# Tensor decompositions for Learning Latent Variable Models

A. Anandkumar R. Ge D. Hsu S. Kakade M. Telegarsky

November 18, 2024

# Motivation - An example latent variable model

Example of a latent variable model: Topic models for text data

The words of a document follow a distribution according to the document's latent topic.

Common method of estimating latent parameters: Expectation-Maximization

Problem: Bad local optima

Can we do better?

Use method-of-momemts.

Express *certain* moments as tensors with a “nice" structure.

Use properties of the “nice" tensors to derive algorithms with good convergence guarantees.

# Preliminaries: Tensor notation

A real $p$-way tensor, say $A \in \otimes_{i=1}^{p} \mathbb{R}^{d_i}$ is a *$p$-way array* of real numbers,

$$[A_{i_1, i_2, \ldots i_p} : i_1 \in [d_1], i_2 \in [d_2], \ldots i_p \in [d_p]]$$

$A$ represents a multilinear map;

Let $V_1 \in \mathbb{R}^{d_1 \times m_1}, V_2 \in \mathbb{R}^{d_2 \times m_2}, \ldots V_p \in \mathbb{R}^{d_p \times m_p}$, then $A(V_1, V_2, \ldots, V_p) \in \otimes_{i=1}^{p} \mathbb{R}^{m_i}$, such that,

$$[A(V_1, V_2, \ldots V_p)]_{i_1, i_2, \ldots i_p} := \sum_{j_1 \in [d_1], j_2 \in [d_2], \ldots, j_p \in [d_p]} A_{j_1, j_2, \ldots j_p} [V_1]_{j_1, i_1} [V_2]_{j_2, i_2} \ldots [V_p]_{j_p, i_p}$$

In this talk, we will only care about two-way tensors (matrices) and three-way tensors.

For $p = 2$,

$$A(V_1, V_2) = V_1^T A V_2$$

For $p = 3$,

$$A(V, I, I) = A \otimes_1 V; \quad A(I, V, I) = A \otimes_2 V; \quad A(I, I, V) = A \otimes_3 V$$
$$A(V_1, V_2, V_3) = A \otimes_1 V_1 \otimes_2 V_2 \otimes_3 V_3$$

# Preliminaries: Eigenvalues and Eigenvectors

For $p = 2$, i.e. *matrices* $A \in \mathbb{R}^{d \times d}$, the vector $u \in \mathbb{R}^d$ is an eigenvector of $A$ with corresponding *eigenvalue* $\lambda \in \mathbb{R}$, if,

$$Au = A(I, u) = \lambda u$$

For $p = 3$, i.e. $A \in \mathbb{R}^{d \times d \times d}$, the vector $u \in \mathbb{R}^d$ is an eigenvector of $A$ with corresponding *eigenvalue* $\lambda \in \mathbb{R}$, if,

$$A(I, u, u) = \lambda u$$

# Preliminaries: Symmetric and ODECO tensors

A $p$-way tensor, $A \in \otimes_p \mathbb{R}^d$, is symmetric if, for any permutation $\pi$,

$$A_{i_1,i_2,\ldots,i_p} = A_{\pi(i_1),\pi(i_2),\ldots,\pi(i_p)}$$

For $k \in \mathbb{N}$, $v_1, v_2, \ldots, v_k \in \mathbb{R}^d, \lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{R}$, tensors of the following form are symmetric,

$$A = \sum_{i=1}^{k} \lambda_i v_i \underbrace{\otimes \ldots \otimes}_{p \text{ times}} v_i$$

A orthogonally decomposable (ODECO) tensor is a tensor, which can be written in the following form,

$$A = \sum_{i=1}^{k} \lambda_i v_i \underbrace{\otimes \ldots \otimes}_{p \text{ times}} v_i$$

where $\{v_i \in \mathbb{R}^d; i \in [k]\}$ for $k \leq d$ are orthogonal vectors.

# Bag of Words Model

Each *document* is assumed to have a *single topic*; and the *distribution of words* in the document is based on the *topic*.

The words in a document are assumed to be *drawn independently*, given the topic.

