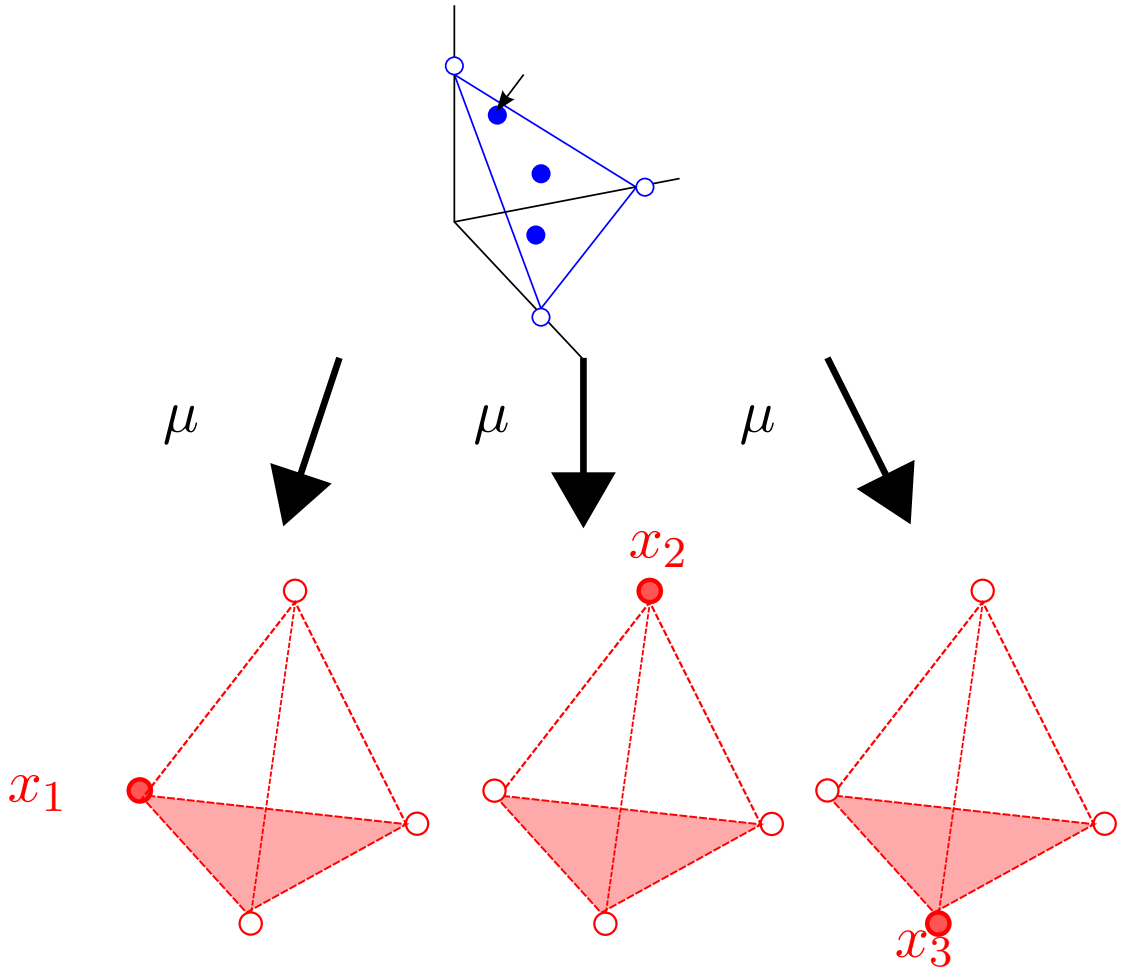
Topic proportions vector (H)



Word generation $(x_1, x_2, \ldots)$

Geometric Picture for Topic Models

Topic proportions vector (H)



Word generation $(x_1, x_2, \ldots)$

Moment-based estimation: co-occurrences of words in documents

3. Approach: the method of moments

The Method of Moments

- (Pearson, 1894): find params consistent with observed moments
- MOGs moments:

$$\mathbb{E}[x],\ \mathbb{E}[xx^\top],\ \mathbb{E}[x \otimes x \otimes x],\ \ldots$$

- Topic model moments:

$$\Pr[x_1], \Pr[x_1, x_2],\ \Pr[x_1, x_2, x_3], \ldots$$

- Identifiability: with exact moments, what order moment suffices?
  - how many words per document suffice?
  - efficient algorithms?

(some) Related Work

- Kruskal's Theorem

  Kruskal (1977), Bhaskara, Charikar, & Vijayaraghavan (2013), ...
- Algebraic Work
  - ICA literature: Cardoso&Common, '96, ...
  - for phylogeny trees: J. T. Chang (1996), E. Mossel & S. Roch (2006),
- Tensor Decomposition Algorithms

  Lathauwer, Moor, & Vandewalle (2000), Zhang & Golub (2001), Anandkumar et. al. (2012), ...
- Structural assumptions/Dictionary learning

  Spielman, Wang & Right (2012), Arora, Ge, & Moitra (2012)

4. Tensor decompositions

With the first moment?

MOGs:

- have:

$$\mathbb{E}[x] = \sum_{i=1}^{k} w_i \mu_i$$

Single Topics:

- with 1 word per document:

$$\Pr[x_1] = \sum_{i=1}^{k} w_i \mu_i$$

ICA:

- define $\mathbb{E}[H_i] := w_i$

$$\mathbb{E}[x] = \sum_{i=1}^{k} w_i \mu_i$$

Not identifiable: only d nums.

With the second moment?

