TENSOR FACTORIZATIONS METHODS : Lecture 3.

I) Tensor Methods for Topic models, Multiview models, GMM. Applications.

II) Matrix Power Method.

III) Tensor Power Method.

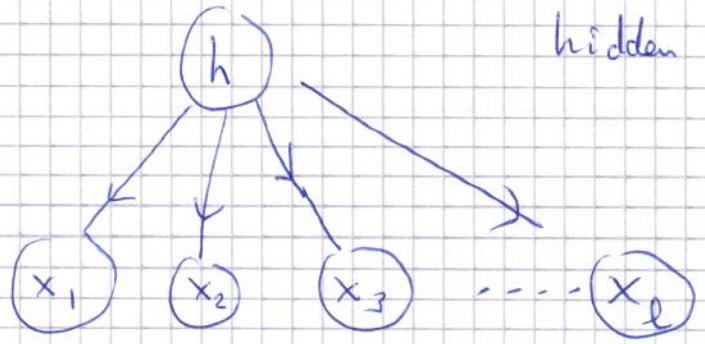
I.a) Single topic models for Documents.

Following is a simplified model for documents. A documents is a collection of ℓ words $x_1, x_2 \cdots x_\ell$

x_1	x_2	x_3
x_4	x_5	x_6
x_7	x_8	x_9
x_{10}	x_{11}	x_{12}
x_{13}	x_{14}	x_{15}

where each words is a random variable taking values in some dictionary of size D. So the fifth word of the document, say x_5, can take D possible values. We assume that each document is about a single topic. There are K possible Topics indexed by $h \in \{1, \ldots, K\}$ (e.g sport, music, politics, cinema).

The probabilistic model of documents is a belief network of the form :



where all conditional probabilities are identical $p(x|h)$.

So the prob distr for a document is

$$p(x_1, x_2, \dots x_\ell, h) = p(h) \prod_{i=1}^{\ell} p(x_i | h)$$

Given a Topic words are i.i.d.

We use the following formalism :

- $p(h=i) = w_i$ = weight of topic $i \in \{1 \dots K\}$.
- $\underline{x} \in$ Dictionary of D words $= \{\underline{e}_1, \underline{e}_2, \dots \underline{e}_D\}$ where each $\underline{e}_\alpha = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ ith coordinate. So we identify words with canonical vectors of basis of $\mathbb{R}^D$.

We set $p(\underline{x} | i) = \underline{\mu}_i$ = vector of probabilities of words given a topic i.

Remark : words are exchangeable random variables in the sense that $p(x_1, x_2 \dots x_\ell) = \sum_{h=1}^{K} p(h) \prod_{i=1}^{\ell} p(x_i | h)$ is permutation invariant.

Our problem: Given frequencies of occurrences of words in documents

$$\hat{Pr}(x_1) \quad ; \quad \hat{Pr}(x_1, x_2) \quad ; \quad \hat{Pr}(x_1, x_2, x_3) \quad \text{etc}$$

we want to determine the parameters

$$\begin{cases} (w_i)_{i=1\dots K} & \text{weights of Topics} \\ \underline{\mu}_i & \text{weights of words given a Topic } i \in \{1 \dots K\}. \end{cases}$$

This is done by matching empirical moments to Theoretical moments.

Theoretical Moments of our probabilistic model:

$$\mathbb{E}(\underline{x}_1) = \sum_{\alpha=1}^{D} \underbrace{Pr(\underline{x}_1 = \underline{e}_\alpha)}_{\text{prob of word } \underline{e}_\alpha} \cdot \underline{e}_\alpha$$

($\underline{e}_\alpha$: possible value of word in Dict.)

$$= \sum_{\alpha=1}^{D} \sum_{i=1}^{K} Pr(\underline{x}_1 = \underline{e}_\alpha \mid i)\, w_i\, \underline{e}_\alpha$$

$$= \sum_{i=1}^{K} \left\{ \sum_{\alpha=1}^{D} \underbrace{Pr(\underline{x}_1 = e_\alpha \mid i)}_{\mu_i^\alpha} \underline{e}_\alpha \right\} w_i$$

$$\Rightarrow \boxed{\mathbb{E}(\underline{x}_1) = \sum_{i=1}^{K} \underline{\mu}_i\, w_i}$$

Note that this formula is the same for the first, second, third etc word of the document.

