Multiway models

02582

Morten Mørup

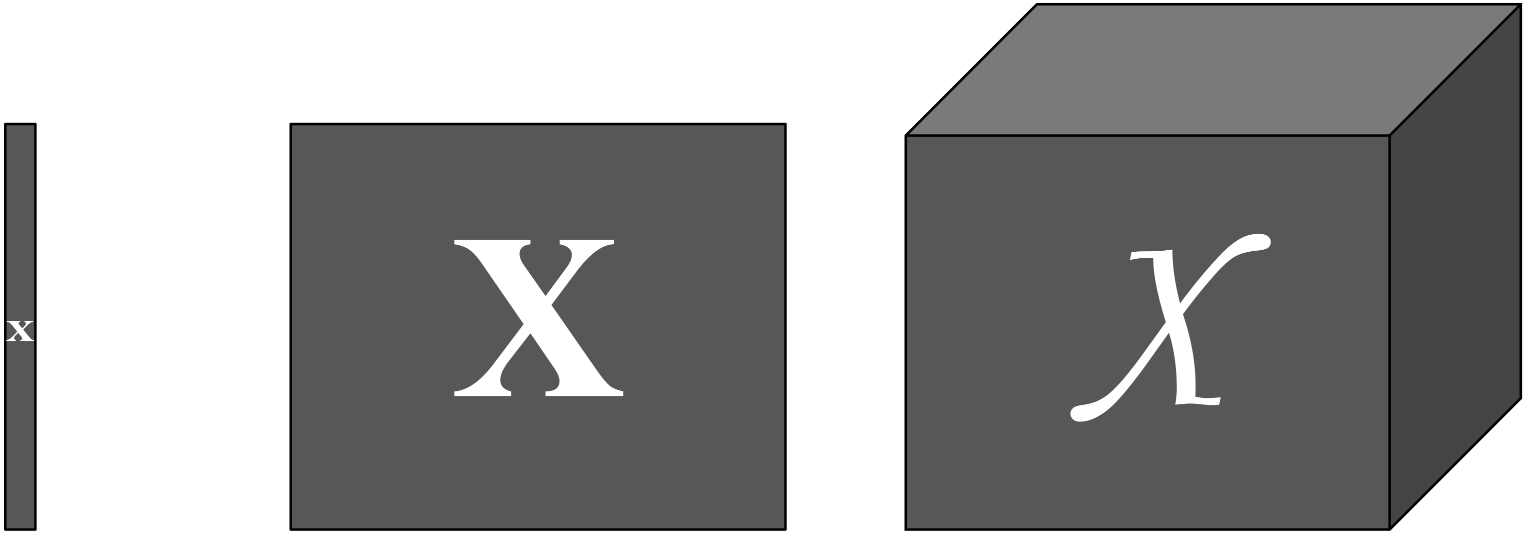
Today's lecture

- A brief history of multi-way analysis
- Tensor Nomenclature
- Tucker Decomposition
- CandeComp/PARAFAC (CP)
- Core Consistency Diagnostic
- Other tensor factorization models
- Missing values
- Software
- Applications

A brief history of multiway analysis

What is a Tensor / Multi-way array?

Tensors, or multi-way arrays, are generalizations of vectors (1st order tensors) and matrices (2nd order tensors) to arrays of general orders ($N > 2$). As such, a 3rd order tensor is an array with elements $x_{i,j,k}$.

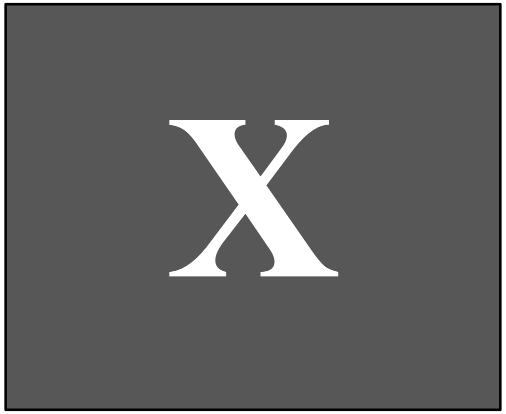


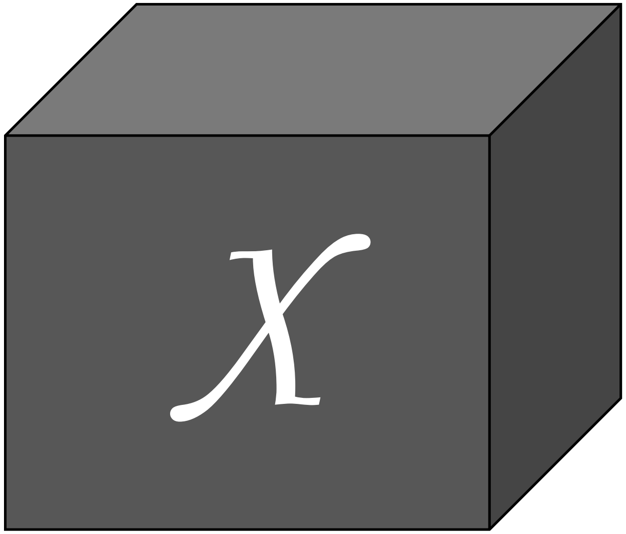
Vector
Order 1 tensor
x_i

Matrix
Order 2 tensor
$x_{i,j}$

3-way array
Order 3 tensor
$x_{i,j,k}$

Three-way structure widely ignored in many fields of research!



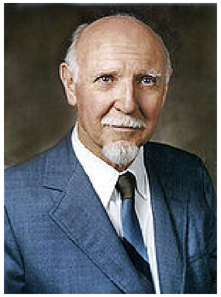


Why bother about tensor decomposition?

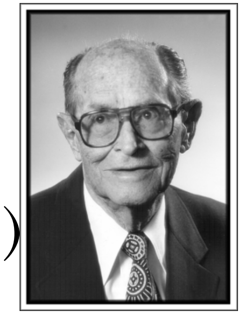
- Tensor decomposition admit uniqueness of the decomposition without additional constraints such as orthogonality and independence
- Tensor decomposition methods can identify components even when facing very poor signal to noise ratios (SNR) and when only a relatively small fraction of all the data is observed.
- Tensor decomposition can explicitly take into account the multi-way structure of the data that would otherwise be lost when analyzing the data by collapsing some of the modes to form a matrix

History of Tensor factorization

- **Hitchcock 1927 (not the filmmaker!)**
  generalized 2-way rank to n-way (i.e. proposed the CP-model, see later slides) as well as introduced
  the notion of n-mode rank
- **Cattell 1944**
  Parallel Proportional Profiles (to resolve rotational indeterminacy in factor analysis)
- **Tucker 1963**
  Proposed the Tucker model (see later slides)
- **Harshman and Carrol & Chang 1970**
  Independently proposed the PARAFAC and CanDecomp models (CP model, see later slides)



Cattell: Also very famous for 16 personality factor model and the 16PF Questionnaire



Tucker

Harshman

Carrol

Tensor vs Matrix decomposition

Factorizing tensors has several advantages over two-way matrix factorization

- Uniqueness
- Component identification even when only a relatively small fraction of all the data is observed.
- Multi-way decomposition techniques can explicitly take into account the multi-way structure of the data that would otherwise be lost when analyzing the data by matrix factorization approaches by collapsing some of the modes.

However, factorizing tensors has its challenges

- Its geometry is not yet fully understood
- The occurrence of so-called degenerate solutions
- Lack of guarantee of finding the optimal solution.