MOGs/ICA:

- additive noise

$$\begin{aligned} & \mathbb{E}[x \otimes x] \\ = \; & \mathbb{E}[(\mu_i + \eta) \otimes (\mu_i + \eta)] \\ = \; & \sum_{i=1}^{k} w_i \; \mu_i \otimes \mu_i + \sigma^2 I \end{aligned}$$

- have a full rank matrix

Single Topics:

- by exchangeability:

$$\begin{aligned} & \Pr[x_1, x_2] \\ = \; & \mathbb{E}[\; \mathbb{E}[x_1|topic] \otimes \mathbb{E}[x_2|topic \\ = \; & \sum_{i=1}^{k} w_i \; \mu_i \otimes \mu_i \end{aligned}$$

- have a low rank matrix!

Still not identifiable!

With three words per document?

- for topics: $d \times d$ matrix, a $d \times d \times d$ tensor:

$$M_2 := \quad \Pr[x_1, x_2] \quad = \sum_{i=1}^{k} w_i \ \mu_i \otimes \mu_i$$

$$M_3 := \quad \Pr[x_1, x_2, x_3] \quad = \sum_{i=1}^{k} w_i \ \mu_i \otimes \mu_i \otimes \mu_i$$

Mixtures of spherical Gaussians

Theorem

The variance σ^2 is is the smallest eigenvalue of the observed covariance matrix $\mathbb{E}[x \otimes x] - \mathbb{E}[x] \otimes \mathbb{E}[x]$. Furthermore, if

$$
\begin{aligned}
M_2 &:= \mathbb{E}[x \otimes x] - \sigma^2 I \\
M_3 &:= \mathbb{E}[x \otimes x \otimes x] \\
&\quad - \sigma^2 \sum_{i=1}^{d} \big(\mathbb{E}[x] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[x] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[x] \big),
\end{aligned}
$$

then

$$
\begin{aligned}
M_2 &= \sum w_i \ \mu_i \otimes \mu_i \\
M_3 &= \sum w_i \ \mu_i \otimes \mu_i \otimes \mu_i .
\end{aligned}
$$

Independent Component Analysis

Theorem

Different higher order moments from MOGs. Use cumulants:

$$M_4 := \mathbb{E}[x \otimes x \otimes x \otimes x] - \big(\mathbb{E}[x \otimes x] \otimes \mathbb{E}[x \otimes x] + \textit{more stuff...}\big),$$

then

$$M_4 = \sum w_i \ \mu_i \otimes \mu_i \otimes \mu_i \otimes \mu_i.$$

Latent Dirichlet Allocation

prior for topic mixture π:

$$p_\alpha(\pi) = \frac{1}{Z} \prod_{i=1}^{k} \pi_i^{\alpha_i - 1}, \quad \alpha_0 := \alpha_1 + \alpha_2 + \cdots + \alpha_k$$

Theorem

Again, three words per doc suffice. Define

$$M_2 := \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0 + 1} \mathbb{E}[x_1] \otimes \mathbb{E}[x_1]$$

$$M_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3] - \frac{\alpha_0}{\alpha_0 + 2} \mathbb{E}[x_1 \otimes x_2 \otimes \mathbb{E}[x_1]] - \textit{more stuff...}$$

Then

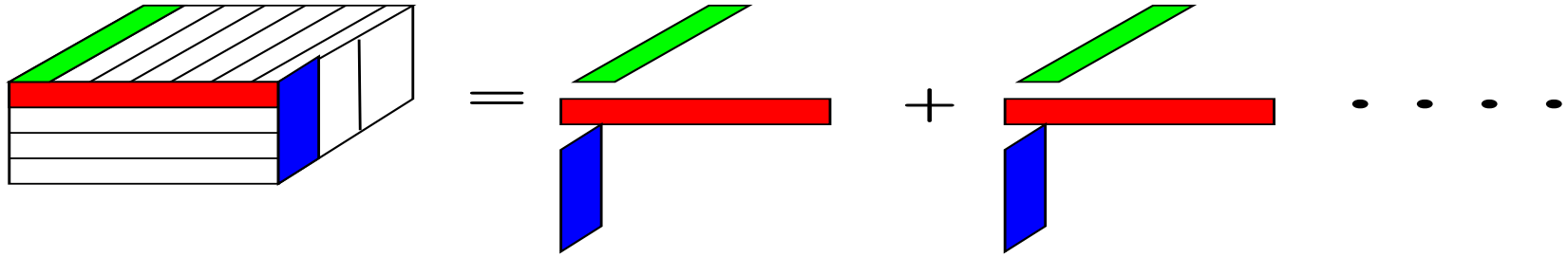
$$M_2 = \sum \tilde{w}_i \ \mu_i \otimes \mu_i$$

$$M_3 = \sum \tilde{w}_i \ \mu_i \otimes \mu_i \otimes \mu_i.$$

Learning without inference!

5. The basic decomposition problem

Low-rank Tensor Decomposition

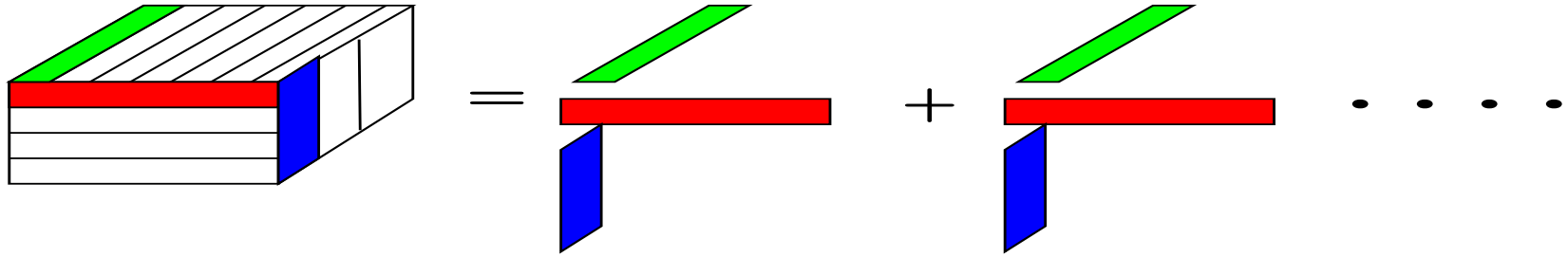


Tensor M_3 $\quad w_1\mu_1^{\otimes 3}$ $\quad w_2\mu_2^{\otimes 3}$

$$\boxed{M_3 = \sum_{i\in[k]} w_i\mu_i^{\otimes 3}}$$

- Rank-k tensor decomposition and typically $k \ll d$
- $u \otimes v \otimes w$ is a rank-1 tensor whose i, j, k^{th} entry is $u_i v_j w_k$.