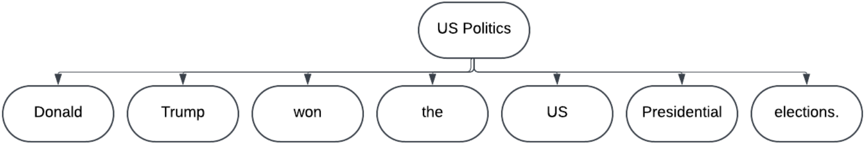



Figure: Bag of words illustration

# Bag of words model

Assume there are,

$k$ distinct topics.

$d$ distinct words in the vocabulary

$l \geq 3$ words in each document

# Bag of words model

The generative process is as follows,

Let the topic of a document be $h$ and the words in the document be $(x_1, x_2, \ldots, x_l); x_j \in \mathbb{R}^d$, where if the $t$-th word in the document is the $i$-th word in the dictionary, we set,

$$x_t = e_i$$

The topic $h \sim \text{Dir}(w)$; $w \in \Delta^{k-1}$, i.e. $w \in [0,1]^k, \sum_{i=1}^{k} w_i = 1$,

$$P(h = j) = w_j; \quad j \in [k]$$

Given the topic $h$, the words $(x_1, x_2, \ldots, x_l)$ are drawn *independently* from $\text{Dir}(\mu_h)$, where $\mu_h \in \Delta^{d-1}$.

# Method of Moments: Symmetric Tensors

Consider the following *two-way* and *three-way* moments.

$$M_2 := \mathbb{E}[x_1 \otimes x_2]$$
$$M_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3]$$

then, by *conditional independence* given topic

$$M_2 = \sum_{i=1}^{k} P(h=i)\mathbb{E}[x_1 \otimes x_2 \| h=i]$$
$$\sum_{i=1}^{k} w_i \mathbb{E}[x_1 \| h=i] \otimes \mathbb{E}[x_2 \| h=i] = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i$$

Similarly,

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

# Reduction to ODECO tensors

We have that the two-way and three-way moments are *symmetric* tensors.

Symmetric tensors of our form following a non-degeneracy condition can be reduced to ODECO tensors.

**Condition (Non-degeneracy)**

*The vectors* $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ *are linearly independent and the scalars* $w_1, w_2, \ldots w_k > 0$ *are strictly positive.*

We will see later that this reduction to ODECO tensors will be convenient for many reasons.

# Reduction to ODECO tensors

The idea is to *whiten* the two-way moment moment, $M_2$.

By the *non-degeneracy condition*, there exists a *whitening* matrix $W$ such that,

$$M_2(W, W) = W^T M_2 W = I$$

Define,

$$\tilde{\mu}_i := \sqrt{w_i} W^T \mu_1$$

Then,

$$M_2(W, W) = \sum_{i=1}^{k} \tilde{\mu}_i \tilde{\mu}_i^T = I$$

Then observe that the projection of the three-way moment, $M_3$, using the same *whitening* is an ODECO tensor.

$$\begin{aligned}\tilde{M}_3 &:= M_3(W, W, W) \in \mathbb{R}^{k \times k \times k} \\ &= \sum_{i=1}^{k} w_i \left(W^T \mu_i\right)^{\otimes 3} = \sum_{i=1}^{k} \frac{1}{\sqrt{w_i}} \tilde{\mu_i}^{\otimes 3}\end{aligned}$$

# Identifiability - Kruskal's Lemma

Let,

$$U := \begin{pmatrix} | & | & & | \\ \tilde{\mu}_1 & \tilde{\mu}_2 & \dots & \tilde{\mu}_k \\ | & | & & | \end{pmatrix}$$

Then the CP decomposition of $\tilde{M}_3 = [U,\ U,\ U]_k$.

The Kruskal rank of the matrix U is defined as the maximum number, $k_U$, such that any subset of columns of $U$ of cardinality $k_U$ are linearly independent.

Since $U$ has *orthogonal columns*, $k_U = k$.

## Lemma (Kruskal, 1977)

*Let* $T = [A,\ B,\ C]_R$ *and suppose that*

$$2R + 2 \leq k_A + k_B + k_C$$

*then the CPD of* $T$ *is unique.*

By Kruskal's lemma, our model is identifiable if

$$\begin{aligned} 2k + 2 &\leq 3k \\ \implies k &\geq 2 \end{aligned}$$

Thus, as long as we have *two or more latent topics*, we are solving an identifiable problem.