Tensor Nomenclature

A general tensor of order N is written

$\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$,

$\mathcal{X}^{I_1 \times I_2 \times \ldots \times I_N}$

a given element of the tensor $\mathcal{X}$ is denoted by

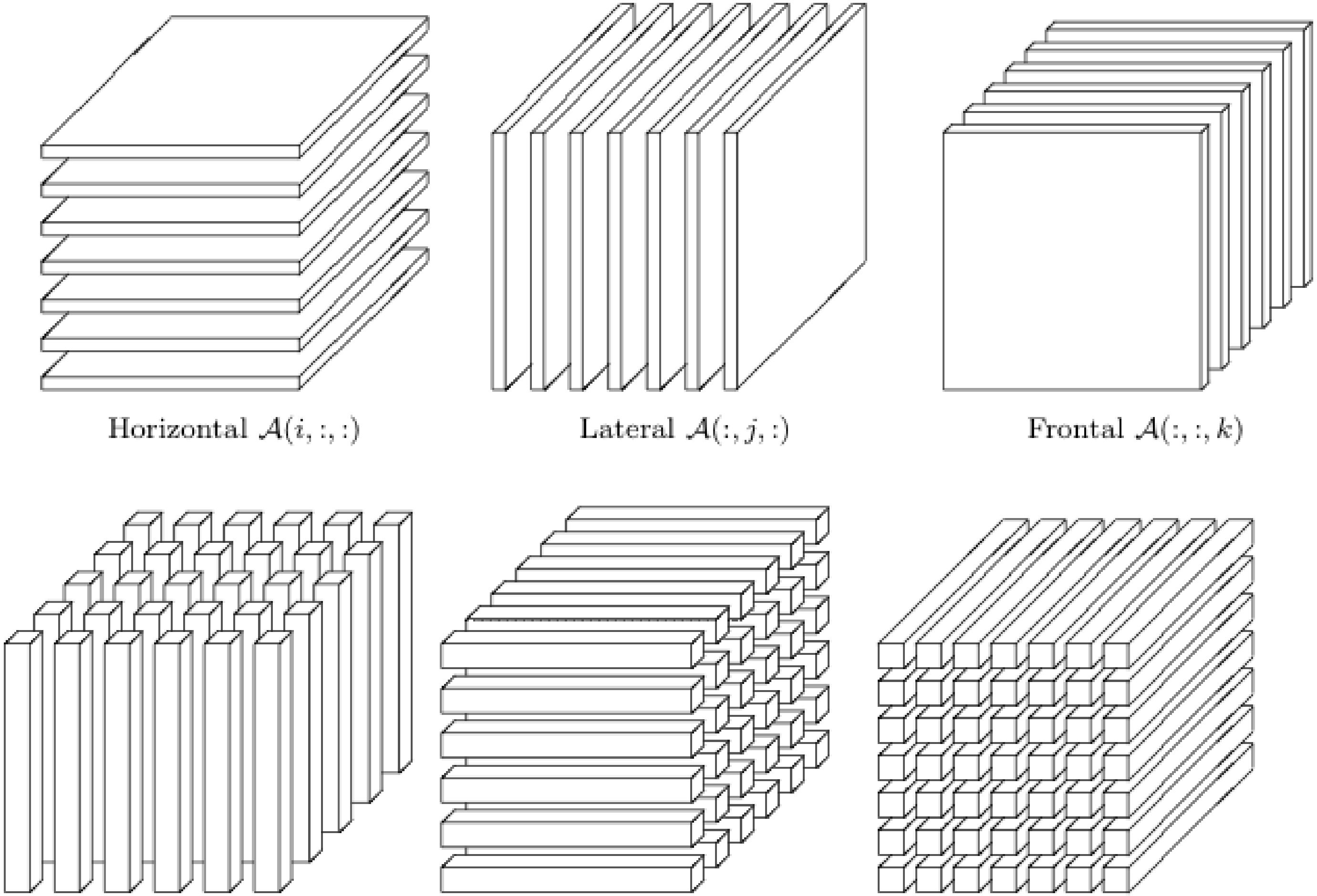
$x_{i_1, i_2, \ldots, i_N}$

Consider the third order tensor $\mathcal{A}^{I \times J \times K}$ and $\mathcal{B}^{I \times J \times K}$. Scalar multiplication, addition of two tensors and the inner product between two tensors are given by

$$\begin{aligned} \alpha\mathcal{B} &= \mathcal{C}, \quad \text{where} \quad c_{i,j,k} = \alpha b_{i,j,k} && (1) \\ \mathcal{A} + \mathcal{B} &= \mathcal{C}, \quad \text{where} \quad c_{i,j,k} = a_{i,j,k} + b_{i,j,k} && (2) \\ \langle \mathcal{A}, \mathcal{B} \rangle &= \sum_{i,j,k} a_{i,j,k} b_{i,j,k} && (3) \end{aligned}$$

As such, the Frobenius norm of a tensor is given by $\|\mathcal{A}\|_F = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$.