Similarly for second and third order moments:

$$\mathbb{E}(\underline{X}_1 \otimes \underline{X}_2) = \sum_{\alpha=1}^{D}\sum_{\beta=1}^{D} \overbrace{pr(\underline{X}_1 = \underline{e}_\alpha, \underline{X}_2 = \underline{e}_\beta)}^{M^{\alpha\beta}} \underline{e}_\alpha \otimes \underline{e}_\beta$$

$$= \sum_{\alpha,\beta} \sum_{i} w_i \underbrace{pr(\underline{X}_1 = \underline{e}_\alpha, X_2 = e_\beta | i)}_{pr(X_1 = e_\alpha | i)\, pr(X_2 = e_\beta | i)} \underline{e}_\alpha \otimes \underline{e}_\beta .$$

$$= \sum_{i=1}^{K} w_i \left\{\sum_\alpha p(X_1 = e_\alpha | i)\right\}\left\{\sum_\beta p(X_2 = e_\beta | i)\right\}$$

$$= \sum_{i=1}^{K} w_i \, \underline{\mu}_i \otimes \underline{\mu}_i$$

$$\Rightarrow \boxed{\mathbb{E}(\underline{X}_1 \otimes \underline{X}_2) = \sum_{i=1}^{K} w_i \, \underline{\mu}_i \otimes \underline{\mu}_i .}$$

Also,

$$\boxed{\mathbb{E}(\underline{X}_1 \otimes \underline{X}_2 \otimes \underline{X}_3) = \sum_{i=1}^{K} w_i \, \underline{\mu}_i \otimes \underline{\mu}_i \otimes \underline{\mu}_i}$$

$$= \sum_{\alpha,\beta,\gamma} pr(X_1 = e_\alpha, X_2 = e_\beta, X_3 = e_\gamma)\, \underline{e}_\alpha \otimes \underline{e}_\beta \otimes \underline{e}_\gamma$$

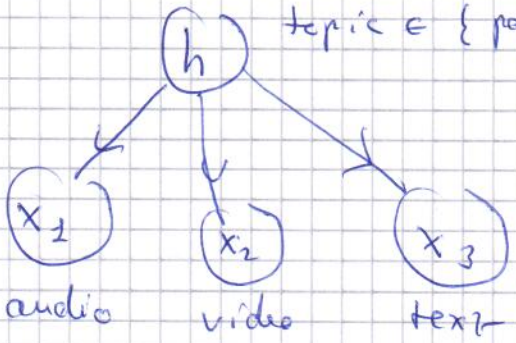
Same formulas are valid for any pair and/or triple of words.

The analysis of documents can be done by making decomposition of empirical Tensors

$$Pr_{emp}(\underbrace{X_1 = e_\alpha}_{= e_\beta} X_2) = M_{\alpha}^{\alpha\beta} \; ; \; Pr_{emp}(X_1 = e_\alpha, X_2 = e_\beta, X_3 = e_\gamma) = \overline{M}_{\alpha}^{\alpha\beta\gamma} .$$

I. b) Multiview Models.

Very similar but slightly more general than previous model. Suppose we have a hidden Topic $h \in \{1 \dots K\}$ for a set of K possible Topics (politics, cinema, sport, travel) and suppose we have many views (say: audio, video, text) of a topic:



$$P_1(\underline{x}_1 | h) \qquad P_2(\underline{x}_2 | h) \qquad P_3(\underline{x}_3 | h).$$

Each $\underline{x}_i \in \{\underline{e}_1, \underline{e}_2, \dots \underline{e}_{D_i}\}$ D_i possible values encoded in canonical basis vectors of $\mathbb{R}^{D_i}$. $i = 1, 2, 3$

Joint p.d:

$$P(\underline{x}_1, \underline{x}_2, \underline{x}_3, h) = P(h) \underbrace{P_1(\underline{x}_1 | h) P_2(\underline{x}_2 | h) P_3(\underline{x}_3 | h)}$$

Now P_1, P_2, P_3 can be unequal (more general than before) so we do not have exchangeability.