## Analogy between eigenvector of matrices and three-way tensors

Consider $\lambda_1, \lambda_2, \ldots, \lambda_d \geq 0$, and $v_1, v_2, \ldots v_d \in \mathbb{R}^d$ an *orthogonal* basis.

$M \coloneqq \sum_{i=1}^{k} \lambda_i v_i \otimes v_i$ is a *positive-semidefinite* matrix.

Consider $u \in \mathbb{R}^d; u = \sum_{i=1}^{d} c_i v_i$. Then,

$$M(I, u) = \sum_{i=1}^{d} \lambda_i c_i v_i$$

Thus, the operator $u \rightarrow M(I, u)$ scales the projections of $u$ along the eigenvectors, $v_i$, by their corresponding eigenvalue, $\lambda_i$.

It can be easily checked that the operator, $u \rightarrow M(I, u)$, is linear, i.e.,

$$M(I, u + v) = M(I, u) + M(I, v)$$

How can we interpret eigenvalues and eigenvectors of three-way tensors?

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

It is clear that $v_1, v_2, \ldots, v_d$ are *eigenvectors* with *eigenvalues* $\lambda_1, \lambda_2, \ldots, \lambda_d$ since they satisfy $M(I, u, u) = \lambda u$.

Consider the operator $u \rightarrow T(I, u, u)$, recall that $u = \sum_{i=1}^{d} c_i v_i$,

$$T(I, u, u) = \sum_{i=1}^{d} \lambda_i c_i^2 v_i$$

Thus, the operator $T(I, u, u)$ scales the projection along each eigenvector, $v_i$, by $\lambda_i c_i$ where $\lambda_i$ is the corresponding eigenvalue and $c_i$ is the corresponding projection itself.

# Eigenvectors and eigenvalues of ODECO tensors

If $\lambda_i$s are distinct, are $v_i$s the only eigenvectors?

The answer is no. Let's see why.

$u = \sum_{i=1}^{d} c_i v_i$ is an eigenvector of $T$ as long as all the *nonzero* projections along $v_i$s with *nonzero* $\lambda_i$s are scaled evenly, i.e. for some constant $c \in \mathbb{R}$,

$$\forall k \in [d] \text{ s.t. } \lambda_k, c_k \neq 0,\ \lambda_k c_k = c$$

Thus we may select any subset $S \subseteq [d]$ such that $\forall i \in S, \lambda_i \neq 0$, and set,

$$c_i = \begin{cases} \frac{c}{\lambda_i}, & \text{if } i \in S, \\ 0, & \text{if } i \notin S. \end{cases}$$

and $u = \sum_{i=1}^{d} c_i v_i$.

This phenomenon does not happen in the *matrix case*, as then for the scaling of any subset of projections, $c_i;\ i \in S$, to match, we would need $\lambda_i = c\ \forall i \in S$, where $c \in \mathbb{R}$ is a constant.

Without loss of generality, for some $k \in [d]$, let $\lambda_1, \lambda_2, \ldots \lambda_k > 0;\ \lambda_i = 0\ \forall i > k, i \leq d$.

Among all eigenvectors of T, $v_1, v_2, \ldots v_k$ are said to be the robust eigenvectors, defined later.

Another noteworthy observation is that the operator $u \to T(I, u, u)$ is *not linear*. For $x, y \in \mathbb{R}^d$,

$$T(I, x+y, x+y) \neq T(I, x, x) + T(I, y, y)$$

# Power iteration

Consider a tensor $T$, which has an orthogonal decomposition of the form,

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

where $v_1 v_2, \ldots, v_k$ are orthogonal and $\lambda_1, \lambda_2, \ldots, \lambda_k > 0$.
Define the power iteration on $T$ starting from $\theta \in \mathbb{R}^d$, as, repeated iterations of,

$$\theta \to \frac{T(I, \theta, \theta)}{\|T(I, \theta, \theta)\|}$$

(Characterization of robust eigenvectors) A vector $u$ is a *robust eigenvector* A vector $u$ is a robust eigenvector of $T$ if there exists an $\epsilon > 0$ such that for all $\theta \in \{u' \in \mathbb{R}^d : \|u' - u\| \leq \epsilon\}$, the power iteration starting from $\theta$ converges to $u$.