Slices and fibers



Horizontal $\mathcal{A}(i,:,:)$

Lateral $\mathcal{A}(:,j,:)$

Frontal $\mathcal{A}(:,:,k)$

Mode-1 — Columns
$\mathcal{A}(:,j,k)$

Mode-2 — Rows
$\mathcal{A}(i,:,k)$

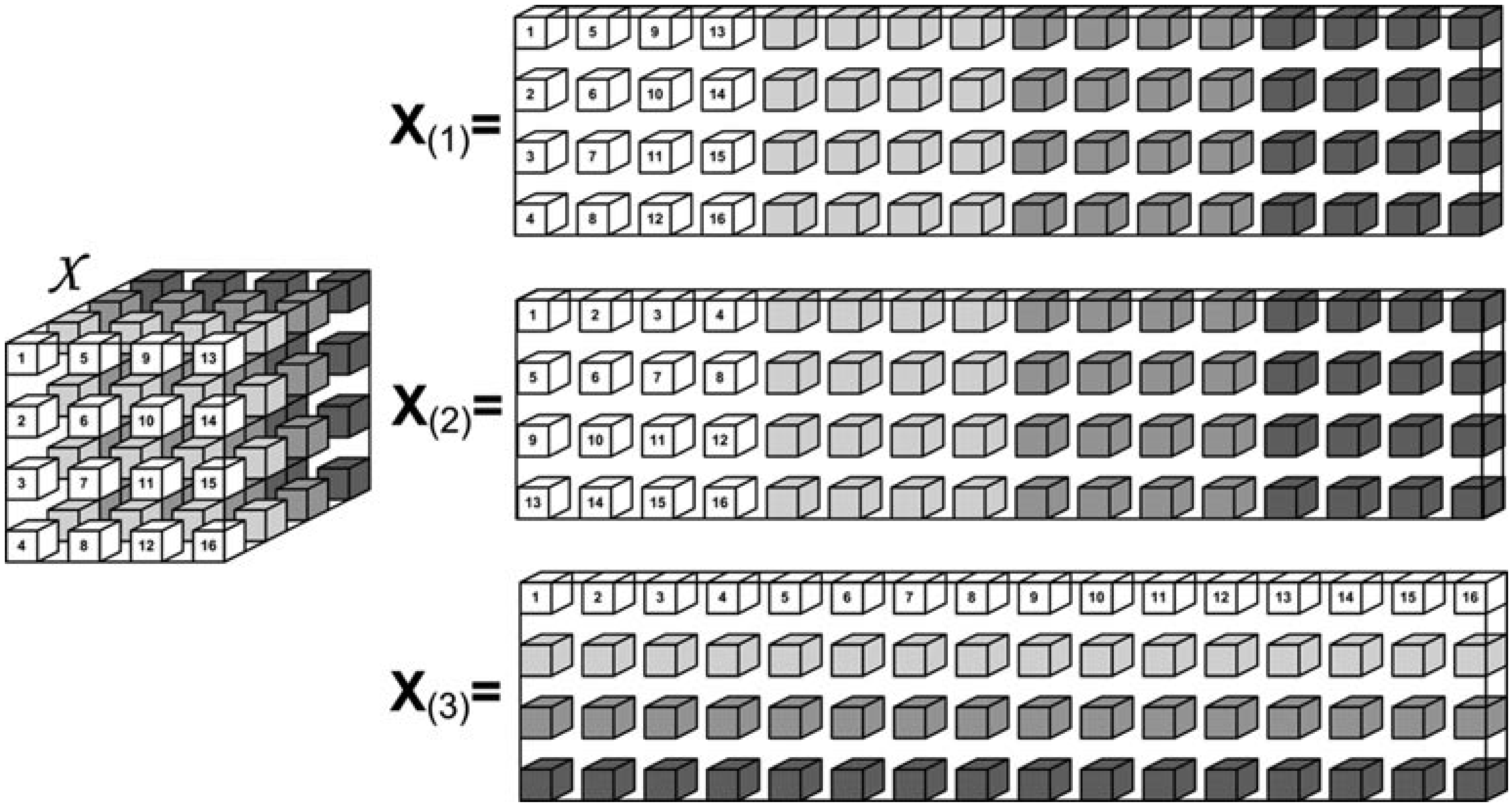
Mode-3 — Tubes
$\mathcal{A}(i,j,:)$

Matricizing and un-matricizing

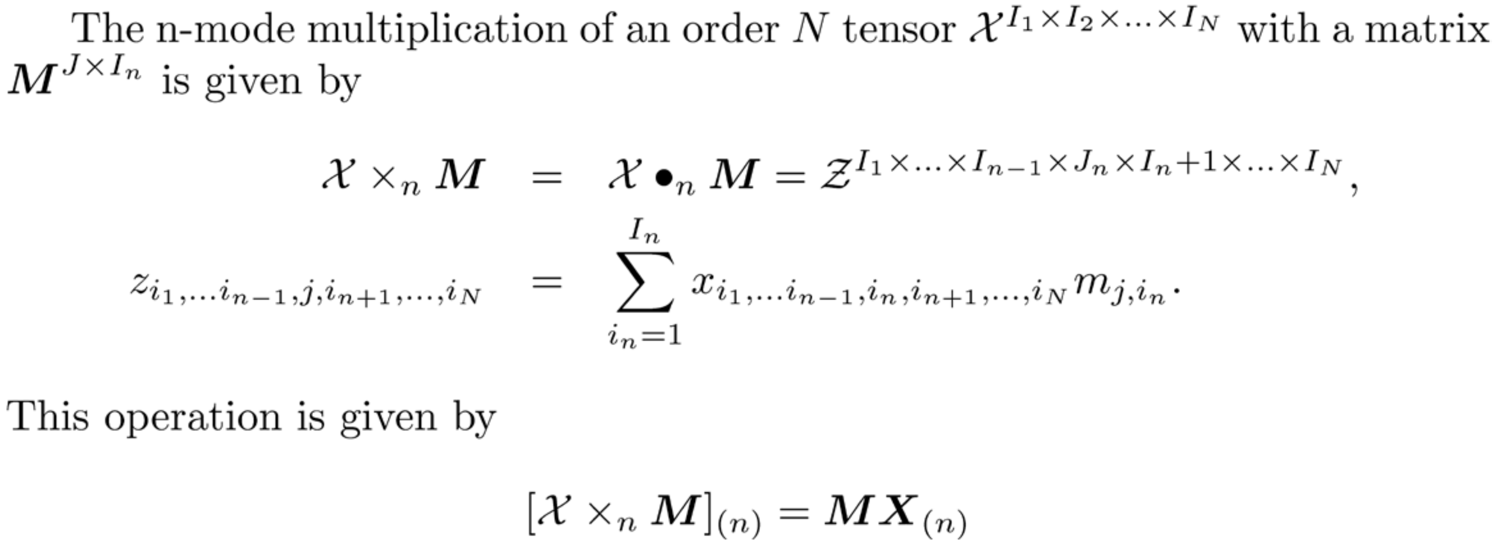
The n^{th} mode matricizing and un-matricizing operation maps a tensor into a matrix and a matrix into a tensor respectively, i.e.

$$\mathcal{X}^{I_1 \times I_2 \times \ldots \times I_N} \quad \underset{matricizing}{\rightarrow} \quad \boldsymbol{X}_{(n)}^{I_n \times I_1 \cdot I_2 \cdots I_{n-1} \cdot I_{n+1} \cdots I_N}$$

$$\boldsymbol{X}_{(n)}^{I_n \times I_1 \cdot I_2 \cdots I_{n-1} \cdot I_{n+1} \cdots I_N} \quad \underset{un-matricizing}{\rightarrow} \quad \mathcal{X}^{I_1 \times I_2 \times \ldots \times I_N}$$



N-mode multiplication



The n-mode multiplication of an order N tensor $\mathcal{X}^{I_1 \times I_2 \times \ldots \times I_N}$ with a matrix $\boldsymbol{M}^{J \times I_n}$ is given by

$$\begin{aligned} \mathcal{X} \times_n \boldsymbol{M} &= \mathcal{X} \bullet_n \boldsymbol{M} = \mathcal{Z}^{I_1 \times \ldots \times I_{n-1} \times J_n \times I_n + 1 \times \ldots \times I_N}, \\ z_{i_1, \ldots i_{n-1}, j, i_{n+1}, \ldots, i_N} &= \sum_{i_n=1}^{I_n} x_{i_1, \ldots i_{n-1}, i_n, i_{n+1}, \ldots, i_N} m_{j, i_n}. \end{aligned}$$

This operation is given by

$$[\mathcal{X} \times_n \boldsymbol{M}]_{(n)} = \boldsymbol{M} \boldsymbol{X}_{(n)}$$

The SVD as n-mode multiplication

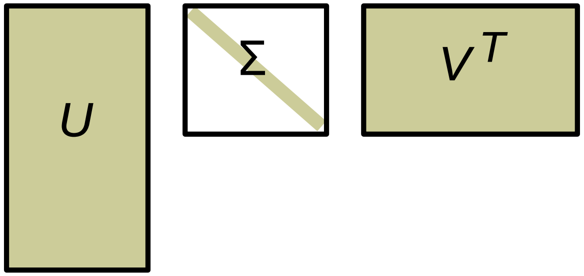
- Notation generalizes matrix-matrix products:

$$\mathbf{A} \times_1 \mathbf{U} = \mathbf{U}\mathbf{A}$$

$$\mathbf{A} \times_2 \mathbf{V} = \mathbf{A}\mathbf{V}^T$$

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{\Sigma} \times_1 \mathbf{U} \times_2 \mathbf{V}.$$

- Can express matrix SVD with n-mode products:



$$Z = X \pounds_n M \quad \$ \quad Z_{(n)} = M X_{(n)}$$

Kronecker and Khatri-Rao product

Kronecker product

$$\boldsymbol{P}^{\mathrm{I}\times\mathrm{J}} \otimes \boldsymbol{Q}^{\mathrm{K}\times\mathrm{L}} = \boldsymbol{R}^{\mathrm{IK}\times\mathrm{JL}}, \qquad \text{such that} \qquad r_{k+K(i-1),l+L(j-1)} = p_{ij}q_{kl},$$

$$\boldsymbol{P} \otimes \boldsymbol{Q} = \begin{bmatrix} p_{11}\boldsymbol{Q} & \cdots & p_{1J}\boldsymbol{Q} \\ \vdots & \ddots & \vdots \\ p_{I1}\boldsymbol{Q} & \cdots & p_{IJ}\boldsymbol{Q} \end{bmatrix}$$

Khatri-Rao product

$$\boldsymbol{A}^{\mathrm{I}\times\mathrm{J}} | \otimes | \boldsymbol{B}^{\mathrm{K}\times\mathrm{J}} = \boldsymbol{A}^{\mathrm{I}\times\mathrm{J}} \odot \boldsymbol{B}^{\mathrm{K}\times\mathrm{J}} = \boldsymbol{C}^{\mathrm{IK}\times\mathrm{J}}, \qquad \text{such that} \qquad c_{k+K(i-1),j} = a_{ij}b_{kj}.$$

$$\boldsymbol{A} \odot \boldsymbol{B} = \left[\begin{array}{ccc} \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 & \cdots & \boldsymbol{a}_J \otimes \boldsymbol{b}_J \end{array}\right] = \begin{bmatrix} \boldsymbol{a}_{11}\boldsymbol{b}_1 & \cdots & a_{1J}\boldsymbol{b}_J \\ \vdots & \vdots & \vdots \\ a_{I1}\boldsymbol{b}_1 & \cdots & a_{IJ}\boldsymbol{b}_J \end{bmatrix}$$

.