- Again it is easy to see :

$$\mathbb{E}(\underline{x}_i) = \sum_{\alpha=1}^{D_i} pr(\underline{x}_i = \underline{e}_\alpha)\, \underline{e}_\alpha .$$

$$= \text{vector}\left(pr(\underline{x}_i = e_\alpha) \right)_{\alpha=1}^{D_i} .$$

$$\mathbb{E}(\underline{x}_i \otimes \underline{x}_j) = \sum_{\alpha=1}^{D_i} \sum_{\beta=1}^{D_j} pr(\underline{x}_i = e_\alpha, \underline{x}_j = \underline{e}_\beta)\, \underline{e}_\alpha \otimes e_\beta$$

$$= \text{Tensor a Matrix } \mu^{\alpha\beta} = pr(x_i = e_\alpha, x_j = e_\beta)$$

↑ for each pair $(i,j) \in \{1,2,3\}^2$.

$$\mathbb{E}(\underline{x}_1 \otimes \underline{x}_2 \otimes x_3) = \sum_{\alpha\beta\gamma}^{D_1 D_2 D_3} pr(x_1 = e_\alpha, x_2 = e_\beta, x_3 = e_\gamma) \cdot \underline{e}_\alpha \otimes \underline{e}_\beta \otimes \underline{e}_\gamma$$

= Tensor or 3d array of numbers

$$pr(\underline{x}_1 = e_\alpha ; x_2 = \underline{e}_\beta ; x_3 = e_\gamma) .$$

- We have the decompositions:

$$\begin{cases} \mathbb{E}(\underline{x}_1) = \sum_{i=1}^{K} w_i \underline{\mu}_i^{(1)} & \underline{\mu}_i^{(1)} = pr(\underline{x} | i) \\ \mathbb{E}(\underline{x}_1 \otimes \underline{x}_2) = \sum_{i=1}^{K} w_i \underline{\mu}_i^{(1)} \otimes \underline{\mu}_i^{(2)} \\ \mathbb{E}(\underline{x}_1 \otimes \underline{x}_2) = \sum_{i=1}^{K} w_i \underline{\mu}_i^{(1)} \otimes \underline{\mu}_i^{(2)} \otimes \underline{\mu}_i^{(3)} . \end{cases}$$

I, C) Gaussian Mixture Models.

Let $p(\underline{x}) = \sum_{i=1}^{K} w_i \frac{e^{-\frac{\|\underline{x} - \underline{a}_i\|^2}{2\sigma^2}}}{(2\pi\sigma^2)^{D/2}}$

= convex combination of spherical Gaussian clusters around $\underline{a}_1 \dots \underline{a}_K$.

In the exercices you show that:

$$\mathbb{E}(\underline{x}) = \sum_{i=1}^{K} w_i \underline{a}_i \equiv \underline{m}.$$

$$\mathbb{E}(\underline{x} \otimes \underline{x}) = \sigma^2 I_{D\times D} + \sum_{i=1}^{K} w_i \underline{a}_i \otimes \underline{a}_i$$

$$\mathbb{E}(\underline{x} \otimes \underline{x} \otimes \underline{x}) = \sum_{i=1}^{K} w_i \underline{a}_i \otimes \underline{a}_i \otimes \underline{a}_i + \sigma^2 \sum_{i=1}^{K} (\underline{m} \otimes \underline{e}_i \otimes \underline{e}_i + \underline{e}_i \otimes \underline{m} \otimes \underline{e}_i + \underline{e}_i \otimes \underline{e}_i \otimes \underline{m}).$$

If we have a way to estimate $\underline{m}$ & σ^2 then we can determine w_i & $\underline{a}_i$'s by looking at decomposing of the Tensor $\mathbb{E}(x^\alpha x^\beta x^\gamma) = T^{\alpha\beta\gamma}$.

Remark: $\underline{\mu}$ is estimated by empirical mean. To estimate σ^2 we can compute smallest e.v of $\mathbb{E}(x^\alpha x^\beta)$ for the case when $K < D$ so we know the matrix $\sum_{i=1}^{K} w_i \underline{a}_i \underline{a}_i^T$ is not full rank and has a zero e.v.

