TENSOR FACTORIZATION METHODS : Lecture 2.

I. Tools : Matricizations, Kronecker product, Khatri-Rao product.

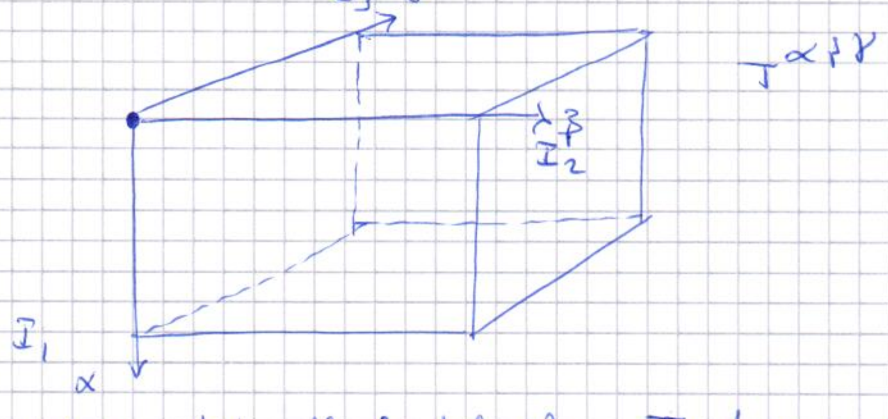
II. Alternating Least Square (ALS) algorithm.

III. Tucker Decomposition and Higher Order Singular value Decomposition (HOSVD).

I. Tools.

We will need various extra tools to go on and introduce them in this paragraph.

I.a) Matricizations of tensors.



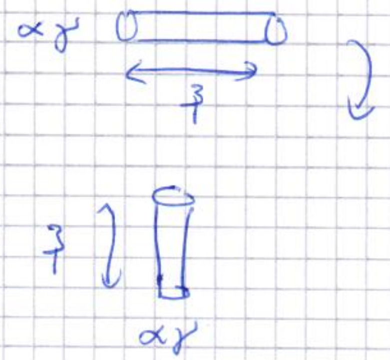
We can take all frontal slices And make out a matrix :

By definition $T_{(1)}$ = matrix of fibers $T^{\bullet(\beta\gamma)}$ ordered as indicated above

$= I_1 \times I_2 I_3$ matrix.

We can also take lateral slices and construct two other matrices:

$T_{(2)}$ = matrix of fibers $T^{\alpha\bullet\gamma}$ put as columns :



This is an $I_2 \times I_1 I_3$ matrix.

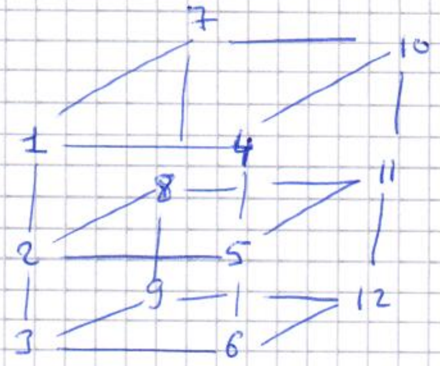
$T_{(3)}$ = matrix of fibers $T^{\alpha\beta\bullet}$ put as columns :

This is an $I_3 \times I_1 I_2$ matrix.

We say that $T_{(1)}$, $T_{(2)}$, $T_{(3)}$ are the three matricizations of the tensor T. See example.

Example.

$$T^{\alpha\beta\gamma} =$$



$$T_{(1)} = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}$$

$$T_{(2)} = \begin{bmatrix} 1 & 2 & 3 & 7 & 8 & 9 \\ 4 & 5 & 6 & 10 & 11 & 12 \end{bmatrix}$$

$$T_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 8 & 9 & 10 & 11 & 12 \end{bmatrix}.$$

∎

We will need to express matricizations of

$$\sum_{r=1}^{R} \underline{a}_r \otimes \underline{b}_r \otimes \underline{c}_r$$

For this handy tools will be the Khatri-Rao product, and the Kronecker product.