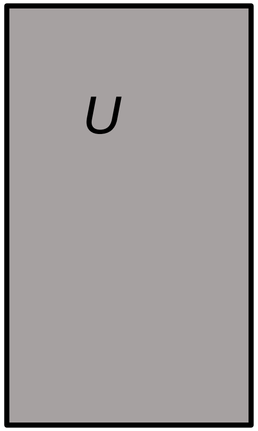
$$(\boldsymbol{A} \odot \boldsymbol{B})^{\dagger} \quad = \quad [(\boldsymbol{A}^{\top}\boldsymbol{A}) * (\boldsymbol{B}^{\top}\boldsymbol{B})]^{-1}(\boldsymbol{A} \odot \boldsymbol{B})^{\top}$$

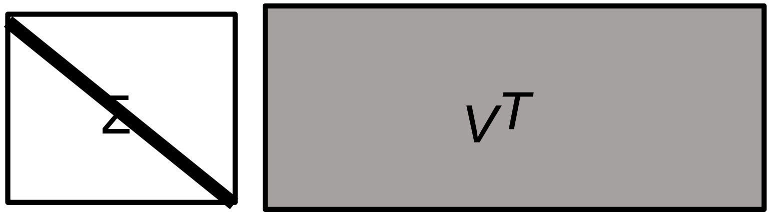
Let $\boldsymbol{P} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$

What is $\boldsymbol{P} \otimes \boldsymbol{Q} \quad \textit{and} \quad \boldsymbol{P} \odot \boldsymbol{Q}$?

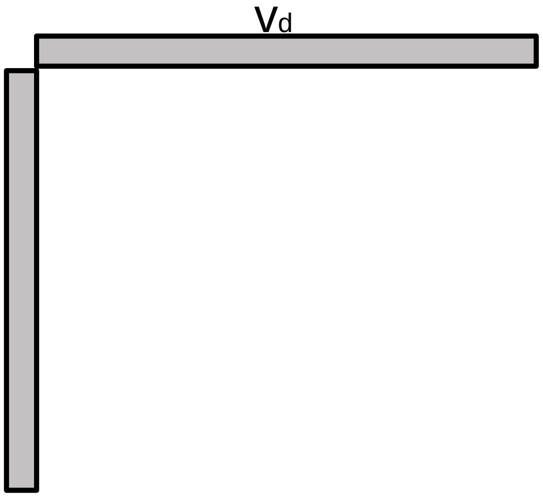
Outer product

$$\boldsymbol{a} \circ \boldsymbol{b} = \boldsymbol{Z} \quad \text{such that} \quad z_{i_1 i_2} = a_{i_1} b_{i_2}$$

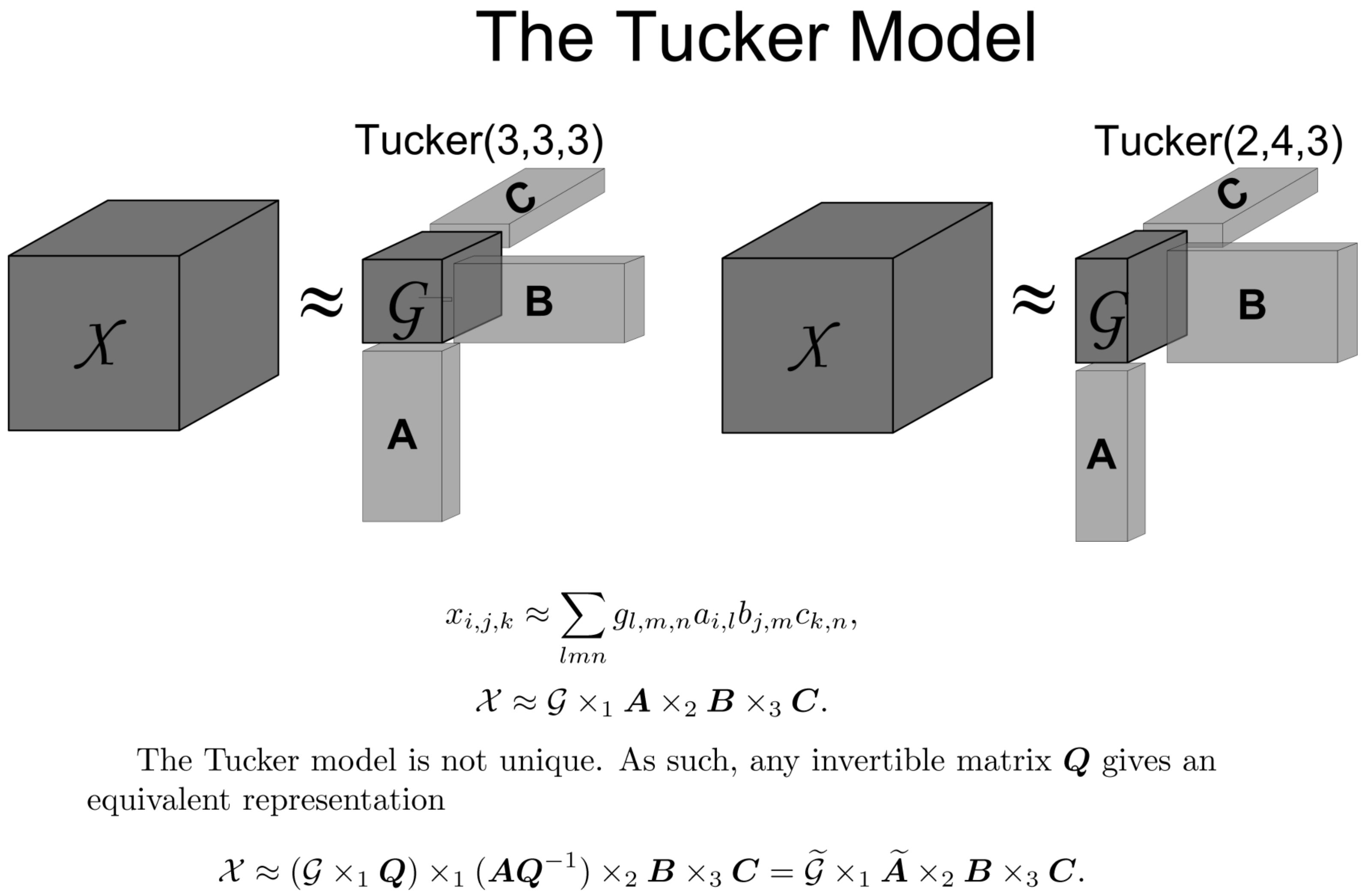




$$\begin{aligned} \boldsymbol{U\Sigma V}^\top &= \sum_d \sigma_d \boldsymbol{u}_d \circ \boldsymbol{v}_d \\ &= \sum_d \sigma_d \end{aligned}$$