For the GMM and single topic document model we see that the Tensor decomposition is special because it involves only terms $\underline{\mu}_i \otimes \underline{\mu}_i \otimes \underline{\mu}_i$.

The "power method" has been investigated in detail for this class of tensors. Here we give a brief introduction to the basics of the power method.

II. Matrix power method.

Let M a symmetric real matrix $N \times N$. It has a set of real eigenvalues & orthonormal eigenvectors

$$M \underline{v}_i = \lambda_i \underline{v}_i \qquad \lambda_i \in \mathbb{R}, \quad \underline{v}_i^T \cdot \underline{v}_j = \delta_{ij} = \begin{cases} 1 & (i=j) \\ 0 & (i \neq j) \end{cases}$$

The spectral decomposition of M is

$$M = \sum_{i=1}^{N} \lambda_i \underline{v}_i \underline{v}_i^T = \sum_{i=1}^{N} \lambda_i \underline{v}_i \otimes \underline{v}_i .$$

and we assume $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq \lambda_N$.

The power method is an iterative method that allows to find eigenvalues and eigenvectors assuming No degeneracy.

$\longrightarrow$

Power Method for top eigenvalue / eigenvector:

- Choose initial unit vector $\underline{x}^{(0)}$ at time $t=0$ s.t $\underline{x}^{(0)}$ is NOT orthogonal to $\underline{v}_1$.

- Iterate $$\underline{x}^{(t)} = \frac{M\,\underline{x}^{(t-1)}}{\|M\,\underline{x}^{(t-1)}\|_2} \qquad t \geq 1.$$

Lemma. Assume $\lambda_1 > \lambda_2$ (so in particular $\underline{v}_1$ is unique). Then

$$\begin{cases} \underline{x}^{(t)} \to \underline{v}_1 \text{ in } L_2 \text{ norm as } t \to +\infty \\ \underline{x}^{(t)} M\, \underline{x}^{(t)} \to \lambda_1 \quad \text{as } t \to +\infty \end{cases}$$

and the convergence rate is $\simeq O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^t\right)$. exp fast.

Remark: In very high dimensions $D \gg 1$ it is not so easy to choose $\underline{x}^{(0)}$ that is "far from orthogonal to $\underline{v}_1$" (even for a random choice!) so there is a big constant C_D in front of $O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^t\right)$ which make initial convergence rate slow.

Proof of Lemma. First we remark that

$$\underline{x}^{t-1} = \frac{M \underline{x}^{t-2}}{\|M \underline{x}^{t-2}\|_2} \Rightarrow \underline{x}^{t} = \frac{M x^{t-1}}{\|M x^{t-1}\|_2}$$

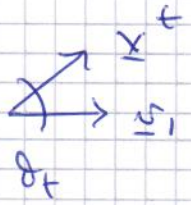
$$= \frac{M^2 x^{t-2}}{\|M^2 x^{t-2}\|_2} \cdot \frac{1}{\left\| \frac{M^2 x^{t-2}}{\|M x^{t-2}\|} \right\|}$$

$$\Rightarrow \quad \underline{x}^{t} = \frac{M^2 x^{t-2}}{\|M^2 x^{t-2}\|} \Rightarrow \boxed{x^t = \frac{M^t x^0}{\|M^t x^0\|_2}}$$

From the spectral decomposition we get :

$$x^t = \frac{\sum_{i=1}^{N} \lambda_i^t \, \underline{v}_i \, (\underline{v}_i^T \cdot \underline{x}^0)}{\left\{ \sum_{i=1}^{N} \lambda_i^{2t} \, (\underline{v}_i^T \cdot \underline{x}^{(0)})^2 \right\}^{1/2}}$$

We look at the angle



$$(\cos \vartheta_t)^2 = \underline{v}_1^T \cdot \underline{x}^t = \frac{\lambda_1^{2t} \, (\underline{v}_1^T \cdot \underline{x}^0)^2}{\sum_{i=1}^{N} \lambda_i^{2t} \, (\underline{v}_i^T \cdot x^0)^2}$$

$$\Rightarrow \quad (\cos \vartheta_t)^2 = \frac{1}{1 + \sum_{i=2}^{N} \left(\frac{d_i}{d_1} \right)^{2t} \frac{(\underline{v}_i^T \cdot x^0)^2}{(\underline{v}_1^T \cdot x^0)^2}}$$