I, b) Kronecker Product of vectors.

$$\underline{c} \otimes_{Kro} \underline{b} \equiv \underbrace{\begin{bmatrix} c^1 \\ \vdots \\ c^{I_3} \end{bmatrix}}_{I_3 \text{ vect}} \otimes_{Kro} \underbrace{\begin{bmatrix} b^1 \\ \vdots \\ b^{I_2} \end{bmatrix}}_{I_2 \text{ vect.}} = \underbrace{\begin{bmatrix} c^1 \underline{b} \\ \vdots \\ c^{I_3} \underline{b} \end{bmatrix}}_{I_3 I_2 \text{ vector.}}$$

$$(\underline{c} \otimes_{Kro} \underline{b})^T \underbrace{\equiv}_{\text{by definition}} \underline{c}^T \otimes_{Kro} \underline{b}^T \underbrace{\equiv}_{\text{by def}} \underbrace{[\, c^1 \underline{b}^T, \dots c^{I_3} \underline{b}^T \,]}_{I_3 I_2 \text{ line vector.}}$$

Exercise! Show that

$$(\underline{e} \otimes_{Kro} \underline{d})^T \underbrace{\cdot}_{\text{inner product}} (\underline{c} \otimes_{Kro} \underline{b}) = (\underline{e}^T \cdot \underline{c})(\underline{d}^T \cdot \underline{b})$$

usual inner products

Notation and Remark.

1) $\otimes$ and $\otimes_{Kro}$ are basically equivalent except that the resulting object is presented as a matrix with $\underline{a} \otimes \underline{b}$ and as a column vect with $\underline{a} \otimes_{Kro} \underline{b}$. But these are isomorphic.

2) Sometimes $\otimes_{Kro}$ is called $\otimes$ and $\otimes$ is called $\circledcirc$.

II.c) Khatri-Rao product.

$C = [\underline{c}_1, \dots \underline{c}_R]$ $\quad I_3 \times R$

$B = [\underline{b}_1, \dots \underline{b}_R]$ $\quad I_2 \times R$.

$C \otimes_{KhR} B = [\underline{c}_1 \otimes_{Kro} \underline{b}_1, \underline{c}_2 \otimes_{Kro} \underline{b}_2, \dots, \underline{c}_R \otimes_{Kro} \underline{b}_R]$.

$I_2 I_3 \times R$ matrix.

Often one also use the relation $\otimes_{KhR} = \odot$.

Properties (exercise!)

1) $$(E \otimes_{KhR} D)^T (C \otimes_{KhR} B) = (E^T C) * (D^T B)$$

where $*$ is the Hadamard pointwise product between matrices: $(A * B)_{ij} = A_{ij} B_{ij}$. For example

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} * \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} b_{11} & a_{12} b_{12} \\ a_{21} b_{21} & a_{22} b_{22} \end{pmatrix}.$$

2) If C & B are full column rank (i.e their columns are lin-independent) Then $C \otimes_{KhR} B$ is also full column rank i.e $\underline{c}_1 \otimes_{Kro} \underline{b}_1, \dots, \underline{c}_R \otimes_{Kro} \underline{b}_R$ are lin-independent.

I.d) Matricizations of $T = \sum_{r=1}^{R} \underline{a}_r \otimes \underline{b}_r \otimes \underline{c}_r$;

We have
$$\begin{cases} T_{(1)} = A \, (C \otimes_{KhR} B)^T . \\ T_{(2)} = B \, (A \otimes_{KhR} C)^T \\ T_{(3)} = C \, (B \otimes_{KhR} A)^T . \end{cases}$$

Check of the first equation ;

$$T_{(1)} = \left[\sum_{r=1}^{R} a_r^\alpha b_r^\beta c_r^1 \;\Big|\; \cdots \;\Big|\; \sum_{r=1}^{R} a_r^\alpha b_r^\beta c_r^\gamma \right].$$

$$= \left[a_1^\alpha \cdots a_R^\alpha \right] \left[\begin{matrix} b_1^\beta c_1^1 \\ \vdots \\ b_R^\beta c_R^1 \end{matrix} \;\Bigg|\; \cdots \;\Bigg|\; \begin{matrix} b_1^\beta c_1^{I_3} \\ \vdots \\ b_R^\beta c_R^{I_3} \end{matrix} \right]$$

$$= A \left[\begin{matrix} \underline{b}_1^T c_1^1 \\ \\ \underline{b}_R^T c_R^1 \end{matrix} \;\Bigg|\; \cdots \;\Bigg|\; \begin{matrix} \underline{b}_1^T c_1^{I_3} \\ \\ \underline{b}_R^T c_R^{I_3} \end{matrix} \right]$$

$$= A \begin{bmatrix} \underline{c}_1^T \otimes \underline{b}_1^T \\ \vdots \\ \underline{c}_R^T \otimes \underline{b}_R^T \end{bmatrix} = A \, (C \otimes_{KhR} B)^T .$$

II. ALS algorithm for Tensor Decomposition.

We first introduce the definition of the Frobenius norm for tensors. This is nothing else than the Euclidean norm for the array $T^{\alpha\beta\gamma}$ of numbers.