Tucker model



The Tucker Model

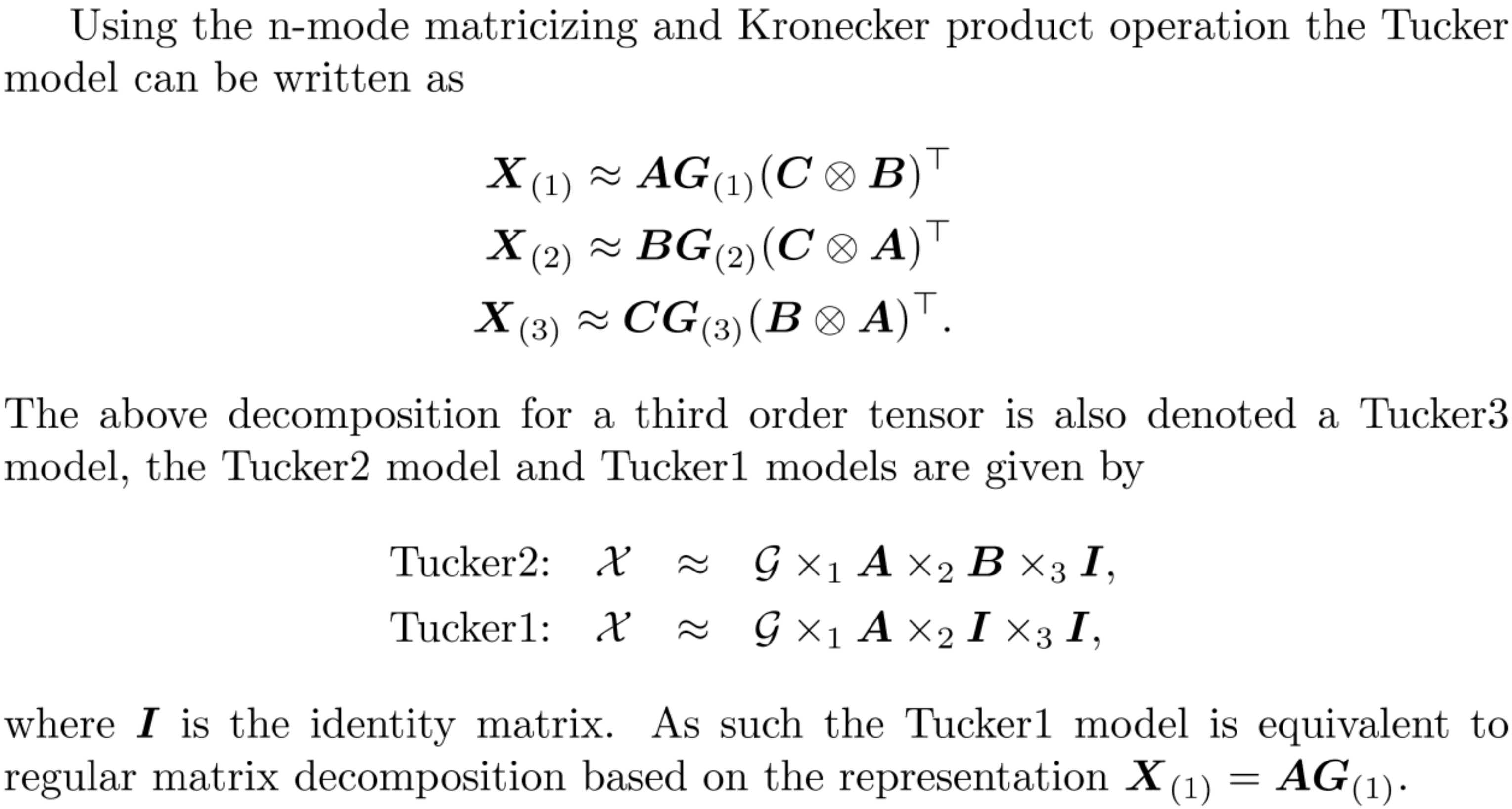
$$x_{i,j,k} \approx \sum_{lmn} g_{l,m,n} a_{i,l} b_{j,m} c_{k,n},$$

$$\mathcal{X} \approx \mathcal{G} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}.$$

The Tucker model is not unique. As such, any invertible matrix $\boldsymbol{Q}$ gives an equivalent representation

$$\mathcal{X} \approx (\mathcal{G} \times_1 \boldsymbol{Q}) \times_1 (\boldsymbol{A}\boldsymbol{Q}^{-1}) \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} = \widetilde{\mathcal{G}} \times_1 \widetilde{\boldsymbol{A}} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}.$$

Tucker model in matrix notation



Using the n-mode matricizing and Kronecker product operation the Tucker model can be written as

$$\begin{aligned}\boldsymbol{X}_{(1)} &\approx \boldsymbol{A}\boldsymbol{G}_{(1)}(\boldsymbol{C}\otimes\boldsymbol{B})^\top\\ \boldsymbol{X}_{(2)} &\approx \boldsymbol{B}\boldsymbol{G}_{(2)}(\boldsymbol{C}\otimes\boldsymbol{A})^\top\\ \boldsymbol{X}_{(3)} &\approx \boldsymbol{C}\boldsymbol{G}_{(3)}(\boldsymbol{B}\otimes\boldsymbol{A})^\top.\end{aligned}$$

The above decomposition for a third order tensor is also denoted a Tucker3 model, the Tucker2 model and Tucker1 models are given by

$$\begin{array}{lccc}\text{Tucker2:} & \mathcal{X} & \approx & \mathcal{G}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{I},\\ \text{Tucker1:} & \mathcal{X} & \approx & \mathcal{G}\times_1\boldsymbol{A}\times_2\boldsymbol{I}\times_3\boldsymbol{I},\end{array}$$

where $\boldsymbol{I}$ is the identity matrix. As such the Tucker1 model is equivalent to regular matrix decomposition based on the representation $\boldsymbol{X}_{(1)} = \boldsymbol{A}\boldsymbol{G}_{(1)}$.

Model estimation

$$\begin{aligned}
\boldsymbol{A} &\leftarrow \boldsymbol{X}_{(1)}(\boldsymbol{G}_{(1)}(\boldsymbol{C}\otimes\boldsymbol{B})^\top)^\dagger = \boldsymbol{X}_{(n)}((\boldsymbol{C}^\dagger\otimes\boldsymbol{B}^\dagger)^\top\boldsymbol{G}_{(1)}^\dagger)\\
\boldsymbol{B} &\leftarrow \boldsymbol{X}_{(2)}(\boldsymbol{G}_{(2)}(\boldsymbol{C}\otimes\boldsymbol{A})^\top)^\dagger = \boldsymbol{X}_{(2)}((\boldsymbol{C}^\dagger\otimes\boldsymbol{A}^\dagger)^\top\boldsymbol{G}_{(2)}^\dagger)\\
\boldsymbol{C} &\leftarrow \boldsymbol{X}_{(3)}(\boldsymbol{G}_{(3)}(\boldsymbol{B}\otimes\boldsymbol{A})^\top)^\dagger = \boldsymbol{X}_{(3)}((\boldsymbol{B}^\dagger\otimes\boldsymbol{A}^\dagger)^\top\boldsymbol{G}_{(3)}^\dagger)\\
\mathcal{G} &\leftarrow \mathcal{X}\times_1\boldsymbol{A}^\dagger\times_2\boldsymbol{B}^\dagger\times_3\boldsymbol{C}^\dagger.
\end{aligned}$$

Imposing orthogonality we find

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{S}^{(1)}\boldsymbol{V}^{(1)^\top} &= \boldsymbol{X}_{(1)}(\boldsymbol{C}\otimes\boldsymbol{B}),\\
\boldsymbol{B}\boldsymbol{S}^{(2)}\boldsymbol{V}^{(2)^\top} &= \boldsymbol{X}_{(2)}(\boldsymbol{C}\otimes\boldsymbol{A}),\\
\boldsymbol{C}\boldsymbol{S}^{(3)}\boldsymbol{V}^{(3)^\top} &= \boldsymbol{X}_{(3)}(\boldsymbol{B}\otimes\boldsymbol{A}).
\end{aligned}$$