Low-rank Tensor Decomposition



Tensor M_3 $\quad w_1\mu_1^{\otimes 3}$ $\quad w_2\mu_2^{\otimes 3}$

$$M_3 = \sum_{i\in[k]} w_i\mu_i^{\otimes 3}$$

- Rank-k tensor decomposition and typically $k \ll d$
- $u \otimes v \otimes w$ is a rank-1 tensor whose i, j, k^{th} entry is $u_i v_j w_k$.

Challenges

- Guaranteed algorithm for tensor decomposition?
- Efficient and scalable implementation?
- Noisy tensor decomposition: exact moments not available
- Sample complexity? How large d compared to k?

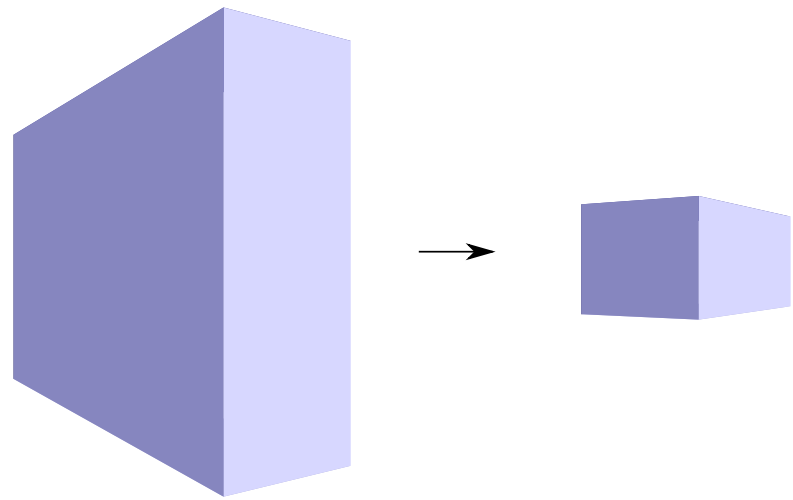
Dimensionality Reduction for Tensor Decomposition

$$M_3 = \sum_{i \in [k]} w_i \mu_i^{\otimes 3}$$

- Rank-k tensor decomposition and typically $k \ll d$
- M_3 has size $O(d^3)$ but number of free parameters: $dk + k$

First Step: Dimensionality Reduction

- Convert M_3 of size $d \times d \times d$ to a tensor T of size $k \times k \times k$
- Carry out decomposition of T



Tensor M_3 Tensor T

Advantages

- Reduced computation
- Robustness to noise

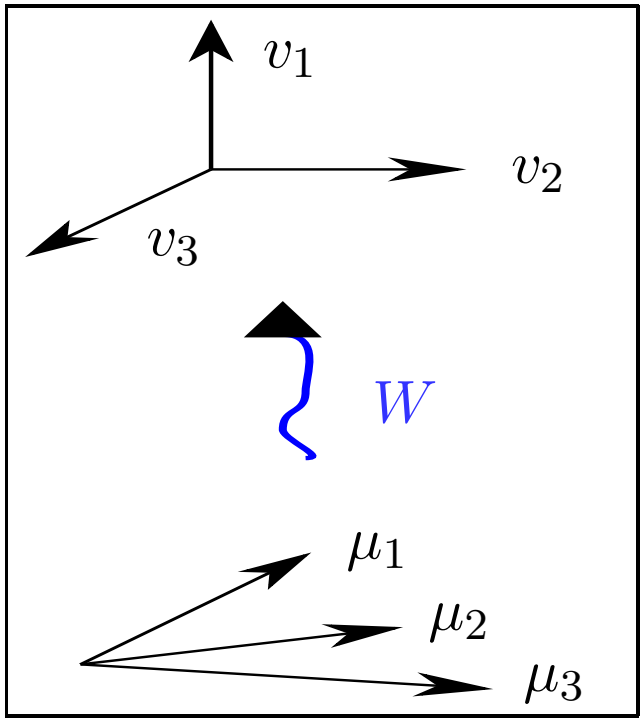
Dimensionality reduction through multi-linear transforms

Dimensionality Reduction through Whitening

$$M_3 = \sum_{i \in [k]} w_i \mu_i^{\otimes 3}$$

Whitening: Conversion of (expected) M_3 to Orthogonal Symmetric Tensor T

- Whitening: W s.t. $W^\top \mu \text{Diag}(w)^{1/2} = R$, $R^\top R = I \in \mathbb{R}^{k \times k}$.
- SVD of M_2 gives Col(μ):
  $M_2 = \mu \text{Diag}(w) \mu^\top = USU^\top$ and
  $W := US^{1/2}$



Multi-linear Transformation of Third Moment Tensor

$$T := M_3(W, W, W) = \sum_i \lambda_i v_i^{\otimes 3}$$

- T is symmetric orthogonal tensor: $\{v_i\}$ are orthonormal.

The basic decomposition problem

The basic decomposition problem

Problem: Given $T \in \mathbb{R}^{k \times k \times k}$ with the promise that

$$T = \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$$

for some orthonormal basis $\{\vec{v}_t\}$ of $\mathbb{R}^k$ and positive scalars $\{\lambda_t > 0\}$, approximately find $\{(\vec{v}_t, \lambda_t)\}$
(up to some desired precision).