Def: Frobenius Norm

$$\|T\|_F^2 = \sum_{\alpha=1}^{I_1} \sum_{\beta=1}^{I_2} \sum_{\gamma=1}^{I_3} |T^{\alpha\beta\gamma}|^2 .$$

Property: $\|\cdot\|_F$ is rotation invariant in the sence that $T(R_1, R_2, R_3)$ has the same norm for rotation (orthogonal) matrices R_1, R_2, R_3. (see exercises).

Now suppose we are given an array of numbers $T^{\alpha\beta\gamma}$ and suppose we know there exist

$A = [\underline{a}_1 \dots \underline{a}_R]$, $B = [\underline{b}_1 \dots \underline{b}_R]$, $C = [\underline{c}_1 \dots \underline{c}_R]$

s.t. $T = \sum_{r=1}^{R} \underline{a}_r \otimes \underline{b}_r \otimes \underline{c}_r$. In order to find A, B, C we could try to minimize

$$\| T - \sum_{r=1}^{R} \underline{a}_r \otimes \underline{b}_r \otimes \underline{c}_r \|_F^2$$

over unknowns A, B, C.

This is a highly non convex minimization problem and in general we do not have algos with convergence garantees. Convergence will in particular be highly dependent ~~upon~~ upon initialization.

Note that obviously:

$$\| T - \sum_{r=1}^{R} \underline{a}_r \otimes \underline{b}_r \otimes \underline{c}_r \|_F^2$$

$$= \| T_{(1)} - A (C \otimes_{KhR} B)^T \|_F^2$$

$$= \| T_{(2)} - B (A \otimes_{KhR} C)^T \|_F^2$$

$$= \| T_{(3)} - C (B \otimes_{KhR} A)^T \|_F^2$$

(where on the right-hand side we have Matrix Frob Norms).

<u>Main idea of ALS:</u> Go through the steps:

$$A \leftarrow \arg\min_A \| T_{(1)} - A (C \otimes_{KhR} B)^T \|_F^2$$

$$B \leftarrow \arg\min_B \| T_{(2)} - B (A \otimes_{KhR} C)^T \|_F^2$$

$$C \leftarrow \arg\min_C \| T_{(3)} - C (B \otimes_{KhR} A)^T \|_F^2 .$$

Solving the least square problems (see later on) we can cast the ALS algorithm in the form:

- Input ~~A~~ $T^{\alpha\beta\gamma}$; Output A, B, C
- Initialize with $B^{(0)}$, $C^{(0)}$ of full column rank and do:

$$\begin{cases} A^{m+1} = T_{(1)} \left(C^{(m)} \otimes_{KhR} B^{(m)} \right) \left(C^{mT} C^{m} * B^{mT} B \right)^{-1} \\ B^{m+1} = T_{(2)} \left(A^{m+1} \otimes_{KhR} B^{m} \right) \left(A^{m+1T} A^{m+1} * C^{mT} C^{m} \right)^{-1} \\ C^{m+1} = T_{(3)} \left(B^{m+1} \otimes A^{m+1} \right) \left(B^{m+1T} B^{m+1} * A^{m+1} A^{m+1} \right)^{-1} \end{cases}$$

(Comment on ALS1 ; ALS2 .)

Derivation of These Equations.

To derive these equations it suffices to see that

$$\underset{A}{\arg\min} \| T_{(1)} - A (C \otimes_{KhR} B)^T \|_F^2$$

$$= T_{(1)} \left((C \otimes_{KhR} B)^T \right)^+$$ ← Moore Penrose pseudo inverse

and for C & B full column rank $\Rightarrow$ $C \otimes_{KhR} B$ full column Rank $\Rightarrow$ $(C \otimes_{KhR} B)^T$ full Row rank $\Rightarrow$

$$\left((C \otimes_{KhR} B)^T \right)^+ = (C \otimes_{KhR} B) \left((C \otimes_{KhR} B)^T (C \otimes_{KhR} B) \right)^{-1}$$
$$= (C \otimes_{KhR} B) \left(C^T C * B^T B \right)^{-1}.$$

Recap of least square solution.

