TENSOR FACTORISATION METHODS : Lecture 1.

I. Motivations

II. Introduction to the language of tensors.

III. Tensor Rank.

IV. Basic tensor factorization theorem.

I. Motivations.

I.a. Gaussian Mixture Model.

$\underline{x} \in \mathbb{R}^D \quad p(\underline{x}) = \sum_{i=1}^{K} w_i \exp\left\{ - \frac{\|\underline{x} - \underline{a}_i\|^2}{2\sigma^2} \right\} (2\pi\sigma^2)^{-D/2}$

isotropic gaussian mixture model. We want to learn hidden variables w_i, $\underline{a}_i$ (weights and means of clusters) from samples $\underline{x}^{(1)} \dots \underline{x}^{(N)}$.

A common method is to match empirical moments to theorical moments. For example let us look at the second cross-moment (exercise)

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

we may match it to the empirical moment

$$\frac{1}{N} \sum_{n=1}^{N} \underline{x}^{(n)} \underline{x}^{(n)T} \equiv M.$$

Suppose $K \leq N$. Since $\underline{a}_i \underline{a}_i^T$ is positive definite $\sum_{i=1}^{K} w_i \underline{a}_i \underline{a}_i^T$ only has ≥ 0 eigenvalues and there exist one zero eigenvalue (since $K < N$). Thus σ^2 is the smallest eigenvalue of $\mathbb{E}(\underline{x}\,\underline{x}^T)$ and we can determine an estimate from the empirical moment. Therefore to determine w_i and $\underline{a}_i$ we attempt to solve for a matrix factorization or decomposition problem

$$M - \sigma^2 I_{D\times D} \simeq \sum_{i=1}^{K} w_i \underline{a}_i \underline{a}_i^T .$$

<u>There is a general problem with this method because typically the solution is not unique.</u> Consider the simpler problem when $w_i = \frac{1}{K}$ (uniform weight). So we have an expression of the form

$$\sum_{i=1}^{K} \underline{b}_i \underline{b}_i^T = \underbrace{[\underline{b}_1 \cdots \underline{b}_K]}_{N\times K} \underbrace{\begin{bmatrix} \underline{b}_1^T \\ \vdots \\ \underline{b}_K^T \end{bmatrix}}_{K\times N}$$

$$= B B^T$$

$$= B R R^T B^T = (BR)(BR)^T$$

for R an $K \times K$ rotation matrix ($RR^T = R^T R = I$)

Therefore the solution (BR) is not unique.

This is called the "<u>Rotation Problem</u>"

Idea : Look at third moments !

In the exercises we prove that :

$$\mathbb{E}\left(x^\alpha x^\beta x^\gamma\right) = \sum_{i=1}^{k} w_i \, a_i^\alpha \, a_i^\beta \, a_i^\gamma + \dots$$

(where $\underline{x} = (x^\alpha)_{\alpha=1}^{D}$ $\quad \underline{a}_i = (a_i^\alpha)_{\alpha=1}^{D}$).

The idea is to find w_i and $\underline{a}_i$ from the empirical estimate

$$\frac{1}{N} \sum_{n=1}^{N} x^{\alpha(n)} \, x^{\beta(n)} \, x^{\gamma(n)}$$

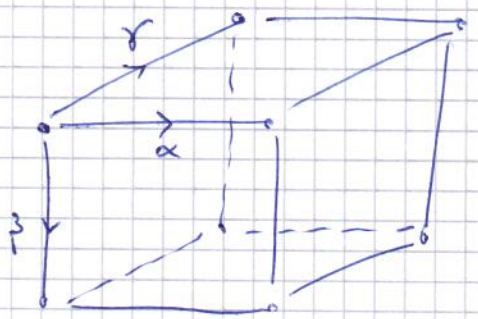
of the third moment. We will see that the rotation problem disappears.