Basic questions

Basic questions

1. Is $\{(\vec{v}_t, \lambda_t)\}$ uniquely determined?

Basic questions

1. Is $\{(\vec{v}_t, \lambda_t)\}$ uniquely determined?
2. If so, is there an efficient algorithm for finding the decomposition?

Basic questions

1. Is $\{(\vec{v}_t, \lambda_t)\}$ uniquely determined?
2. If so, is there an efficient algorithm for finding the decomposition?
3. What if T is perturbed by some small amount?

> Perturbed problem: Same as the original problem, except instead of T, we are given $T + E$ for some "error tensor" E.

How "large" can E be if we want ε precision?

Analogous matrix problem

Matrix problem: Given $M \in \mathbb{R}^{k \times k}$ with the promise that

$$M = \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \ \vec{v}_t{}^{\top}$$

for some orthonormal basis $\{\vec{v}_t\}$ of $\mathbb{R}^k$ (w.r.t. standard inner product) and positive scalars $\{\lambda_t > 0\}$, approximately find $\{(\vec{v}_t, \lambda_t)\}$
(up to some desired precision).

Analogous matrix problem

- We're promised that M is symmetric and positive definite, so requested decomposition is an **eigendecomposition**.

Analogous matrix problem

- We're promised that M is symmetric and positive definite, so requested decomposition is an **eigendecomposition**.

  In this case, an eigendecomposition **always exists**, and **can be found efficiently**.

Analogous matrix problem

- We're promised that M is symmetric and positive definite, so requested decomposition is an **eigendecomposition**.

  In this case, an eigendecomposition **always exists**, and **can be found efficiently**.

  It is **unique** if and only if the $\{\lambda_i\}$ are distinct.

Analogous matrix problem

- We're promised that M is symmetric and positive definite, so requested decomposition is an **eigendecomposition**.

  In this case, an eigendecomposition **always exists**, and **can be found efficiently**.

  It is **unique** if and only if the $\{\lambda_i\}$ are distinct.

- What if M is perturbed by some small amount?

> Perturbed matrix problem: Same as the original problem, except instead of M, we are given $M + E$ for some "error matrix" E

Analogous matrix problem

- We're promised that M is symmetric and positive definite, so requested decomposition is an **eigendecomposition**.

  In this case, an eigendecomposition **always exists**, and **can be found efficiently**.

  It is **unique** if and only if the $\{\lambda_i\}$ are distinct.

- What if M is perturbed by some small amount?

  Perturbed matrix problem: Same as the original problem, except instead of M, we are given $M + E$ for some "error matrix" E

  Answer provided by **matrix perturbation theory** (*e.g.*, Davis-Kahan), which requires $\|E\|_2 < \min_{i \neq j} |\lambda_i - \lambda_j|$.

Back to the original problem

Problem: Given $T \in \mathbb{R}^{k \times k \times k}$ with the promise that

$$T = \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$$

for some orthonormal basis $\{\vec{v}_t\}$ of $\mathbb{R}^k$ (w.r.t. standard inner product) and positive scalars $\{\lambda_t > 0\}$, approximately find $\{(\vec{v}_t, \lambda_t)\}$
(up to some desired precision).

Back to the original problem

Problem: Given $T \in \mathbb{R}^{k \times k \times k}$ with the promise that

$$T = \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$$

for some orthonormal basis $\{\vec{v}_t\}$ of $\mathbb{R}^k$ (w.r.t. standard inner product) and positive scalars $\{\lambda_t > 0\}$, approximately find $\{(\vec{v}_t, \lambda_t)\}$
(up to some desired precision).

Such decompositions **do not necessarily exist**, even for symmetric tensors.

Back to the original problem

Problem: Given $T \in \mathbb{R}^{k \times k \times k}$ with the promise that

$$T = \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$$

for some orthonormal basis $\{\vec{v}_t\}$ of $\mathbb{R}^k$ (w.r.t. standard inner product) and positive scalars $\{\lambda_t > 0\}$, approximately find $\{(\vec{v}_t, \lambda_t)\}$ (up to some desired precision).

Such decompositions **do not necessarily exist**, even for symmetric tensors.

Where the decompositions do exist, the Perturbed problem asks if they are "robust".

Main ideas

Easy claim: Repeated application of a certain quadratic operator (based on T) recovers a single $(\vec{v}_t, \lambda_t)$ up to any desired precision.

- Orthogonal symmetric tensor: $T = \sum_i \lambda_i v_i^{\otimes 3}$
- $T(I, v_i, v_i) = \sum_j \lambda_j \langle v_i, v_j \rangle^2 v_j = \lambda_i v_i$

Obtaining eigenvectors through power iterations

$$\boxed{u \mapsto \frac{T(I, u, u)}{\|T(I, u, u)\|}}$$

Main ideas

Easy claim: Repeated application of a certain quadratic operator (based on T) recovers a single $(\vec{v}_t, \lambda_t)$ up to any desired precision.

- Orthogonal symmetric tensor: $T = \sum_i \lambda_i v_i^{\otimes 3}$
- $T(I, v_i, v_i) = \sum_j \lambda_j \langle v_i, v_j \rangle^2 v_j = \lambda_i v_i$

Obtaining eigenvectors through power iterations

$$\boxed{u \mapsto \frac{T(I, u, u)}{\|T(I, u, u)\|}}$$

Self-reduction: Replace T with $T - \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$.

Main ideas

Easy claim: Repeated application of a certain quadratic operator (based on T) recovers a single $(\vec{v}_t, \lambda_t)$ up to any desired precision.

- Orthogonal symmetric tensor: $T = \sum_i \lambda_i v_i^{\otimes 3}$
- $T(I, v_i, v_i) = \sum_j \lambda_j \langle v_i, v_j \rangle^2 v_j = \lambda_i v_i$

Obtaining eigenvectors through power iterations

$$\boxed{u \mapsto \frac{T(I, u, u)}{\|T(I, u, u)\|}}$$

Self-reduction: Replace T with $T - \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$.