Since $\left(\frac{\lambda_i}{\lambda_1}\right)^{2t} \leq \left(\frac{\lambda_2}{\lambda_1}\right)^{2t}$ for $i \geq 2$ we have

$$\sum_{i=2}^{N} \left(\frac{\lambda_i}{\lambda_1}\right)^{2t} \left(\frac{v_i^T \cdot x^0}{v_1^T \cdot x^0}\right)^2 \leq \left(\frac{\lambda_2}{\lambda_1}\right)^{2t} \frac{\sum_{i=2}^{N} (v_i^T x^0)^2}{(v_1^T \cdot x^0)^2}$$

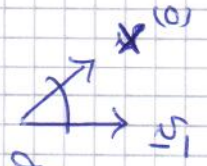
and since $\sum_{i=1}^{N} (v_i^T \cdot x^0)^2 = 1 \Rightarrow \sum_{i=2}^{N} (v_i^T \cdot x^0)^2 = 1 - v_1^T \cdot x^0$

we finely get

$$(\cos\vartheta_t)^2 \geq \frac{1}{1 + \left(\frac{\lambda_2}{\lambda_1}\right)^{2t} \frac{1-(\cos\vartheta_0)^2}{(\cos\vartheta_0)^2}}$$

$$\Rightarrow \quad 1 - (\cos\vartheta_t)^2 \leq \left(\frac{\lambda_2}{\lambda_1}\right)^{2t} (\operatorname{tg}\vartheta_0)^2 \cdot \frac{1}{1 + \left(\frac{\lambda_2}{\lambda_1}\right)^{2t} (\operatorname{tg}\vartheta_0)^2}$$

$$\leq \left(\frac{\lambda_2}{\lambda_1}\right)^{2t} (\operatorname{tg}\vartheta_0)^2$$



end of proof.

If ϑ_0 is away from $\frac{\pi}{2}$ this tends to zero exp fast as $\left(\frac{\lambda_2}{\lambda_1}\right)^{2t}$. Note that

$$\| x^t - v_1 \|^2 = 1 + 1 - 2\, x^t \cdot v_1 = 2(1 - \cos\vartheta_t)$$

For $x^t M x^t$ we show that it tends to $v_1 M v_1 = \lambda_1$.
(details left as exercise).

□

Power Method for other eigenvalues / eigenvectors of the Matrix:

- "Deflate" the matrix $M' \leftarrow M - \lambda_1 \underline{v}_1 \underline{v}_1^T$.

  Since $M' = \sum_{i=2}^{N} \lambda_i \underline{v}_i \underline{v}_i^T$ we can apply the same power method to M' to get $\lambda_2, \underline{v}_2$ as long as $\lambda_2 > \lambda_3$.

- and so on for other λ_i, v_i's as long as we have always a gap between eigenvalues.

Remark:

Stability of the algorithm with respect to noise will require good gaps between eigenvalues [see for example Golub & van Loan 1996, Sec 8.2-3].

III Tensor Power Method.

III.a) Method for "orthogonal tensors".

First we look at tensors that have a decomposition of the form

$$T = \sum_{i=1}^{k} \lambda_i \, \underline{v}_i \otimes \underline{v}_i \otimes \underline{v}_i$$

where $\lambda_i \in \mathbb{R}$ and $[\underline{v}_1 \dots \underline{v}_k]$ is an array of orthonormal vectors in $\mathbb{R}^D$. Note that necessarily we must have $k \leq D$ and that Jennrich's thm applies. So this polyadic decomposition is unique.

We will generalize from matrix case. The main idea is as follows:

input $T^{\times 3}$; output λ_i & $\underline{v}_i$.

1) Take $\underline{x}^{(0)}$ unit vector in $\mathbb{R}^D$ at random.