The 3-d array of numbers

$$T^{\alpha\beta\gamma} = \mathbb{E}\left(x^\alpha x^\beta x^\gamma\right)$$

for $\alpha = 1 \dots D$, $\beta = 1 \dots D$, $\gamma = 1 \dots D$ is called a tensor. T

$T =$

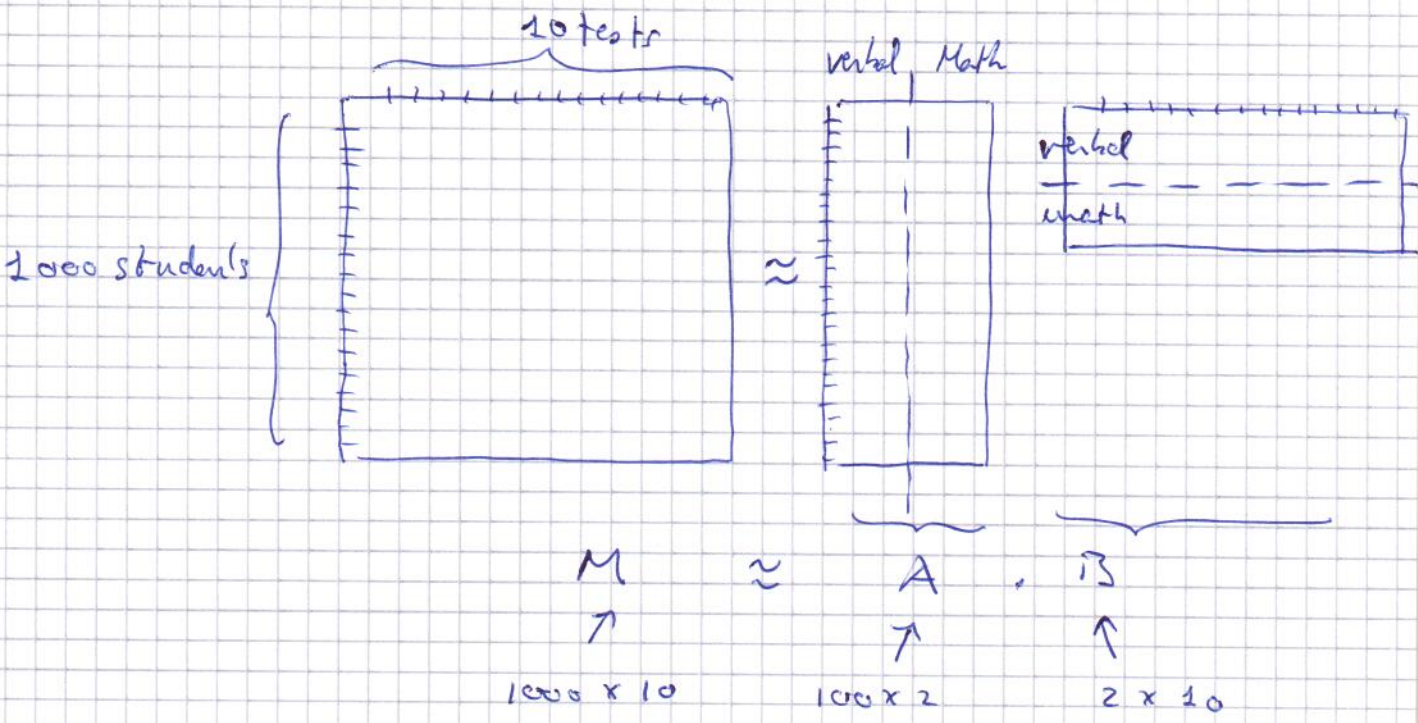


(Note that $\mathbb{E}(x^\alpha x^\beta)$ is a matrix $\alpha = 1 \dots D$, $\beta = 1 \dots D$.)

I.b. Spearman's Problem. (psychology).

Spearman's hypothesis to test "intelligence" : is that there are (say) two types of intelligence : "verbal" and "mathematical".

Data Matrix M = students x test matrix



The factorization of M extracts for each student a two dim vector that characterizes his verbal & math ability and similarly for each test the verbal & math components.