- Why?: $T - \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t \ = \ \sum_{\tau \neq t} \lambda_\tau \ \vec{v}_\tau \otimes \vec{v}_\tau \otimes \vec{v}_\tau$.

Main ideas

Easy claim: Repeated application of a certain quadratic operator (based on T) recovers a single $(\vec{v}_t, \lambda_t)$ up to any desired precision.

- Orthogonal symmetric tensor: $T = \sum_i \lambda_i v_i^{\otimes 3}$
- $T(I, v_i, v_i) = \sum_j \lambda_j \langle v_i, v_j \rangle^2 v_j = \lambda_i v_i$

Obtaining eigenvectors through power iterations

$$\boxed{u \mapsto \frac{T(I, u, u)}{\|T(I, u, u)\|}}$$

Self-reduction: Replace T with $T - \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$.

- Why?: $T - \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t \ = \ \sum_{\tau \neq t} \lambda_\tau \ \vec{v}_\tau \otimes \vec{v}_\tau \otimes \vec{v}_\tau$.
- Catch: We don't recover $(\vec{v}_t, \lambda_t)$ exactly, so we actually can only replace T with

  $$T - \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t + E_t$$

  for some "error tensor" E_t.

Rest of this talk

1. Identifiability of decomposition $\{(\vec{v}_t, \lambda_t)\}$ from T.
2. A power iteration algorithm for finding the decomposition.
3. Perturbation analysis.

Identifiability of the decomposition

Orthonormal basis $\{\vec{v}_t\}$ of $\mathbb{R}^k$, positive scalars $\{\lambda_t > 0\}$:

$$T = \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$$

In what sense is $\{(\vec{v}_t, \lambda_t)\}$ uniquely determined?

Identifiability of the decomposition

Orthonormal basis $\{\vec{v}_t\}$ of $\mathbb{R}^k$, positive scalars $\{\lambda_t > 0\}$:

$$T = \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$$

In what sense is $\{(\vec{v}_t, \lambda_t)\}$ uniquely determined?

Claim: $\{\vec{v}_t\}$ are isolated local maximizers of certain cubic form $f_T : \mathbb{S}^{k-1} \to \mathbb{R}^k$, and $f_T(\vec{v}_t) = \lambda_t$.

Review: Rayleigh quotient

Recall Rayleigh quotient for matrix $M := \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \vec{v}_t^{\top}$ (assuming $\vec{x} \in \mathbb{S}^{k-1}$):

$$R_M(\vec{x}) \quad := \quad \vec{x}^{\top} M \vec{x} \quad = \quad \sum_{t=1}^{k} \lambda_t \ (\vec{v}_t^{\top} \vec{x})^2.$$

Review: Rayleigh quotient

Recall Rayleigh quotient for matrix $M := \sum_{t=1}^{k} \lambda_t \, \vec{v}_t \vec{v}_t^{\top}$ (assuming $\vec{x} \in \mathbb{S}^{k-1}$):

$$R_M(\vec{x}) \quad := \quad \vec{x}^{\top} M \vec{x} \quad = \quad \sum_{t=1}^{k} \lambda_t \, (\vec{v}_t^{\top} \vec{x})^2.$$

Every $\vec{v}_t$ such that $|\lambda_t| = \max!$ is a maximizer of R_M.

(These are also the only local maximizers.)

The natural cubic form

Consider the function $f_T \colon \mathbb{S}^{k-1} \to \mathbb{R}^k$ given by

$$\vec{x} = (x_1, x_2, \ldots, x_k) \quad \mapsto \quad f_T(\vec{x}) = \sum_{i_1, i_2, i_3} T_{i_1, i_2, i_3} \; x_{i_1} x_{i_2} x_{i_3}.$$

The natural cubic form

Consider the function $f_T \colon \mathbb{S}^{k-1} \to \mathbb{R}^k$ given by

$$\vec{x} = (x_1, x_2, \ldots, x_k) \quad \mapsto \quad f_T(\vec{x}) = \sum_{i_1, i_2, i_3} T_{i_1, i_2, i_3} \; x_{i_1} x_{i_2} x_{i_3}.$$

For our promised $T = \sum_{t=1}^{k} \lambda_t \; \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$, f_T becomes

The natural cubic form

Consider the function $f_T \colon \mathbb{S}^{k-1} \to \mathbb{R}^k$ given by

$$\vec{x} = (x_1, x_2, \ldots, x_k) \quad \mapsto \quad f_T(\vec{x}) = \sum_{i_1,i_2,i_3} T_{i_1,i_2,i_3}\ x_{i_1} x_{i_2} x_{i_3}.$$

For our promised $T = \sum_{t=1}^{k} \lambda_t\ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$, f_T becomes

$$f_T(\vec{x}) = \sum_{t=1}^{k} \lambda_t \sum_{i_1,i_2,i_3} \left(\vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t\right)_{i_1,i_2,i_3} x_{i_1} x_{i_2} x_{i_3}$$

The natural cubic form

Consider the function $f_T \colon \mathbb{S}^{k-1} \to \mathbb{R}^k$ given by

$$\vec{x} = (x_1, x_2, \ldots, x_k) \quad \mapsto \quad f_T(\vec{x}) = \sum_{i_1,i_2,i_3} T_{i_1,i_2,i_3}\ x_{i_1} x_{i_2} x_{i_3}.$$

For our promised $T = \sum_{t=1}^{k} \lambda_t\ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$, f_T becomes

$$\begin{aligned} f_T(\vec{x}) &= \sum_{t=1}^{k} \lambda_t \sum_{i_1,i_2,i_3} \left(\vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t\right)_{i_1,i_2,i_3} x_{i_1} x_{i_2} x_{i_3} \\ &= \sum_{t=1}^{n} \lambda_t \sum_{i_1,i_2,i_3} (\vec{v}_t)_{i_1} (\vec{v}_t)_{i_2} (\vec{v}_t)_{i_3} x_{i_1} x_{i_2} x_{i_3} \end{aligned}$$