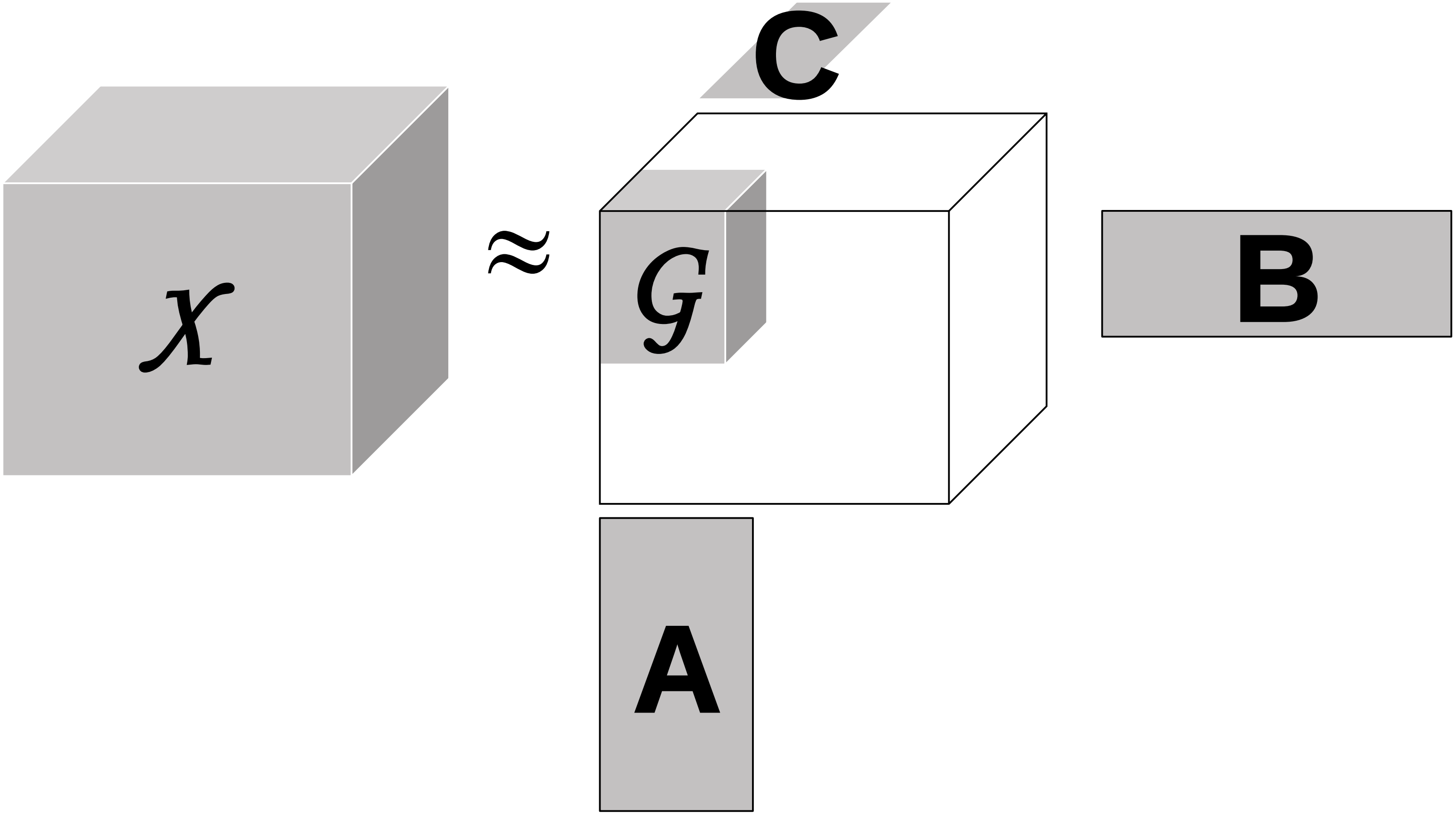
A special case is given by the HOSVD approach where the loadings of each mode is determined by the SVD

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{S}^{(1)}\boldsymbol{V}^{(1)^\top} &= \boldsymbol{X}_{(1)},\\
\boldsymbol{B}\boldsymbol{S}^{(2)}\boldsymbol{V}^{(2)^\top} &= \boldsymbol{X}_{(2)},\\
\boldsymbol{C}\boldsymbol{S}^{(3)}\boldsymbol{V}^{(3)^\top} &= \boldsymbol{X}_{(3)}.
\end{aligned}$$

Show that if **A, B** and **C** are orthonormal, i.e. **A**$^\top$**A=I**, **B**$^\top$**B=I** and **C**$^\top$**C=I**, we obtain

$$\|\mathcal{X}-\mathcal{G}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C}\|_F^2=\sum_{ijk}(x_{i,j,k}-\sum_{lmn}g_{l,m,n}a_{i,l}b_{j,m}c_{k,n})^2=\|\mathcal{X}\|_F^2-\|\mathcal{G}\|_F^2$$

The Tucker model is particularly useful for compression

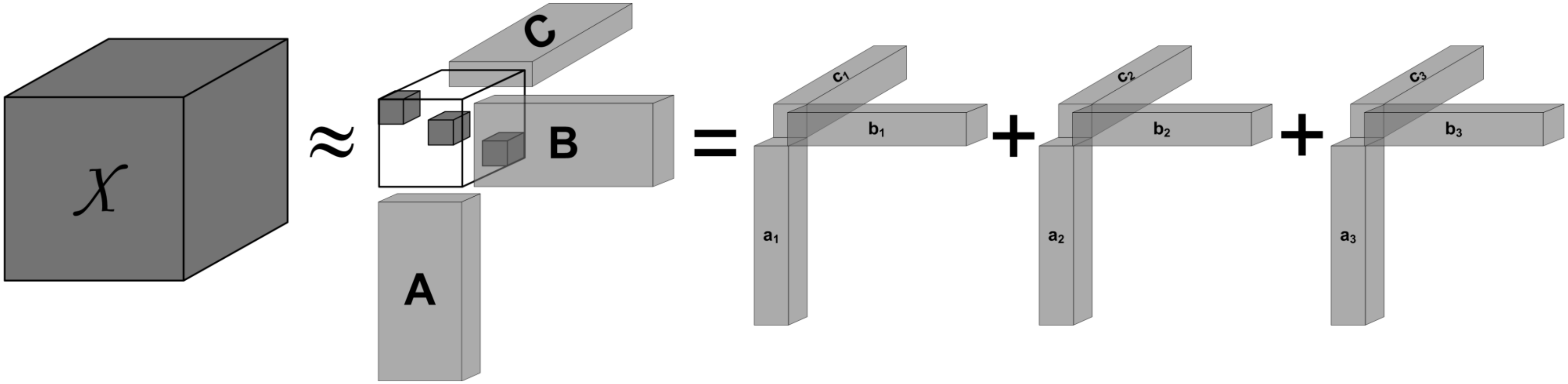


CanDecomp/PARAFAC (CP) model

$$x_{i,j,k} \approx \sum_d a_{i,d} b_{j,d} c_{k,d}$$

$$\mathcal{X} \approx \mathcal{D} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} = \mathcal{I} \times_1 \hat{\mathbf{A}} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C},$$

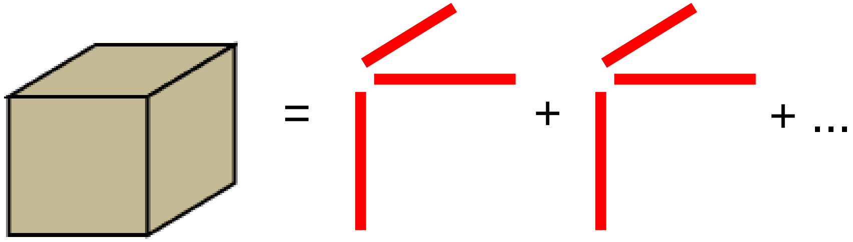
Canoncical Decomposition/PARAFAC (CP)



Many ways of writing the CP model

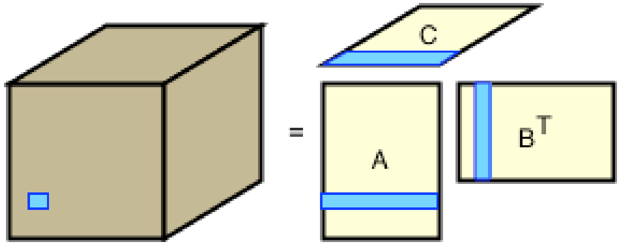
Outer product form

$$\mathcal{X} \approx \sum_{i=1}^{r} \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i$$



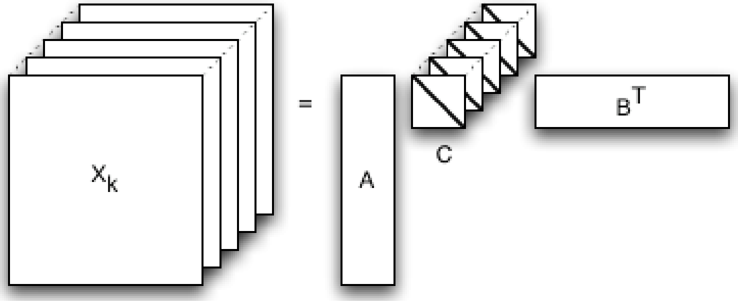
Scalar form

$$x_{ijk} \approx \sum_{i=1}^{r} a_{ir} b_{jr} c_{kr}$$



Tensor slice form

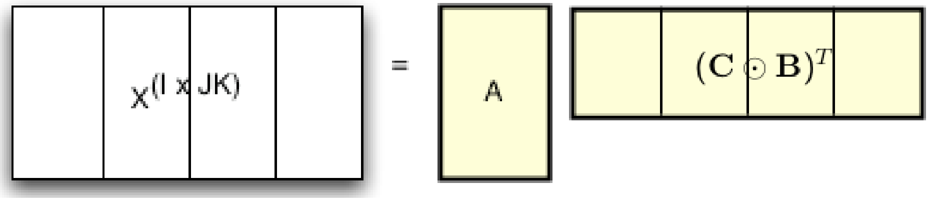
$$\mathbf{X}_k \approx \mathbf{A}\ \mathrm{diag}(\mathbf{c}_{k:})\mathbf{B}^T$$



Matrix form

$$\mathbf{X}_{(1)} \approx \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T$$

Matricized array



When multiplying by an invertible matrix $\boldsymbol{Q}$ we find

$$\mathcal{X} \approx (\mathcal{D} \times_1 \boldsymbol{Q} \times_2 \boldsymbol{R} \times_3 \boldsymbol{S}) \times_1 (\boldsymbol{A}\boldsymbol{Q}^{-1}) \times_2 (\boldsymbol{B}\boldsymbol{R}^{-1}) \times_3 (\boldsymbol{C}\boldsymbol{Q}^{-1}) = \widetilde{\mathcal{D}} \times_1 \widetilde{\boldsymbol{A}} \times_2 \widetilde{\boldsymbol{B}} \times_3 \widetilde{\boldsymbol{C}}.$$

Do you think the CP model is unique?