But again the interpretation of the results is problematic because of the "rotation problem"

$$A\,B = \underbrace{A\,R}_{A'}\,\underbrace{R^T\,B}_{B'} \quad \text{for } R = 2\times 2 \text{ orthogonal matrix.}$$

Adding a third dimension to the data matrix M (turning it to a higher dimensional array) and testing for more than two skills only solves the rotation problem and typically leads to a unique solution.

I.C. Recap on matrix factorizations.

With additional constraints of the factors the matrix factorizations / decompositions can become unique.

E.g.: Singular value decomposition:

Any $M = U \Sigma V^{T,*}$ where U and V are unitary and $\Sigma =$ ~~de~~ singular values. If all singular values are distinct (and non-zero) we can uniquely write

$$M = \sum_{i=1}^{R} \sigma_i \, \underline{u}_i \, \underline{v}_i^{T,*}$$

where $\sigma_1 > \sigma_2 > \cdots > \sigma_R > 0$ (R = rank of M).

(See exercises for more).

E.g.: M is a sym rank-one matrix $M = \underline{a}\,\underline{a}^T$ and $\underline{a}$ is unique up to sign.

II. Introduction to the language of tensors.

- **A column vector** $\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \in \mathbb{R}^D$ is a "one-index" or "one-mode" object with components x^α, $\alpha = 1 \cdots D$.

- **A matrix** $X = \begin{bmatrix} \uparrow I_1 \downarrow \\ \longleftrightarrow I_2 \longleftrightarrow \end{bmatrix} \in \mathbb{R}^{I_1} \times \mathbb{R}^{I_2}$

is a "two-index" or "two-mode" object or array with components $X^{\alpha\beta}$, $\alpha = 1 \cdots I_1$, $\beta = 1 \cdots I_2$.

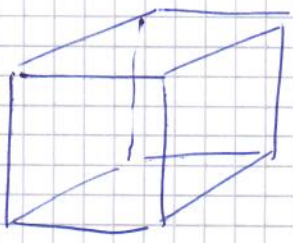
We denote by $X^{\bullet\beta}$ = vector corresponding to column β of the matrix X

$X^{\alpha\bullet}$ = line vector corresponding to ~~column~~ line α of X.

A matrix X can be viewed as a set of columns $\left(\downarrow \; \downarrow \; \downarrow \; \downarrow \right)$ or a set of rows $\left(\longleftrightarrow \right) \} X^{\alpha\bullet}$

$X^{\bullet\beta}$

- **A Tensor** is a multidimensional array of numbers

$T =$  $\in \mathbb{R}^{I_1} \times \mathbb{R}^{I_2} \times \mathbb{R}^{I_3}$

Has 3 indexes or 3 - "modes"

$T^{\alpha\beta\gamma}$ $\alpha = 1 \cdots I_1$, $\beta = 1 \cdots I_2$, $\gamma = 1 \cdots I_3$.

Generalization → p - dimensional array. T with components $T^{\alpha_1 \alpha_2 \dots \alpha_p} \in \mathbb{R}^{I_1} \times \mathbb{R}^{I_2} \times \dots \times \mathbb{R}^{I_p}$.

Terminology : p is the "order", the "mode", the "way".