The natural cubic form

Consider the function $f_T \colon \mathbb{S}^{k-1} \to \mathbb{R}^k$ given by

$$\vec{x} = (x_1, x_2, \ldots, x_k) \quad \mapsto \quad f_T(\vec{x}) = \sum_{i_1,i_2,i_3} T_{i_1,i_2,i_3}\ x_{i_1} x_{i_2} x_{i_3}.$$

For our promised $T = \sum_{t=1}^{k} \lambda_t\ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$, f_T becomes

$$\begin{aligned} f_T(\vec{x}) &= \sum_{t=1}^{k} \lambda_t \sum_{i_1,i_2,i_3} \left(\vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t\right)_{i_1,i_2,i_3} x_{i_1} x_{i_2} x_{i_3} \\ &= \sum_{t=1}^{n} \lambda_t \sum_{i_1,i_2,i_3} (\vec{v}_t)_{i_1} (\vec{v}_t)_{i_2} (\vec{v}_t)_{i_3} x_{i_1} x_{i_2} x_{i_3} \\ &= \sum_{t=1}^{k} \lambda_t\ (\vec{v}_t{}^{\top} \vec{x})^3. \end{aligned}$$

The natural cubic form

Consider the function $f_T \colon \mathbb{S}^{k-1} \to \mathbb{R}^k$ given by

$$\vec{x} = (x_1, x_2, \ldots, x_k) \quad \mapsto \quad f_T(\vec{x}) = \sum_{i_1,i_2,i_3} T_{i_1,i_2,i_3}\ x_{i_1} x_{i_2} x_{i_3}.$$

For our promised $T = \sum_{t=1}^{k} \lambda_t\ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t$, f_T becomes

$$\begin{aligned} f_T(\vec{x}) &= \sum_{t=1}^{k} \lambda_t \sum_{i_1,i_2,i_3} \left(\vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t\right)_{i_1,i_2,i_3} x_{i_1} x_{i_2} x_{i_3} \\ &= \sum_{t=1}^{n} \lambda_t \sum_{i_1,i_2,i_3} (\vec{v}_t)_{i_1} (\vec{v}_t)_{i_2} (\vec{v}_t)_{i_3} x_{i_1} x_{i_2} x_{i_3} \\ &= \sum_{t=1}^{k} \lambda_t\ (\vec{v}_t{}^{\top} \vec{x})^3. \end{aligned}$$

Observation: $f_T(\vec{v}_t) = \lambda_t$.

Variational characterization

Claim: Isolated local maximizers of f_T on $\mathbb{S}^{k-1}$ are $\{\vec{v}_t\}$.

Variational characterization

Claim: Isolated local maximizers of f_T on $\mathbb{S}^{k-1}$ are $\{\vec{v}_t\}$.

Objective function (with constraint):

$$\vec{x} \mapsto \inf_{\lambda \neq 0} \sum_{t=1}^{k} \lambda_t \ (\vec{v}_t{}^\top \vec{x})^3 - 1.5\lambda(\|\vec{x}\|_2^2 - 1).$$

Variational characterization

Claim: Isolated local maximizers of f_T on $\mathbb{S}^{k-1}$ are $\{\vec{v}_t\}$.

Objective function (with constraint):

$$\vec{x} \mapsto \inf_{\lambda \neq 0} \sum_{t=1}^{k} \lambda_t \ (\vec{v}_t{}^\top \vec{x})^3 - 1.5\lambda(\|\vec{x}\|_2^2 - 1).$$

First-order condition for local maxima:

$$\sum_{t=1}^{k} \lambda_t \ (\vec{v}_t{}^\top \vec{x})^2 \ \vec{v}_t = \lambda \vec{x}.$$

Variational characterization

Claim: Isolated local maximizers of f_T on $\mathbb{S}^{k-1}$ are $\{\vec{v}_t\}$.

Objective function (with constraint):

$$\vec{x} \mapsto \inf_{\lambda \neq 0} \sum_{t=1}^{k} \lambda_t \ (\vec{v}_t{}^\top \vec{x})^3 - 1.5\lambda(\|\vec{x}\|_2^2 - 1).$$

First-order condition for local maxima:

$$\sum_{t=1}^{k} \lambda_t \ (\vec{v}_t{}^\top \vec{x})^2 \ \vec{v}_t = \lambda \vec{x}.$$

Second-order condition for isolated local maxima:

$$\vec{w}^\top \left(2 \sum_{t=1}^{k} \lambda_t \ (\vec{v}_t{}^\top \vec{x}) \vec{v}_t \vec{v}_t{}^\top - \lambda I \right) \vec{w} < 0, \qquad \vec{w} \perp \vec{x}.$$

Intuition behind variational characterization

May as well assume $\vec{v}_t$ is t^{th} coordinate basis vector, so

$$\max_{\vec{x} \in \mathbb{R}^k} f_T(\vec{x}) = \sum_{t=1}^{k} \lambda_t \, x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{k} x_t^2 = 1.$$

Intuition behind variational characterization

May as well assume $\vec{v}_t$ is t^{th} coordinate basis vector, so

$$\max_{\vec{x}\in\mathbb{R}^k} f_T(\vec{x}) = \sum_{t=1}^{k} \lambda_t \, x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{k} x_t^2 = 1.$$

Intuition: Suppose $\operatorname{supp}(\vec{x}) = \{1, 2\}$, and $x_1, x_2 > 0$.