2) Iterate $\underline{x}^t = \dfrac{T(I, \underline{x}^t, \underline{x}^t)}{\| T(I, \underline{x}^t, \underline{x}^t) \|_2}$ where by definition $T(I, \underline{x}^t, \underline{x}^t) = \sum_{i=1}^{k} \lambda_i \underline{v}_i (\underline{v}_i^T \cdot \underline{x}^t)^2$

3) Under suitable conditions: $\underline{x}^t \to$ some $\underline{v}_*$ and $\lambda^t \to$ some λ_* (see later).

4) Deflate the Tensor

$$T' \leftarrow T - \lambda_* \underline{v}_* \otimes \underline{v}_* \otimes \underline{v}_* .$$

5) and so on.

Remarks:

a) Convergence to $\underline{v}_*$, λ_* depends on the initialization among other things and one does not converge to "top" "eigenvalue" "eigenvector".

b) Steps 3 and 4 work out if a certain non-degeneracy condition is met (see Thm later).

c) This method can be applied to tensors s.t the $\underline{v}_i$'s are NOT orthogonal after a suitable orthonormalization procedure called "whitening" (see last paragraph).

Theorem on Tensor Power Method.

Let $T = \sum_{i=1}^{K} \lambda_i \, \underline{v}_i \otimes \underline{v}_i \otimes \underline{v}_i$ with $[\underline{v}_1 \dots \underline{v}_K]$ $D \times K$ an orthonormal set of vectors and $\lambda_1 \dots \lambda_K \in \mathbb{R}$.

Suppose we initialize the method with $\underline{x}^{(0)} \in \mathbb{R}^D$ s.t the set of numbers

$$|\lambda_1 \, \underline{v}_1^T \cdot \underline{x}^{(0)}|, \dots, |\lambda_K \, \underline{v}_K^T \cdot x^{(0)}|$$

has a unique largest element. Without loss of generality suppose

$$|\lambda_1 \, \underline{v}_1^T \cdot \underline{x}^{(0)}| > |\lambda_2 \, \underline{v}_2^T \cdot \underline{x}^{(0)}| \geq \dots \geq \dots$$

↑ strict for this one

Set $$\underline{x}^t = \frac{T(I, \underline{x}^{t-1}, \underline{x}^{t-1})}{\| T(I, \underline{x}^{t-1}, \underline{x}^{t-1} \|} \qquad t \geq 1$$

Then $$\| \underline{x}^t - \underline{v}_1 \|^2 \leq \left(2 \lambda_1^2 \sum_{i=2}^{K} \lambda_i^{-2} \right) \left| \frac{\lambda_2 \, \underline{v}_2^T \cdot \underline{x}^0}{\lambda_1 \, \underline{v}_1^T \cdot \underline{x}^0} \right|^{2^{t+1}}.$$

and similarly for

$$\lambda^t = T(x^t, x^t, x^t) \to \lambda_1 \quad \text{fast.}$$

Proof.

First note that by definition:

$$T(I, x^t, x^t) = \sum_{i=1}^{k} \lambda_i \, \underline{v}_i \, (\underline{v}_i^T \cdot \underline{x}^t)^2$$

By looking at iteration t & $t-1$ we check (as for matrices) first that

$$\underline{x}^t = \frac{\sum_{i=1}^{k} \lambda_i^3 \, (v_i^T x^{t-1})^2 \, v_i}{\left\| \sum_{i=1}^{k} \lambda_i \, (v_i^T \cdot x^{t-1})^2 \, v_i \right\|_2}$$

$$= \frac{\sum_{i=1}^{k} \lambda_i^3 \, (v_i^T \cdot x^{t-2})^4 \, v_i}{\left\| \sum_{i=1}^{k} \lambda_i^3 \, (v_i^T \cdot x^{t-2})^4 \, v_i \right\|_2}$$

$$= \frac{\sum_{i=1}^{k} \lambda_i^{2^t - 1} \, (v_i^T \cdot x^{(0)})^{2^t} \, v_i}{\left\| \sum_{i=1}^{k} \lambda_i^{2^t - 1} \, (v_i^T \cdot x^{(0)})^{2^t} \, v_i \right\|_2}$$