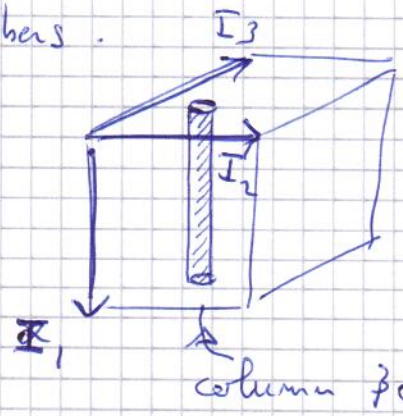
$p = 1 \rightarrow$ vector

$p = 2 \rightarrow$ matrix

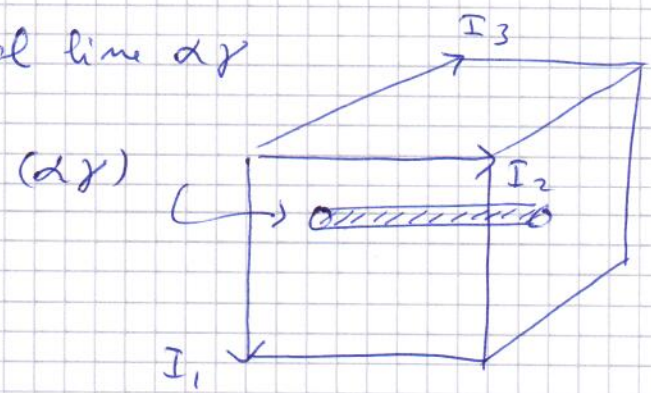
$p \geq 3 \rightarrow$ tensor in general.

• Views of a tensor : collection of fibers.

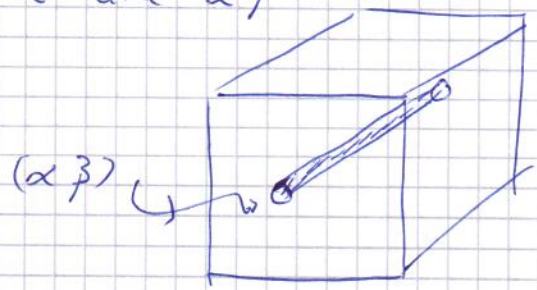
$X^{\bullet \beta\gamma}$ = fiber at column $\beta\gamma$



$X^{\alpha \bullet \gamma}$ = fiber at horizontal line $\alpha\gamma$

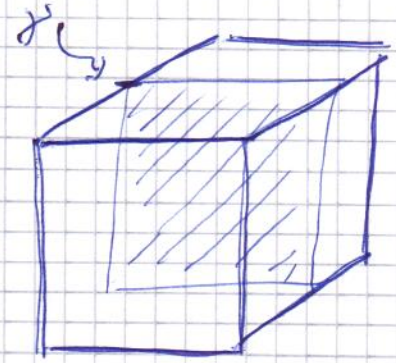


$X^{\alpha\beta \bullet}$ = fiber at horizontal line $\alpha\beta$

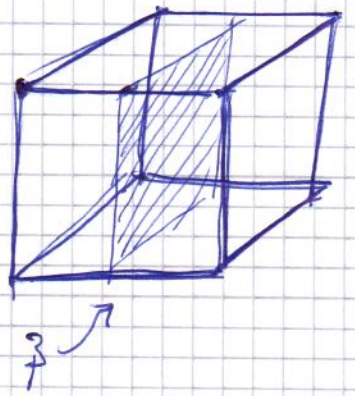


But we can also view the tensor as a collection of matrices

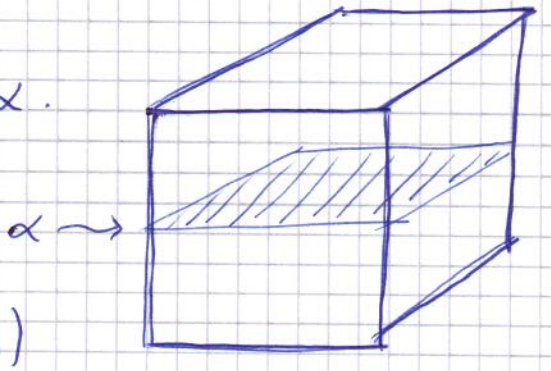
$X^{\cdot\cdot k}$ = Matrix of frontal slice k



$X^{\cdot j \cdot}$ = Matrix of lateral slice j



$X^{i \cdot\cdot}$ = Matrix of ~~lateral~~ horizontal slice i.



(The tensor is a collection of matrices)

- Remark:

A general approach to the analysis of tensors will be to reduce its study to collections of vectors and/or matrices and apply known results of linear algebra. Of course these vectors and/or matrices are not completely independent and this introduces constraints that make the decompositions more unique in general.

• Tensor Product of vectors.