- The problem is of the form

$$\underset{X}{\arg\min} \, \| \underset{I_1 \times I_2 I_3}{\Upsilon} - \underset{I_1 \times R}{X} \, \underset{R \times I_2 I_3}{\Phi} \|_F^2$$

and Φ is full Row-Rank.

- Set $X = \Upsilon \underbrace{\phi^T(\phi\phi^T)^{-1}}_{\text{Moore Penrose pseudo inverse for } \phi \text{ full Row Rank.}} + Z$

$$\|\Upsilon - X\phi\|_F^2 = \|\Upsilon - \Upsilon\phi^T(\phi\phi^T)^{-1}\phi - Z\phi\|_F^2$$

$$= \|\Upsilon - \Upsilon\phi^T(\phi\phi^T)^{-1}\phi\|_F^2 + \|Z\phi\|_F^2 \qquad (*)$$

$$\geqslant \|\Upsilon - \Upsilon\phi^T(\phi\phi^T)^{-1}\phi\|_F^2$$

where we used in $(*)$ that the term:

$$2\,\mathrm{Tr}\,[\Upsilon - \Upsilon\,\phi^T(\phi\phi^T)^{-1}\phi]\,[Z\phi]^T$$

$$= 2\,\mathrm{Tr}\,\Upsilon\,[I - \phi^T(\phi\phi^T)^{-1}\phi]\,\phi^T Z^T$$

$$= 2\,\mathrm{Tr}\,\Upsilon\,[\underbrace{\phi - \phi^T(\phi\phi^T)^{-1}\phi\phi^T}_{0}]\,Z^T = 0.$$

- Thus $\|\Upsilon - X\phi\|_F^2 \geqslant \|\Upsilon - \Upsilon\,\phi^T(\phi\phi^T)^{-1}\phi\|_F^2$

for all X and equality is attained with $X = \Upsilon\phi^T(\phi\phi^T)^{-1}$.

III. Tucker Decomposition(s) and HOSVD.

III. a) Recap of Matrix SVD.

* A ; $M \times N$ real matrix. We can alway find U, V orthogonal (i.e s.t $U^T U = U U^T = I$, $V^T V = V^T V = I$)

s.t $\underbrace{A}_{M\times N} = \underbrace{U}_{M\times M} \underbrace{\Sigma}_{M\times N} \underbrace{V^T}_{N\times N}$

where Σ_{ij} has $\Sigma_{ii} = \sigma_i$; $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min(M,N)}$ and zero elsewhere are the singular values.

* Suppose $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_R > 0$ for $R \leq \min(M,N)$ (so possibly some singular values are zero) we can write SVD as

$$\underbrace{A}_{M\times N} = \bar{U}_{M\times R} \, \Sigma_{R\times R} \, \bar{V}^T_{R\times N}$$

$$= \sum_{r=1}^{R} \sigma_r \, \underline{u}_r \, \underline{v}_r^T$$

left sing vector right singular vectors

$U_{M\times R} = [\, \underline{u}_1 \; \dots \; \underline{u}_R \,]$ $\underset{N\times R}{V} = [\, \underline{v}_1 \; \dots \; \underline{v}_R \,]$

are $M \times R$ and $N \times R$ arrays of orthonormal vectors.

$U^T_{M\times R} U_{M\times R} = I_{R\times R}$ and $V^T_{N\times R} V_{R\times N} = I_{R\times R}$.

* If the singular values are all distinct in this (last) form the SVD decomposition is unique.

* For matrices the Eckart – Young Theorem solves the following problem: Find the Best Rank Approximation to a Matrix.

Eckart - Young Theorem

Let $K \leq R$, Then

$$\underset{A_K : \, \mathrm{Rank}(A_K) \leq K}{\arg\min} \| A - A_K \|_F^2 = \sum_{i=1}^{K} \sigma_i \, \underline{u}_i \, \underline{v}_i^T$$

where we truncated the SVD to the first K highest singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_K$.

This Theorem is at basis of dimensionality reduction techniques (e.g PCA).

* For Tensors the problem of finding

$$\underset{T_K \text{ of rank } K}{\arg\min} \| T - T_K \|_F^2$$

is not well defined. Indeed (see exercises) the space of tensors of rank K is not closed: one can find sequences of rank K that converge to tensors of rank $K+1$. (This jump of rank cannot happen for matrices).

III. b) Concept of Multilinear Rank.

To formulate analogs of SVD we need another concept of rank ; the multilinear rank.