As such, the new core $\widetilde{\mathcal{D}}$ must be non-zero only along the diagonal for the representation to remain a CP model. This in practice has the consequence that $\boldsymbol{Q}$ can only be a scale and permutation matrix. The uniqueness properties of the CP model were thoroughly investigated by J.B. Kruskal and among several results the following uniqueness criterion derived

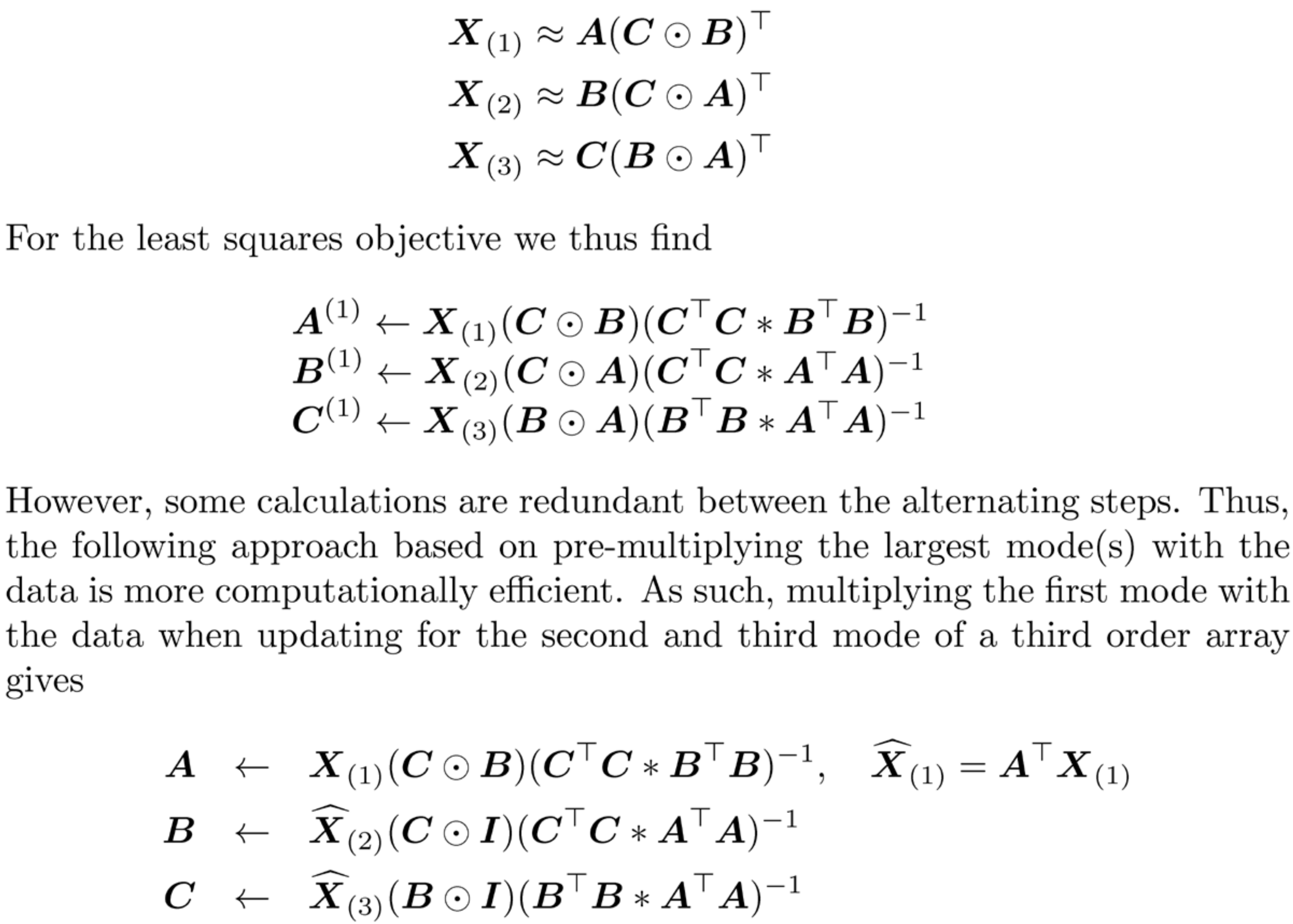
$$k_{\boldsymbol{A}} + k_{\boldsymbol{B}} + k_{\boldsymbol{C}} \geq 2D + 2. \tag{1}$$

Here, the Kruskal rank or k-rank $k_{\boldsymbol{A}}$ of a matrix $\boldsymbol{A}$ is the maximal number r such that any set of r columns of the matrix $\boldsymbol{A}$ is linearly independent. As such, $k_{\boldsymbol{A}} \leq rank(\boldsymbol{A})$. As Kruskal wrote in, struck by his own uniqueness criterion,

"A surprising fact is that the nonrotatability characteristic can hold even when the number of factors extracted is greater than every dimension of the three-way array."

The criterion has been generalized to order N arrays.

CP Model estimation



$$\boldsymbol{X}_{(1)} \approx \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^\top$$
$$\boldsymbol{X}_{(2)} \approx \boldsymbol{B}(\boldsymbol{C} \odot \boldsymbol{A})^\top$$
$$\boldsymbol{X}_{(3)} \approx \boldsymbol{C}(\boldsymbol{B} \odot \boldsymbol{A})^\top$$