A basic operation to form tensors from vectors is called the tensor product or outer product. By definition for $\underline{a} \in \mathbb{R}^{I_1}$, $\underline{b} \in \mathbb{R}^{I_2}$ we form the tensor

$$\underline{a} \otimes \underline{b} \quad \text{with components:} \quad (\underline{a} \otimes \underline{b})^{\alpha\beta} = a^{\alpha} b^{\beta}$$

For $\underline{a} \in \mathbb{R}^{I_1}$, $\underline{b} \in \mathbb{R}^{I_2}$, $\underline{c} \in \mathbb{R}^{I_3}$:

$$\underline{a} \otimes \underline{b} \otimes \underline{c} \quad \text{with components:} \quad (\underline{a} \otimes \underline{b} \otimes \underline{c})^{\alpha\beta\gamma} = a^{\alpha} b^{\beta} c^{\gamma}$$

Remark that $\underline{a} \otimes \underline{b} = \underline{a}\,\underline{b}^T$ = column vect × line vect = matrix.

(Remark also that $\underline{a}^T \underline{b} = \sum_{i=1}^{I} a_i b_i$ = inner product.).

Notation: we denote the tensor product as $\otimes$ as usual in math. In some papers it is denoted by $\odot$.

We will introduce many other useful products later on when we need them [Kronecker product, Khatri-Rao product, Hadamard product ...]

Parenthesis :

- Tensors viewed as multilinear transformations.

A matrix can be viewed as a linear transformation acting on vectors. For example :

$$(M\underline{x})^{\beta} = \sum_{\alpha=1}^{D} M^{\beta\alpha} x^{\alpha} \; ; \; \underline{y}^{T} M \underline{x} = \sum_{\alpha,\beta} y^{\beta} M^{\beta\alpha} x^{\alpha}$$

Similarly a tensor is a "multilinear transformation but a richer one" :

$$T(I, I, \underline{x})^{\alpha\beta} = \sum_{\gamma} T^{\alpha\beta\gamma} x^{\gamma} \quad \text{a matrix}$$

$$T(I, \underline{x}, I)^{\alpha\gamma} = \sum_{\beta} T^{\alpha\beta\gamma} x^{\beta} \quad \text{a matrix}$$

$$T(\underline{x}, I, I)^{\beta\gamma} = \sum_{\alpha} T^{\alpha\beta\gamma} x^{\alpha} = \text{a matrix}$$

$$T(I, \underline{x}, \underline{y})^{\alpha} = \sum_{\beta,\gamma} T^{\alpha\beta\gamma} x^{\beta} y^{\gamma} = \text{a vector}$$

$$T(\underline{x}, I, \underline{y})^{\beta} = \sum_{\alpha\gamma} T^{\alpha\beta\gamma} x^{\alpha} y^{\gamma} = \text{a vector}$$

$$T(\underline{x}, \underline{y}, I) = \sum_{\alpha\beta} T^{\alpha\beta\gamma} x^{\alpha} y^{\beta} = \text{a vector}$$

$$T(\underline{x}, \underline{y}, \underline{z}) = \sum_{\alpha\beta\gamma} T^{\alpha\beta\gamma} x^{\alpha} y^{\beta} z^{\gamma} = \text{a scalar.}$$

$$T(M_1, M_2, M_3)^{\alpha'\beta'\gamma'} = \sum_{\alpha\beta\gamma} T^{\alpha\beta\gamma} M_1^{\alpha\alpha'} M_2^{\beta\beta'} M_3^{\gamma\gamma'}$$

a new tensor ; etc ...

III. Tensor Rank.

So far we formed elementary tensors $\underline{a} \otimes \underline{b} \otimes \underline{c}$ from vectors $\underline{a}$, $\underline{b}$, $\underline{c}$. These are called rank-one tensors. This is by analogy with matrices for which rank-one matrices are of the form $\underline{a}\,\underline{b}^T = \underline{a} \otimes \underline{b}$

Given sufficiently many terms, any tensor can be decomposed as a sum of rank-one Tensors:

$$T = \sum_i \underline{a}_i \otimes \underline{b}_i \otimes \underline{c}_i$$

Indeed Given the number $T^{\alpha\beta\gamma} \in \mathbb{R}^{I_1} \times \mathbb{R}^{I_2} \times \mathbb{R}^{I_3}$ with sufficiently many vectors $\{\underline{a}_i, \underline{b}_i, \underline{c}_i\}$ we can always solve for these vectors. Such decompositions are often called "polyadic decompositions"