**Theorem (Power iteration)**

*The set $\theta \in \mathbb{R}^d$ which does not converge to some $v_i$ under the power iteration starting at $\theta$ has measure zero.*

*The set of robust eigenvectors is equal to $\{v_1, v_2, \ldots, v_k\}$.*

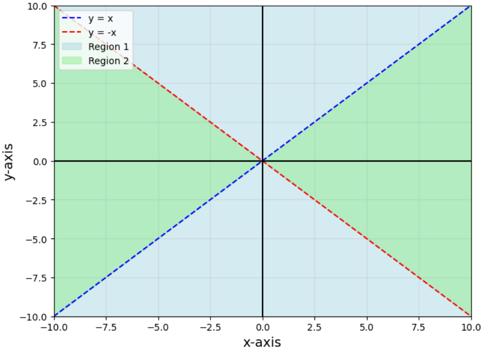
## Power iteration: Proof(Intuition)

For easier visulaization, let $d = k = 2, v_1 = e_1, v_2 = e_2, \lambda_1 = \lambda_2 = 1$.

For $\theta$ in the blue region, the power iteration converges to $e_1$.

For $\theta$ in the green region, the power iteration converges to $e_2$.

For $\theta$ on the dashed lines, which is a *measure zero* set, the power iteration does not converge.




Power iteration visualization

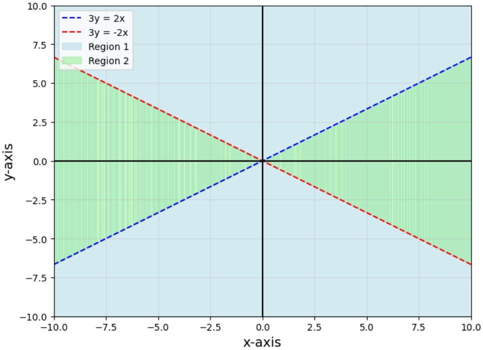
## Power iteration: Proof(Intuition)

Another example, let $d = k = 2, v_1 = e_1, v_2 = e_2, \lambda_1 = 2, \lambda_2 = 3.$

For $\theta$ in the blue region, the power iteration converges to $e_1$.

For $\theta$ in the green region, the power iteration converges to $e_2$.

For $\theta$ on the dashed lines, which is a *measure zero* set, the power iteration does not converge.




Power iteration visualization

# Power iteration: Proof(Intuition)

For $c_1 = \langle \theta, e_1 \rangle, c_2 = \langle \theta, e_2 \rangle$,

For $\theta$ in the blue region,

$|\lambda_1 c_1| > |\lambda_2 c_2|$

Hence, the size of the projection along $e_1$ grows in comparison to the projection along $e_2$.

The power iteration converges to $e_1$.

For $\theta$ in the green region,

$|\lambda_2 c_2| > |\lambda_1 c_1|$

Hence, the size of the projection along $e_2$ grows in comparison to the projection along $e_1$.

The power iteration converges to $e_2$.

# Power iteration: Proof(Intuition)

The only points of convergence of the power iteration are $e_1$ and $e_2$ since,

Let $u$ be a convergence point of the power iteration.

Then there exists an $\epsilon > 0$ such than, for all $\theta \in \{u' \in \mathbb{R}^d : \|u' - u\| \leq \epsilon\}$, the power iteration converges to $u$.

Such points of convergence by definition, are the robust eigenvectors $e_1$ and $e_2$.

The proof follows similarly for higher dimensions, i.e. $d > 2$.