Consider a tensor T and its matricisations $T_{(1)}$, $T_{(2)}$, $T_{(3)}$ . . .

Let R_1 = dim space spanned by columns of $T_{(1)}$
= column rank of $T_{(1)}$.

R_2 = dim space spanned by columns of $T_{(2)}$
= column rank of $T_{(2)}$

R_3 = idem .

(and so on for higher order tensors).

Definition : Multilinear rank :

$$\mathrm{rank}_{\boxplus}(T) = (R_1, R_2, R_3, \ldots)$$

So the multilinear rank is a collection of numbers representing $\underbrace{\text{the ranks}}_{\text{matrix}}$ of the matricizations.

(Recall : Tensor Rank $= R =$ smallest number of terms in polyadic decomposition.
For Matrices the two concepts are equivalent $R_1 = R_2 = R$)

III. C) Orthogonal Tucker Decomposition.

Theorem: Let $\text{Rank}_{\boxplus}(T) = (R_1, R_2, R_3)$. It is always possible to decompose T as follows:

$$T = \sum_{p=1}^{R_1} \sum_{q=1}^{R_2} \sum_{r=1}^{R_3} G_{pqr} \; \underline{u}_p \otimes \underline{v}_p \otimes \underline{w}_r$$

where $[\underline{u}_1, \ldots, \underline{u}_{R_1}]$ $[\underline{v}_1, \ldots \underline{v}_{R_2}]$ $[\underline{w}_1 \ldots w_{R_3}]$ are orthogonal arrays of vectors and

G_{pqr} is an $R_1 \times R_2 \times R_3$ array of numbers called the core Tensor.

Remarks:

a) The core tensor is not diagonal. So this decomp is different from the polyadic one. Also here $\underline{u}$'s, $\underline{v}$'s, $\underline{w}$'s are $\perp$ but in the polyadic they are not necessarily $\perp$.

b) This decomposition is not unique. (unlike sometimes the SVD).

c) Truncations do NOT give a "best low-multilinear rank approx" but a pretty good one.

Remark b) Continued:

If we the non-unicity take three rotation orthogonal matrices M^u $R_1 \times R_2$, M^v $R_2 \times R_2$, M^w $R_3 \times R_3$.

Define
$$\tilde{\underline{u}}_p = \sum_{p'=1}^{R_1} (M^{uT})_{pp'} \underline{u}_{p'}$$

$$\tilde{\underline{v}}_p = \sum_{p'=1}^{R_2} (M^{vT})_{pp'} \underline{v}_{p'}$$

$$\tilde{\underline{w}}_p = \sum_{p'=1}^{R_3} (M^{wT})_{pp'} \underline{w}_{p'}$$

and
$$\tilde{G}_{pqr} = \sum_{p''q''r''} G_{p''q''r''} M^u_{pp''} M^v_{qq''} M^w_{rr''}.$$

Check that

$$\sum_{p,q,r} \tilde{G}_{pqr} \tilde{\underline{u}}_p \otimes \tilde{\underline{v}}_q \otimes \tilde{\underline{w}}_r = \sum_{pqr} G_{pqr} \underline{u}_p \otimes \underline{v}_q \otimes \underline{w}_r$$

From the fact that $M^u M^{uT} = M^{uT} M^u = I$

idem for M^v, M^w.

Remark c) continued: we will not prove here but just state two interesting facts:

- $\exists$ best multilinear Rank Approx. In other words given T of mult lin rank (R_1, R_2, R_3) of $K_1 \leq R_1$, $K_2 \leq R_2$, $K_3 \leq R_3$ the following problem is well defined (The min is attained):

$$\underset{\tilde{T}:\, \text{rank}\,(K_1, K_2, K_3)}{\arg\min} \| T - \tilde{T} \|_F^2 = T_*$$

- Although the above problem is well defined we do not in general know of good algorithms. But the Tucker - HOSVD (see next paragraph) gives a pretty good approximation in the following sense:

$$\| T - \underbrace{\tilde{T}_{K_1 K_2 K_3}}_{\substack{\text{Tensor} \\ \text{obtained by} \\ \text{truncating HOSVD} \\ \text{to } \sum\limits_{p \leq K_1} \sum\limits_{q \leq K_2} \sum\limits_{r \leq K_3}}} \|_F^2 \leq \sqrt{(\text{mode}) - 1}\; \| T - \underset{\substack{\uparrow \\ \text{Best}}}{\tilde{T}_*} \|_F^2$$

III . d) Proof of Theorem (Tucker) and HOSVD.