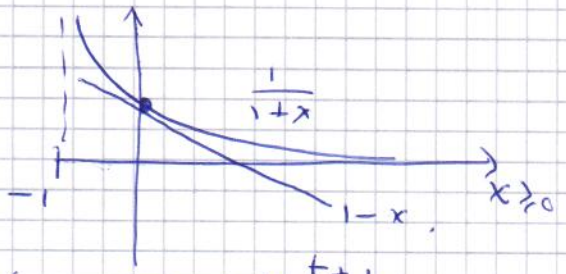
Multiplying by $\underline{v}_1^T$ we find

$$(\underline{v}_1^T \cdot \underline{x}^t)^2 = \frac{\lambda_2^{2^{t+1} - 2} \, (\underline{v}_1^T \cdot \underline{x}^{(0)})^{2^{t+1}}}{\sum_{i=1}^{k} \lambda_i^{2^{t+1} - 2} \, (\underline{v}_i^T \cdot \underline{x}^{(0)})^{2^{t+1}}} .$$

$$\Rightarrow \quad (\underline{v}_1^T \cdot \underline{x}^t)^2 = \frac{1}{1 + \sum_{i=2}^{K} \left(\frac{\lambda_i \, \underline{v}_i^T \underline{x}^0}{\lambda_1 \, \underline{v}_1^T \cdot x^0} \right)^2 \left(\frac{\lambda_1}{\lambda_i} \right)^{2t+1}}$$

$$\geq 1 - \lambda_1^2 \sum_{i=2}^{K} \lambda_i^{-2} \left(\frac{\lambda_i \, \underline{v}_i^T \cdot \underline{x}^0}{\lambda_1 \, \underline{v}_1^T \cdot \underline{x}^0} \right)^{2t+1}$$

where we used $x \geq 0 \Rightarrow \frac{1}{1+x} \geq 1 - x$.



$$\Rightarrow \quad 1 - (\underline{v}_1^T \cdot \underline{x}^t)^2 \leq \lambda_1^2 \left(\sum_{i=2}^{K} \lambda_i^{-2} \right) \cdot \left(\frac{\lambda_2 \, \underline{v}_2^T \cdot x^0}{\lambda_1 \, \underline{v}_1^T \cdot x^0} \right)^{2t+1}$$

$$1 - (\cos\vartheta_t)^2 \geq 1 - \cos\vartheta_t \quad (\text{for } \cos\vartheta_t < 1).$$

$$\Rightarrow \quad 2(1 - \cos\vartheta_t) \leq 2 \cdot \lambda_1^2 \left(\sum_{i=2}^{K} \lambda_i^{-2} \right) \left(\frac{\lambda_2 \, \underline{v}_2^T \cdot \underline{x}^0}{\lambda_1 \, \underline{v}_1^T \cdot \underline{x}^0} \right)^{2t+1}$$

$$\| \underline{x}^t - \underline{v}_1 \|_2^2 .$$

For $\lambda_t = T(x^t, x^t, x^t)$ we have

$$\lambda_t = \sum_{i=1}^{K} \lambda_i (\underline{v}_i \cdot \underline{x}^t)^3$$

$$= \lambda_1 (\underline{v}_1^T \cdot x^t)^3 \left\{ 1 + \sum_{i=2}^{K} \frac{\lambda_i (\underline{v}_i^T \cdot x^t)^3}{\lambda_1 (\underline{v}_i^T \cdot x^t)^3} \right\}$$

$$\to 0 \quad \text{for } \underline{v}_i^T \cdot \underline{v}_1 = 0, \; i \geq 2$$

$$\to \lambda_1 .$$

III. b) Whitening procedure.

Often in applications we have

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

where μ_i's are not necessarily orthogonal.
Before applying the power method we would like to change basis in order to transform T to an "orthogonal tensor".

This is possible when

$$M = \sum_{i=1}^{k} \lambda_i \, \mu_i \otimes \mu_i = \sum_{i=1}^{k} \lambda_i \, \mu_i \mu_i^T$$

is available (e.g. Topic Models, GMM).