Definition: Tensor Rank. The minimal possible number of terms in a polyadic decomposition is called the rank of the tensor. We will denote it by R. So R = min possible # of terms such that

$$T = \sum_{i=1}^{R} \underline{a}_i \otimes \underline{b}_i \otimes \underline{c}_i$$

for some $A = [\underline{a}_1 \dots \underline{a}_R]$ $B = [\underline{b}_1 \dots \underline{b}_R]$

$C = [\underline{c}_1 \dots \underline{c}_R]$.

Remarks:

a) For matrices this definition reduces to usual definition as dim of span of columns = column rank

dim of span of rows = row rank

recall that always row rank = column rank.

b) For matrices all definitions of rank are equivalent and there is only one "good" ~~def~~ notion of rank. But for tensors, as we will see, there is not a unique useful such definition. Tensor Rank defined above is our first such definition - [other ones we will encounter are Kruskal-rank, Multilinear-Rank].

c) For matrices it is possible to systematically determine the rank, e.g. by row and/or column operations or by SVD. For Tensors there is no such systematic method. Determining the rank of a tensor is difficult.

d) Tensor Rank has some counter-intuitive properties (see examples & exercises).

e) For order Three $R \leq \min(I_1 I_2, I_1 I_3, I_2 I_3)$ but nothing much more can be said.

IV. Basic Tensor factorization / decomposition Theorem.

Theorem [Jennrich, Carroll, Harshman 1970's ...].

Let $A = [\underline{a}_1 \dots \underline{a}_R]$ $I_1 \times R$ matrix of column vectors

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

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

Let $R \leq \min(I_1, I_2)$ and A, B full column Rank (in other words $\underline{a}_1 \dots \underline{a}_R$ are lin-indep and $\underline{b}_1 \dots \underline{b}_R$ are lin-indep).

Let also vectors $\underline{c}_1 \dots \underline{c}_R$ be pair-wise lin indep. (in other words $\forall i,j \quad \underline{c}_i \neq \lambda \underline{c}_j$).

Suppose $T = \sum_{r=1}^{R} \underline{a}_i \otimes \underline{b}_i \otimes \underline{c}_i$ is a tensor formed out of these arrays of vectors.

Then the decomposition of T as a sum of rank-one tensors is unique and its rank is R.

(up to scalings $\lambda \underline{a}_i \otimes \frac{1}{\lambda} \underline{b}_i \otimes \underline{c}$ and permutation of terms in the sum.)

Remarks:

a) This theorem will be proven in a constructive way and the proof gives us an algorithm to find A, B, C from the multi-array $T^{\alpha\beta\gamma}$ of numbers.

b) In practice we are given a noisy version of $T^{\alpha\beta\gamma}$ or some empirical estimate. It is an interesting problem to determine the stability of the algorithm for decomposition → (research problem currently). Basically stability requires good gaps between singular values of A and B.

c) For $R > \min(I_1, I_2)$ but not too large there exist unicity results but not algorithmic. For R very large unicity breaks down. There are also current research directions.

d) One result for higher rank is the following:

* Define the Kruskal rank of a set of vectors as the max K s.t. all K-subsets are lin-indep. and $\exists$ $K+1$-subset that is lin-dependent.

* Let K_a, K_B, K_C the Kruskal ranks of A, B, C. If $2R+2 \leq K_a + K_B + K_C$ then T has rank R and the decomposition is unique.

An important tool for the proof of Jennrich's theorem is the Moore-Penrose pseudo inverse (see exercises).

Let $A \in \mathbb{R}^{M \times N}$, $A^T \in \mathbb{R}^{N \times M}$, real matrix.