# Power iteration: Convergence

## Theorem (Power iterations convergence)

*Consider orthogonally decomposable* $T = \sum_{i=1}^{k} \lambda_i v_i \otimes v_i \otimes v_i$*, where* $v_1 v_2, \ldots, v_k$ *are orthogonal and* $\lambda_1, \lambda_2, \ldots, \lambda_k > 0$*.*

*For* $\theta_0 \in \mathbb{R}^d$*, define the projections* $c_i := \langle \theta_0, v_i \rangle$ *let the set of numbers* $\{|\lambda_1 c_1|, |\lambda_2 c_2\rangle|, \ldots, |\lambda_k c_k\rangle|\}$ *have a unique largest value. Wlog,* $|\lambda_1 c_1| > |\lambda_2 c_2|$ *are the two largest values.*

*Denote the outcome of the power iteration starting from* $\theta_0$ *after* $t$ *steps as* $\theta_t$*, i.e.*

$$\theta_t := \frac{T(I, \theta_{t-1}, \theta_{t-1})}{\|T(I, \theta_{t-1}, \theta_{t-1})\|}$$

*Then,*

$$\|v_1 - \theta_t\|^2 \leq C \cdot \left| \frac{\lambda_2 c_2}{\lambda_1 c_1} \right|^{2^t + 1}$$

*where* $C = 2\lambda_1^2 \sum_{i=2}^{k} \lambda_i^{-2}$ *is a constant.*

To obtain all robust eigenvectors, we may simply proceed iteratively using deflation, executing the power method on $T - \sum_j \lambda_j v_j \otimes v_j \otimes v_j$ after having obtained the robust eigenvector/eigenvalue pairs $\{(\lambda_j, v_j)\}$.

# Power iteration: Convergence (Interpretation)

Here, $|\lambda_1 c_1| > |\lambda_2 c_2|$ are the two *largest* scalings applied to the projections on the robust eigenvectors.

The error scales quadratically in the ratio of the scalings, $\left|\frac{\lambda_2 c_2}{\lambda_1 c_1}\right|$.

# Power iteration: Convergence (Proof)

## Orthogonal Complement

Let $\bar{\theta}_0, \bar{\theta}_1, \ldots, \bar{\theta}_t$ be the sequence such that, $\bar{\theta}_t \coloneqq T(I, \bar{\theta}_{t-1}, \bar{\theta}_{t-1})$.

(a) $\theta_t = \frac{\bar{\theta}_{t-1}}{\|\bar{\theta}_{t-1}\|}$ (b) $\bar{\theta}_t = \sum_{i=1}^{d} \lambda_i^{2^t-1} c_i^{2^t} v_i$

$$1 - \langle v_1, \theta_t \rangle^2 = 1 - \frac{\langle v_1, \bar{\theta}_t \rangle^2}{\|\bar{\theta}_t\|^2} = 1 - \frac{\lambda_1^{2^t+1} c_1^{2^t+1}}{\sum_{i=1}^{d} \lambda_i^{2^t+1} c_i^{2^t+1}} = \frac{\sum_{i=2}^{d} \lambda_i^{2^t+1} c_i^{2^t+1}}{\sum_{i=1}^{d} \lambda_i^{2^t+1} c_i^{2^t+1}}$$

$$\leq \lambda_1^2 \sum_{i=2}^{k} \lambda_i^{-2} \left| \frac{\lambda_2 c_2}{\lambda_1 c_1} \right|^{2^t+1}$$

## Pythagoras theorem

Since $\lambda_1 > 0, 0 < \langle v_1, \theta_t \rangle < 1$, we have,

$$\|v_1 - \theta_t\|^2 = 2(1 - \langle v_1, \theta_t \rangle) \leq 2(1 - \langle v_1, \theta_t \rangle^2)$$

and the result follows.

# Tensor perturbation problem specification

$T \in \mathbb{R}^{d \times d \times d}$ is a symmetric tensor with the orthogonal decomposition,

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

, where each $\lambda_i > 0$, and $\{v_1, v_2, \ldots, v_d\}$ form an orthogonal basis.

$$\hat{T} = T + E$$

, is the perturbed tensor and the operator norm of $E$, $\|E\| \leq \epsilon$ for some $\epsilon > 0$.

The operator norm of a three-way tensor is defined as,

$$\|E\| \coloneqq \sup_{\|\theta\|=1} |E(\theta, \theta, \theta)|$$

# Power iterations with Noise: Algorithm

**Algorithm 1** Robust tensor power method

**input** symmetric tensor $\tilde{T} \in \mathbb{R}^{k \times k \times k}$, number of iterations $L$, $N$.
**output** the estimated eigenvector/eigenvalue pair; the deflated tensor.