Assumption: $[\mu_1, \dots, \mu_k]$ are linearly independent and $\lambda_1 \dots \lambda_k$ are strictly positive.
Then M is a symmetric semi-definite positive (square) matrix and we have

$$M_2 = U D U^T$$

with $D = \mathrm{diag}(d_i)$ $d_i > 0$. Since M_2 is rank k we have $d_1 \geq d_2 \geq \dots \geq d_k > 0$ (with other possible eigenvalues zero)

$$\Rightarrow \quad M_2 = U_{N\times K}\ \mathrm{Diag}(d_1 \dots d_K)\ U^T_{K\times N}$$

where $U_{N\times K} = [\underline{u}_1 \dots \underline{u}_K]$ are orthonormal eigenvectors.

Define: $W = U_{N\times K}\ \mathrm{Diag}(d_1^{-1/2} \dots d_K^{-1/2})$.

We claim that the vectors

$$\underline{v}_i = \sqrt{d_i}\, W^T \underline{u}_i \qquad i = 1 \dots k \qquad (*),$$

form an orthonormal system of vectors. (proof below).

Whitening of T is the operation:

$$T(W, W, W) \equiv \sum_{i=1}^{K} \lambda_i\, W^T \underline{u}_i \otimes W^T \underline{u}_i \otimes W^T \underline{u}_i$$

(a multi-linear transformation to a new Tensor)

$$= \sum_{i=1}^{K} \lambda_i^{-1/2}\ \underline{v}_i \otimes \underline{v}_i \otimes \underline{v}_i .$$

After applying e.g. the Tensor Power method to the withened tensor we retrieve $\lambda_i^{-1/2}$ and $\underline{v}_i$. Then we get $\underline{u}_i$ from $\underline{v}_i$ as follows:

$$\underline{u}_i = \frac{1}{\sqrt{d_i}}\ U_{N\times K}\ \mathrm{diag}(d_1^{1/2} \dots d_K^{1/2})\ \underline{v}_i$$

as can be checked from (*) above.

Proof of Claim

Recall $M_2 = U_{N\times K}\ \mathrm{Diag}(d_1 \cdots d_K)\ U^T_{K\times N}$

and by definition $W = U_{N\times K}\ \mathrm{Diag}(d_1^{-1/2} \cdots d_K^{-1/2})$

Thus:

$$W^T M_2 W = \mathrm{Diag}(d_1^{-1/2} \cdots d_K^{-1/2})\ U^T_{K\times N}\ U_{N\times K}$$
$$\cdot\ \mathrm{Diag}(d_1 \cdots d_K)\ U^T_{K\times N}\ U_{N\times K}$$
$$\cdot\ \mathrm{Diag}(d_1^{-1/2} \cdots d_K^{-1/2})$$

$$\Rightarrow \quad W^T M_2 W = I_{K\times K}.$$

which means:

$$\sum_{i=1}^{K} \lambda_i (W^T \underline{\mu}_i)(\underline{\mu}_i^T W) = I$$

$$\Rightarrow \quad \sum_{i=1}^{K} (\sqrt{\lambda_i}\, W^T \underline{\mu}_i)(\sqrt{\lambda_i}\, \underline{\mu}_i^T W) = I$$

$$\Rightarrow \quad \sum_{i=1}^{K} (\sqrt{\lambda_i}\, W^T \underline{\mu}_i)(\sqrt{\lambda_i}\, W^T \underline{\mu}_i)^T = I_{K\times K}.$$

Since $[\underline{\mu}_1 \cdots \underline{\mu}_K]$ are lin indep and W^T is full row-rank $[W^T\underline{\mu}_1 \cdots W^T\underline{\mu}_K]$ are also lin indep. Thus this equation implies they must be orthogonal.

For this last point in more details:

Set $W^T \underline{\mu}_i = \underline{w}_i$. Since $\underline{\mu}_1 \dots \underline{\mu}_K$ form a basis of K-dim space and $\text{Rank}(W^T) = K$ the $\underline{w}_i$ also form a basis of K-dim space.

Moreover $\sum_{i=1}^{K} \underline{w}_i \underline{w}_i^T = I_{K \times K}$ implies $\sum_{i=1}^{K} \underline{w}_i (\underline{w}_i^T \cdot \underline{w}_\ell) = \underline{w}_\ell$. Thus we must have $\underline{w}_i^T \cdot \underline{w}_\ell = \delta_{i\ell}$, i.e they are orthonormal.