We provide a proof of the existence of a Tucker decomposition, which is constructive, and also gives us an algorithm. The particular decomposition obtained here is an "analog" of SVD and is called <u>Higher Order Singular Value Decomposition (HOSVD)</u>

Take the Matricizations $T_{(1)}, T_{(2)}, T_{(3)}$.
Perform usual matrix SVD on these matrices

$$T_{(1)} = U^{(1)}_{I_1 \times R_1} \Sigma_{R_1 \times R_1} V^{(1)T}_{R_1 \times I_2 I_3}$$

$$T_{(2)} = U^{(2)}_{I_2 \times R_2} \Sigma_{R_2 \times R_2} V^{(2)T}_{R_2 \times I_1 I_3}$$

$$T_{(3)} = U^{(3)}_{I_3 \times R_3} \Sigma_{R_3 \times R_3} V^{(3)T}_{R_3 \times I_1 I_2}$$

which exist (recall the dimensions and ranks of $T^{(1)}, T^{(2)}, T^{(3)}$ to get the above).

Consider left singular vectors of $T_{(1)}, T_{(2)}, T_{(3)}$.
These are read of the matrices $U^{(1)}_{I_1 \times R_1}$; $U^{(2)}_{I_2 \times R_2}$; $U^{(3)}_{I_3 \times R_3}$

$\longrightarrow$

$[\underline{u}_1^{(1)} \ldots \underline{u}_{R_1}^{(1)}]$ array for $U^{(1)}_{I_1 \times R_1}$

$[\underline{u}_1^{(2)} \ldots \underline{u}_{R_2}^{(2)}]$ array for $U^{(2)}_{I_2 \times R_2}$

$[\underline{u}_1^{(3)} \ldots \underline{u}_{R_3}^{(3)}]$ array for $U^{(3)}_{I_3 \times R_3}$.

Finally compute the tensor

$$G = T(U^{(1)}, U^{(2)}, U^{(3)})$$

where by definition this means:

$$G_{pqr} = \sum_{\alpha=1}^{I_1} \sum_{\beta=1}^{I_2} \sum_{\gamma=1}^{I_3} T^{\alpha\beta\gamma} U^{T(1)}_{\alpha p} U^{T(2)}_{\beta q} U^{T(3)}_{\gamma r} .$$

The existence of the Tucker decomposition is then established by checking that with this G and these $U^{(1)}, U^{(2)}, U^{(3)}$ we must have (invert the last relation using unitarity):

$$T^{\alpha\beta\gamma} = \sum_{p=1}^{R_1} \sum_{q=1}^{R_2} \sum_{r=1}^{R_3} G_{pqr} \underbrace{U^{(1)}_{\alpha p} U^{(2)}_{\beta q} U^{(3)}_{\gamma r}}_{u^{(1)\alpha}_{p}\, u^{(2)\beta}_{q}\, u^{(3)\gamma}_{r}}$$

i.e. $T = \sum_{pqr} G_{pqr} \underline{u}^{(1)} \otimes \underline{u}^{(2)} \otimes u^{(3)}$.

This last proof gives an algorithm for obtaining a Tucker decomposition

HOSVD algorithm.

input $T^{\alpha\beta\gamma}$; output G, U, V, W.

1) From $T^{\alpha\beta\gamma}$ consider matricizations $T_{(1)}, T_{(2)}, T_{(3)}$.
2) Compute left singular vectors of $T_{(1)}, T_{(2)}, T_{(3)}$ from the matrix SVD. This yields the arrays corresponding to non-zero singular values:

$$[\underline{u}_1 \ldots \underline{u}_{R_1}], \quad [\underline{v}_1 \ldots \underline{v}_{R_2}], \quad [\underline{w}_1 \ldots \underline{w}_{R_3}]$$

Note this also gives a systematic way to compute the multilinear rank (R_1, R_2, R_3).

3) Compute the "core tensor" from

$$G_{pqr} = \sum_{\alpha\beta\gamma} T^{\alpha\beta\gamma} u_p^\alpha v_q^\beta w_r^\gamma .$$

This yields the HOSVD:

$$T = \sum_{p,q,r}^{R_1,R_2,R_3} G_{pqr} \, \underline{u}_p \otimes \underline{v}_q \otimes \underline{w}_r .$$

Remark: as explained before this is just one particular Tucker decomp. These are not unique and can be obtained by rotations.