Intuition behind variational characterization

May as well assume $\vec{v}_t$ is t^{th} coordinate basis vector, so

$$\max_{\vec{x}\in\mathbb{R}^k} f_T(\vec{x}) = \sum_{t=1}^{k} \lambda_t \; x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{k} x_t^2 = 1.$$

Intuition: Suppose $\operatorname{supp}(\vec{x}) = \{1, 2\}$, and $x_1, x_2 > 0$.

$$f_T(\vec{x}) = \lambda_1 x_1^3 + \lambda_2 x_2^3 < \lambda_1 x_1^2 + \lambda_2 x_2^2 \leq \max\{\lambda_1, \; \lambda_2\}.$$

Intuition behind variational characterization

May as well assume $\vec{v}_t$ is t^{th} coordinate basis vector, so

$$\max_{\vec{x}\in\mathbb{R}^k} f_T(\vec{x}) = \sum_{t=1}^{k} \lambda_t \; x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{k} x_t^2 = 1.$$

Intuition: Suppose $\text{supp}(\vec{x}) = \{1, 2\}$, and $x_1, x_2 > 0$.

$$f_T(\vec{x}) = \lambda_1 x_1^3 + \lambda_2 x_2^3 < \lambda_1 x_1^2 + \lambda_2 x_2^2 \leq \max\{\lambda_1, \; \lambda_2\}.$$

Better to have $|\text{supp}(\vec{x})| = 1$, *i.e.*, picking $\vec{x}$ to be a coordinate basis vector. □

Tensor power iteration

Obtaining eigenvectors through power iterations

$$u \mapsto \frac{T(I, u, u)}{\|T(I, u, u)\|}$$

Start with some $\vec{x}^{(0)}$, and for $j = 1, 2, \ldots$:

$$\vec{x}^{(j)} := \phi_T\left(\vec{x}^{(j-1)}\right) = \sum_{t=1}^{k} \lambda_t \left(\vec{v}_t{}^\top \vec{x}^{(j-1)}\right)^2 \vec{v}_t.$$

Tensor power iteration

Obtaining eigenvectors through power iterations

$$\boxed{u \mapsto \frac{T(I,u,u)}{\|T(I,u,u)\|}}$$

Start with some $\vec{x}^{(0)}$, and for $j = 1, 2, \ldots$:

$$\vec{x}^{(j)} := \phi_T\left(\vec{x}^{(j-1)}\right) = \sum_{t=1}^{k} \lambda_t \left(\vec{v}_t{}^\top \vec{x}^{(j-1)}\right)^2 \vec{v}_t.$$

Claim: For almost all initial $\vec{x}^{(0)}$, the sequence $\left(\vec{x}^{(j)}/\|\vec{x}^{(j)}\|\right)_{j=1}^{\infty}$ converges *quadratically fast* to some $\vec{v}_t$.

Review: matrix power iteration

Recall matrix power iteration for matrix $M := \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \vec{v}_t^{\top}$:

Start with some $\vec{x}^{(0)}$, and for $j = 1, 2, \ldots$:

$$\vec{x}^{(i)} \ := \ M\vec{x}^{(j-1)} \ = \ \sum_{t=1}^{k} \lambda_t \ \left(\vec{v}_t^{\top} \vec{x}^{(j-1)}\right) \ \vec{v}_t.$$

i.e., component in $\vec{v}_t$ direction is scaled by λ_t.

Review: matrix power iteration

Recall matrix power iteration for matrix $M := \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \vec{v}_t{}^\top$:

Start with some $\vec{x}^{(0)}$, and for $j = 1, 2, \ldots$:

$$\vec{x}^{(i)} \ := \ M\vec{x}^{(j-1)} \ = \ \sum_{t=1}^{k} \lambda_t \ \left(\vec{v}_t{}^\top \vec{x}^{(j-1)}\right) \ \vec{v}_t.$$

i.e., component in $\vec{v}_t$ direction is scaled by λ_t.

If $\lambda_1 > \lambda_2 \geq \cdots$, then

$$\frac{\left(\vec{v}_1{}^\top \vec{x}^{(j)}\right)^2}{\sum_{t=1}^{k}\left(\vec{v}_t{}^\top \vec{x}^{(j)}\right)^2} \geq 1 - \left(\frac{\lambda_2}{\lambda_1}\right)^{2j}.$$

Converges *linearly* to $\vec{v}_1$ (assuming gap $\lambda_2/\lambda_1 < 1$).

Tensor power iteration convergence analysis

Let $c_t := \vec{v}_t{}^\top \vec{x}^{(0)}$ (initial component in $\vec{v}_t$ direction); assume WLOG

$$\lambda_1 |c_1| > \lambda_2 |c_2| \geq \lambda_3 |c_3| \geq \cdots .$$

Tensor power iteration convergence analysis

Let $c_t := \vec{v}_t{}^\top \vec{x}^{(0)}$ (initial component in $\vec{v}_t$ direction); assume WLOG

$$\lambda_1 |c_1| > \lambda_2 |c_2| \geq \lambda_3 |c_3| \geq \cdots .$$

Then

$$\vec{x}^{(1)} = \sum_{t=1}^{k} \lambda_t \left(\vec{v}_t{}^\top \vec{x}^{(0)}\right)^2 \vec{v}_t = \sum_{t=1}^{k} \lambda_t c_t^2 \ \vec{v}_t$$

i.e., component in $\vec{v}_t$ direction is squared then scaled by λ_t.

Tensor power iteration convergence analysis

Let $c_t := \vec{v}_t{}^\top \vec{x}^{(0)}$ (initial component in $\vec{v}_t$ direction); assume WLOG

$$\lambda_1 |c_1| > \lambda_2 |c_2| \geq \lambda_3 |c_3| \geq \cdots .$$

Then

$$\vec{x}^{(1)} = \sum_{t=1}^{k} \lambda_t \left(\vec{v}_t{}^\top \vec{x}^{(0)}\right)^2 \vec{v}_t = \sum_{t=1}^{k} \lambda_t c_t^2 \ \vec{v}_t$$

i.e., component in $\vec{v}_t$ direction is squared then scaled by λ_t.