By definition the Moore-Penrose pseudo inverse is the matrix $A^+ \in \mathbb{R}^{N \times M}$ satisfying the four conditions:

1) $A A^+ A = A$
2) $A^+ A A^+ = A^+$
3) $(A A^+)^T = A A^+$ ($A A^+$ is symetric)
4) $(A^+ A)^T = A^+ A$ ($A^+ A$ is symmetric).

Properties of the Moore-Penrose pseudo inverse.

1) A^+ alway exists and is unique.
2) If A has real entries then A^+ also.
3) $(A^+)^+ = A$ (inverse of the inverse is A).
4) Pseudo inverse of $O_{M \times N}$ is $(O^T)_{N \times M}$ (zero matrix)
5) $(A^+)^T = (A^T)^+$
6) $(\alpha A)^+ = \alpha^{-1} A^+$ for $\alpha \neq 0$.
7) If A has full column rank and B has full row rank then $(AB)^+ = B^+ A^+$
8) If A has full column rank then $A^T A$ is invertible and $A^+ = (A^T A)^{-1} A^T$ and $A^+ A = I$.
9) If A has full row rank then $A A^T$ is invertible and $A^+ = A^T (A A^T)^{-1}$ and $A A^+ = I$.

Proof of Jennrich's algorithm.

We will project $T = \sum_{r=1}^{R} \underline{a}_r \otimes \underline{b}_r \otimes \underline{c}_r$ on two random vectors $\underline{x}, \underline{y} \in \mathbb{R}^{I_3}$ and consider the two matrices $T(I, I, \underline{x})$ and $T(I, I, \underline{y})$ here denoted as $T_{\underline{x}}$ and $T_{\underline{y}}$:

$$T_{\underline{x}} = \sum_{r=1}^{R} \underline{a}_r \otimes \underline{b}_r (\underline{x}^T \cdot \underline{c}_r) \quad \text{where } \underline{x}^T \cdot \underline{c}_r = \sum_{\gamma} x^{\gamma} c_r^{\gamma}$$

$$T_{\underline{y}} = \sum_{r=1}^{R} \underline{a}_r \otimes \underline{b}_r (\underline{y}^T \cdot \underline{c}_r) \text{ and } \underline{y}^T \cdot \underline{c}_r = \sum_{\gamma} y^{\gamma} c_r^{\gamma} .$$

We have :

$$T_{\underline{x}} = \sum_{r=1}^{R} \underline{a}_r \underline{b}_r^T (x^T \cdot c_r) = [\underline{a}_1 \dots \underline{a}_R] \begin{bmatrix} x^T \cdot c_1 & & \\ & \ddots & \\ & & x^T \cdot c_r \end{bmatrix} \begin{bmatrix} \underline{b}_1^T \\ \vdots \\ \underline{b}_R^T \end{bmatrix}$$

$$= \underbrace{A}_{I_1 \times R} \underbrace{\text{Diag}(x^T \cdot c_r)}_{R \times R} \underbrace{B^T}_{R \times I_2}$$

$$T_{\underline{y}} = A \, \text{Diag}(y^T \cdot c_r) \, B^T .$$

We look at the two matrices $T_{\underline{x}} (T_{\underline{y}})^+$ and $(T_{\underline{x}})^+ (T_{\underline{y}})$ where $(T_{\underline{x}})^+$ is the Moore-Penrose pseudo-inverse.

Let us compute $(T_y)^+$ and $(T_x)^+$.

A has full column rank and B^T full row rank. Also taking $y \in \mathbb{R}^J$ random with prob (one) [random according to continuous measure w.r.t Lebesgue measure] we have that $y^T \cdot c_r \neq 0 \quad \forall r = 1 \dots R$. Thus $\mathrm{Diag}(y^T c_r) B^T$ has also full row rank. By property 7) of MP we get:

$$(A \,\mathrm{Diag}(y^T \cdot c_r) B^T)^+ = (\mathrm{Diag}(y^T \cdot c_r) B^T)^+ A^+$$

$$= (B^T)^+ (\mathrm{Diag}(y^T \cdot c_r))^+ A^+$$

$$= (B^T)^+ \,\mathrm{Diag}\left(\frac{1}{(y^T \cdot c_r)}\right) A^+ .$$

$$= (T_y)^+ .$$

Similarly $(T_x)^+ = (B^T)^+ \,\mathrm{Diag}\left(\frac{1}{(x^T \cdot c_r)}\right) A^+ .$