**for** $\tau = 1$ to $L$ **do**
  Draw $\theta_0^{(\tau)}$ uniformly at random from the unit sphere in $\mathbb{R}^k$.
  **for** $t = 1$ to $N$ **do**
    Compute power iteration update

$$\theta_t^{(\tau)} := \frac{\tilde{T}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})}{\|\tilde{T}(I, \theta_{t-1}^{(\tau)}, \theta_{t-1}^{(\tau)})\|} \tag{7}$$

  **end for**
**end for**
Let $\tau^* := \arg\max_{\tau \in [L]} \{\tilde{T}(\theta_N^{(\tau)}, \theta_N^{(\tau)}, \theta_N^{(\tau)})\}$.
Do $N$ power iteration updates (7) starting from $\theta_N^{(\tau^*)}$ to obtain $\hat{\theta}$, and set $\hat{\lambda} := \tilde{T}(\hat{\theta}, \hat{\theta}, \hat{\theta})$.
**return** the estimated eigenvector/eigenvalue pair $(\hat{\theta}, \hat{\lambda})$; the deflated tensor $\tilde{T} - \hat{\lambda}\, \hat{\theta}^{\otimes 3}$.

Figure: Power iteration with noise

# Tensor perturbation problem specification

$T \in \mathbb{R}^{d \times d \times d}$ is a symmetric tensor with the orthogonal decomposition,

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

, where each $\lambda_i > 0$, and $\{v_1, v_2, \ldots, v_d\}$ form an orthogonal basis.

$$\hat{T} = T + E$$

, is the perturbed tensor and the operator norm of $E$, $\|E\| \leq \epsilon$ for some $\epsilon > 0$.

The operator norm of a three-way tensor is defined as,

$$\|E\| \coloneqq \sup_{\|\theta\|=1} |E(\theta, \theta, \theta)|$$

# Power iterations with Noise: Informal Convergence Theorem

Define $\lambda_{\min} := \min\{\lambda_i : i \in [k]\}$, $\lambda_{\max} := \max\{\lambda_i : i \in [k]\}$.

## Theorem (Informal)

*There exist constants $C_1, C_2, C_3 > 0$ such that the following holds. Pick any $\eta \in (0,1)$, and suppose,*

$$\epsilon \leq C_1 \cdot \frac{\lambda_{\min}}{d}, \quad N \geq C_2 \cdot \left(\log d + \log\log\left(\frac{\lambda_{\max}}{\epsilon}\right)\right)$$

*and $L =$ poly$(k)\log(1/\eta)$. Suppose the algorithm is iteratively called $d$ times, deflating $\hat{T}$ upon identifying each eigenvector. Let $(\hat{v}_1, \hat{\lambda}_1), (\hat{v}_2, \hat{\lambda}_2), \ldots (\hat{v}_d, \hat{\lambda}_d)$, then*

$$\|v_{\pi(j)} - \hat{v}_j\| \leq 8\epsilon/\lambda_{\pi(j)}, \quad |\lambda_{\pi(j)} - \hat{\lambda}_j| le 5\epsilon \quad \forall\, j \in [d]$$

$$\left\| T - \sum_{j=1}^{d} \hat{\lambda}_j \hat{v}_j \otimes \hat{v}_j \otimes \hat{v}_j \right\| \leq 55\epsilon$$

# Power iterations with Noise: Proof ideas

The proof of this theorem is not as straightforward as the noiseless case.

Here, $E$ is a symmetric but not necessarily orthogonally decomposable tensor. Errors accrue since we lose the orthogonality structure, in particular in the power iteration and deflation steps. Proof of this theorem requires a careful and systematic analysis of the accumulating errors.

Furthermore, conditions for a "good" initialization need to be studied.

Observe that for this theorem, we need the *stronger* condition, k=d, since for a vector $u \in \mathbb{R}^d$, we may not be able to restrict $E(I, u, u)$ to the linear subspace spanned by eigenvectors of $T$, $\{v_1, v_2, \ldots, v_k\}$ which have *positive* eigenvalues.

# Advantages of this method

The method attempts to avoid the *bad local optima* problem of Expectation-Maximization-style algorithms by parameterizing the problem carefully and identifying a ODECO tensor structure in the moments.

Due to *unique identifiability* properties of such tensors, it is hoped that bad local optima may be evaded with good initialization.

THANK YOU