By induction

$$\vec{x}^{(j)} = \sum_{t=1}^{k} {\lambda_t}^{2^j - 1} c_t^{2^j} \ \vec{v}_t,$$

so

$$\frac{\left(\vec{v}_1{}^\top \vec{x}^{(j)}\right)^2}{\sum_{t=1}^{k} \left(\vec{v}_t{}^\top \vec{x}^{(j)}\right)^2} \geq 1 - k \left(\frac{\lambda_1}{\max_{t \neq 1} \lambda_t}\right)^2 \left|\frac{\lambda_2 c_2}{\lambda_1 c_1}\right|^{2^{j+1}} .$$

Matrix vs. tensor power iteration

Matrix power iteration:

Tensor power iteration:

Matrix vs. tensor power iteration

Matrix power iteration:

1. Requires gap between largest and second-largest λ_t.
   (Property of the matrix only.)

Tensor power iteration:

1. Requires gap between largest and second-largest $\lambda_t|c_t|$.
   (Property of the tensor and initialization $\vec{x}^{(0)}$.)

Matrix vs. tensor power iteration

Matrix power iteration:

1. Requires gap between largest and second-largest λ_t.
   (Property of the matrix only.)
2. Converges to top $\vec{v}_t$.

Tensor power iteration:

1. Requires gap between largest and second-largest $\lambda_t|c_t|$.
   (Property of the tensor and initialization $\vec{x}^{(0)}$.)
2. Converges to $\vec{v}_t$ for which $\lambda_t|c_t| = \max!$ (could be any of them).

Matrix vs. tensor power iteration

Matrix power iteration:

1. Requires gap between largest and second-largest λ_t.
   (Property of the matrix only.)
2. Converges to top $\vec{v}_t$.
3. Linear convergence. (Need $O(\log(1/\epsilon))$ iterations.)

Tensor power iteration:

1. Requires gap between largest and second-largest $\lambda_t|c_t|$.
   (Property of the tensor and initialization $\vec{x}^{(0)}$.)
2. Converges to $\vec{v}_t$ for which $\lambda_t|c_t| = \max!$ (could be any of them).
3. Quadratic convergence. (Need $O(\log\log(1/\epsilon))$ iterations.)

Effect of errors in tensor power iterations

Suppose we are given $\widehat{T} := T + E$, with

$$T = \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t, \qquad \varepsilon := \sup_{\vec{x} \in \mathbb{S}^{k-1}} \|\phi_E(\vec{x})\|.$$

Quadratic operator $\phi_{\widehat{T}}$ with $\widehat{T}$:

$$\phi_{\widehat{T}}(\vec{x}) \quad = \quad \sum_{t=1}^{k} \lambda_t \ \left(\vec{v}_t{}^{\top} \vec{x}\right)^2 \vec{v}_t \quad + \quad \phi_E(\vec{x}).$$

Effect of errors in tensor power iterations

Suppose we are given $\widehat{T} := T + E$, with

$$T = \sum_{t=1}^{k} \lambda_t \ \vec{v}_t \otimes \vec{v}_t \otimes \vec{v}_t, \qquad \varepsilon := \sup_{\vec{x} \in \mathbb{S}^{k-1}} \|\phi_E(\vec{x})\|.$$

Quadratic operator $\phi_{\widehat{T}}$ with $\widehat{T}$:

$$\phi_{\widehat{T}}(\vec{x}) \quad = \quad \sum_{t=1}^{k} \lambda_t \ \left(\vec{v}_t{}^{\top} \vec{x}\right)^2 \vec{v}_t \quad + \quad \phi_E(\vec{x}).$$

Claim: If $\varepsilon \leq O(\frac{\min_t \lambda_t}{k})$ and $N \geq \Omega(\log(k) + \log\log \frac{\max_t \lambda_t}{\varepsilon})$, then N steps of tensor power iteration on $T + E$ (with good initialization) gives

$$\|\hat{v}_i - \vec{v}_i\| \leq O(\varepsilon/\lambda_i), \qquad |\hat{\lambda}_i - \lambda_i| \leq O(\varepsilon).$$

6. Recap and remarks

Recap and remarks

- Many latent variable models have low rank tensor forms

Recap and remarks

- Many latent variable models have low rank tensor forms
- Orthogonally diagonalizable tensors have very nice *identifiability*, *computational*, and *robustness* properties.

Recap and remarks

- Many latent variable models have low rank tensor forms
- Orthogonally diagonalizable tensors have very nice *identifiability*, *computational*, and *robustness* properties.
- Many analogues to matrix SVD, but also many important differences arising from non-linearity.

Recap and remarks

- Many latent variable models have low rank tensor forms
- Orthogonally diagonalizable tensors have very nice *identifiability*, *computational*, and *robustness* properties.
- Many analogues to matrix SVD, but also many important differences arising from non-linearity.
- Greedy algorithm for finding the decomposition can be rigorously analyzed and shown to be effective and efficient.

Recap and remarks

- Many latent variable models have low rank tensor forms
- Orthogonally diagonalizable tensors have very nice *identifiability*, *computational*, and *robustness* properties.
- Many analogues to matrix SVD, but also many important differences arising from non-linearity.
- Greedy algorithm for finding the decomposition can be rigorously analyzed and shown to be effective and efficient.

  Many variants possible (*e.g.*, initialization, deflation).

Recap and remarks

- Many latent variable models have low rank tensor forms
- Orthogonally diagonalizable tensors have very nice *identifiability*, *computational*, and *robustness* properties.
- Many analogues to matrix SVD, but also many important differences arising from non-linearity.
- Greedy algorithm for finding the decomposition can be rigorously analyzed and shown to be effective and efficient.

  Many variants possible (*e.g.*, initialization, deflation).
- Non-orthogonal (*e.g.*, overcomplete) CP decomposition is active area of research.

Questions?