Thus

$$T_x (T_y)^+ = A \,\mathrm{Diag}\left(\frac{(x^T \cdot c_r)}{(y^T \cdot c_r)}\right) A^+$$

$$(T_x)^+ T_y = (B^T)^+ \,\mathrm{Diag}\left(\frac{y^T \cdot c_r}{x^T \cdot c_r}\right) B^T .$$

Using $A^+ A = I$ for A with full column rank and $B^T (B^T)^+$ for B^T with full row rank we find:

$$T_x (T_y)^+ A = A \, \mathrm{Diag}\left(\frac{x^T \cdot c_r}{y^T \cdot c_r} \right)$$

$$B^T (T_x)^+ T_y = \mathrm{Diag}\left(\frac{y^T c_r}{x^T c_r} \right) B^T .$$

which mean also:

$$\begin{cases} T_x (T_y)^+ \underline{a}_r = \left(\dfrac{x^T \cdot c_r}{y^T \cdot c_r} \right) a_r \\ b_r^T (T_x)^+ T_y = \left(\dfrac{y^T \cdot c_r}{x^T \cdot c_r} \right) b_r \end{cases}$$

So $\underline{a}_r$ are right eigenvectors of $T_x (T_y)^+$ and
So b_r are left eigenvectors of $(T_x)^+ T_y$

To compute $\underline{a}_r$ & $\underline{b}_r$ in Jennrich's algorithm you just compute the right and left eigenvectors of $T_x (T_y)^+$ and $(T_x)^+ T_y$. The eigenvalues pair up as inverses of each other $\dfrac{x^T \cdot c_r}{y^T \cdot c_r}$ and $\dfrac{y^T \cdot c_r}{x^T \cdot c_r}$. The a_r's & $\underline{b}_r$'s are unique because all eigenvalues are distinct. This is where $\underline{c}_r \neq d \underline{c}_{r'}$ is important hypothesis.

Finely we have to determine $\underline{c}_r$, $r = 1 \dots R$.

Since we are give $T^{\alpha\beta\gamma}$ we have:

$$T^{\alpha\beta\gamma} = \sum_{r=1}^{R} a_r^{\alpha} b_r^{\beta} c_r^{\gamma}$$

Fix γ. This is a linear system of $I_1 I_2$ equations for R unknown c_r^{γ}, $r = 1 \dots R$. The matrix of this system is $\underbrace{M^{\alpha\beta, r}}_{I_1 I_2 \times R \text{ matrix}} = a_r^{\alpha} b_r^{\beta}$

This matrix can be shown to be full column rank R because A & B are full column rank - [see exercices].

Therefore the system of equations has a unique solution $[\, c_1^{\gamma} \dots c_R^{\gamma} \,]$.

Summary of Jennrich's Algorithm.

[Given $T^{\alpha\beta\gamma}$; Output A, B, C].

1) Compute $(T_{\underline{x}})^{\alpha\beta} = \sum_{\gamma} T^{\alpha\beta\gamma} x^{\gamma}$ for a random $\underline{x}$

   Compute $(T_{\underline{y}})^{\alpha\beta} = \sum_{\gamma} T^{\alpha\beta\gamma} y^{\gamma}$ for a random $\underline{y}$.

2) Compute Moore-Penrose pseudoinverse $(T_x)^+$ and $(T_y)^+$ and also $T_x (T_y)^+$ and $(T_x)^+ T_y$.

3) Compute right eigenvectors $\underline{a}_r$ of $T_x (T_y)^+$ and left eigenvectors $\underline{b}_r$ of $(T_x)^+ T_y$.

4) Pair them up (because they have inverse eigenvalues).

5) Solve for C_r^{γ} the lin syst of equations $T^{\alpha\beta\gamma} = \sum_r a_r^{\alpha} b_r^{\beta} C_r^{\gamma}$.