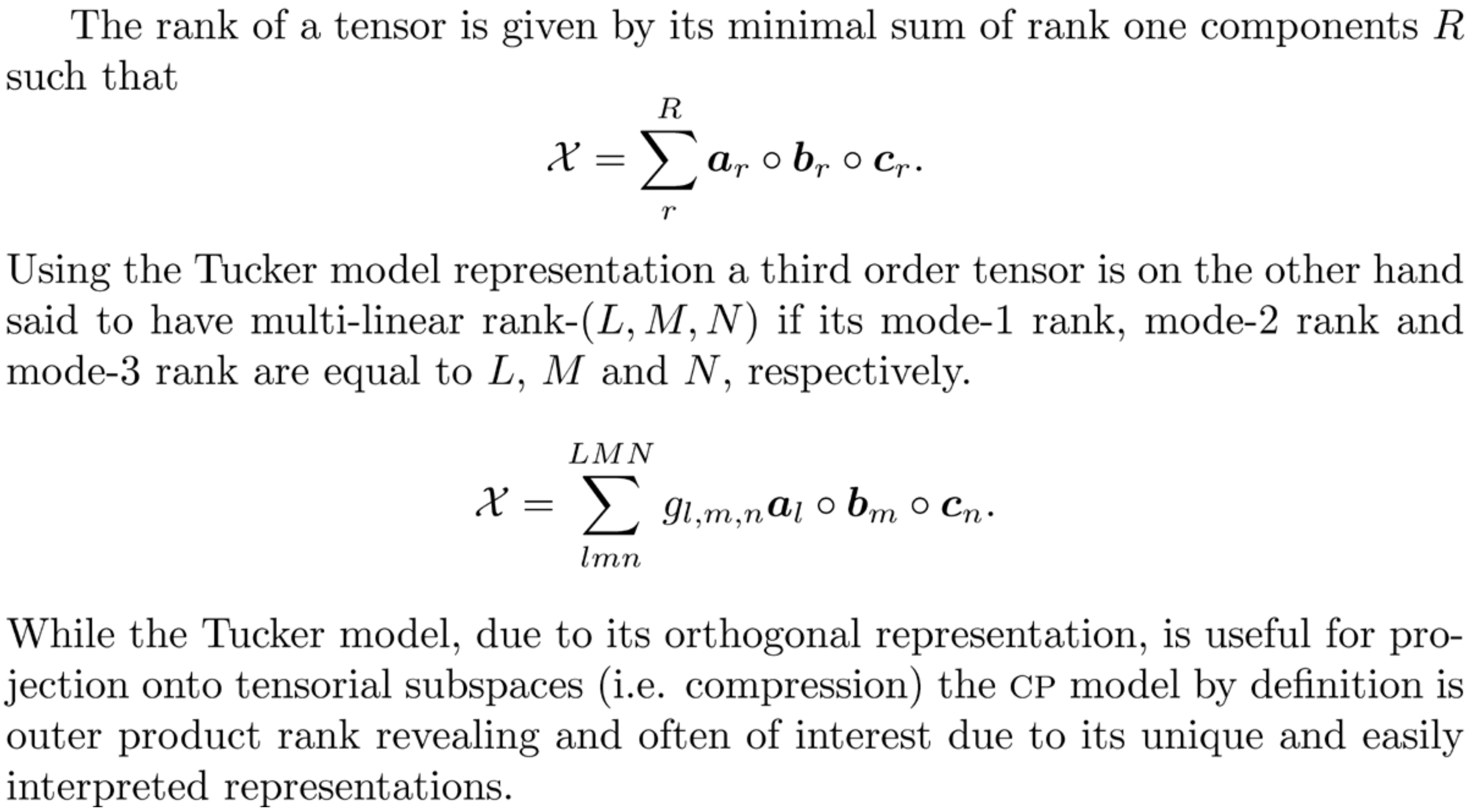
For the least squares objective we thus find

$$\begin{aligned}
\boldsymbol{A}^{(1)} &\leftarrow \boldsymbol{X}_{(1)}(\boldsymbol{C} \odot \boldsymbol{B})(\boldsymbol{C}^\top\boldsymbol{C} * \boldsymbol{B}^\top\boldsymbol{B})^{-1}\\
\boldsymbol{B}^{(1)} &\leftarrow \boldsymbol{X}_{(2)}(\boldsymbol{C} \odot \boldsymbol{A})(\boldsymbol{C}^\top\boldsymbol{C} * \boldsymbol{A}^\top\boldsymbol{A})^{-1}\\
\boldsymbol{C}^{(1)} &\leftarrow \boldsymbol{X}_{(3)}(\boldsymbol{B} \odot \boldsymbol{A})(\boldsymbol{B}^\top\boldsymbol{B} * \boldsymbol{A}^\top\boldsymbol{A})^{-1}
\end{aligned}$$

However, some calculations are redundant between the alternating steps. Thus, the following approach based on pre-multiplying the largest mode(s) with the data is more computationally efficient. As such, multiplying the first mode with the data when updating for the second and third mode of a third order array gives

$$\begin{aligned}
\boldsymbol{A} &\leftarrow \boldsymbol{X}_{(1)}(\boldsymbol{C} \odot \boldsymbol{B})(\boldsymbol{C}^\top\boldsymbol{C} * \boldsymbol{B}^\top\boldsymbol{B})^{-1}, \quad \widehat{\boldsymbol{X}}_{(1)} = \boldsymbol{A}^\top\boldsymbol{X}_{(1)}\\
\boldsymbol{B} &\leftarrow \widehat{\boldsymbol{X}}_{(2)}(\boldsymbol{C} \odot \boldsymbol{I})(\boldsymbol{C}^\top\boldsymbol{C} * \boldsymbol{A}^\top\boldsymbol{A})^{-1}\\
\boldsymbol{C} &\leftarrow \widehat{\boldsymbol{X}}_{(3)}(\boldsymbol{B} \odot \boldsymbol{I})(\boldsymbol{B}^\top\boldsymbol{B} * \boldsymbol{A}^\top\boldsymbol{A})^{-1}
\end{aligned}$$

Rank and Multilinear rank



The rank of a tensor is given by its minimal sum of rank one components R such that

$$\mathcal{X} = \sum_{r}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r.$$

Using the Tucker model representation a third order tensor is on the other hand said to have multi-linear rank-(L, M, N) if its mode-1 rank, mode-2 rank and mode-3 rank are equal to L, M and N, respectively.

$$\mathcal{X} = \sum_{lmn}^{LMN} g_{l,m,n} \boldsymbol{a}_l \circ \boldsymbol{b}_m \circ \boldsymbol{c}_n.$$

While the Tucker model, due to its orthogonal representation, is useful for projection onto tensorial subspaces (i.e. compression) the CP model by definition is outer product rank revealing and often of interest due to its unique and easily interpreted representations.

Core Consistency Diagnostic

For the CP model a common heuristic approach for evaluating the number of components is based on the so-called core consistency diagnostic (CCD) has been proposed

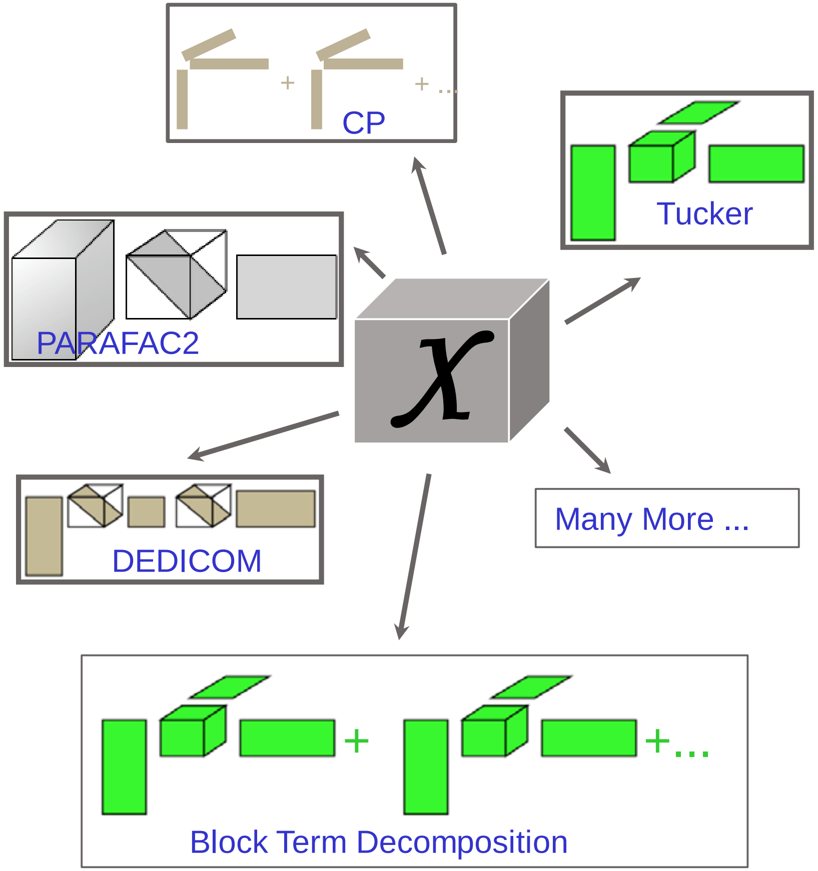
$$\begin{aligned} \mathcal{G} &\leftarrow \mathcal{X} \times_1 \boldsymbol{A}^{\dagger}_{\text{CP}} \times_2 \boldsymbol{B}^{\dagger}_{\text{CP}} \times_3 \boldsymbol{C}^{\dagger}_{\text{CP}}, \\ \text{CCD} &= 100 \cdot \Big(1 - \frac{\|\mathcal{I} - \mathcal{G}\|_F^2}{\|\mathcal{I}\|_F^2}\Big). \end{aligned}$$

What is CCD measuring?

Where $\mathcal{G}$ is the corresponding Tucker core array obtained from the loadings $\boldsymbol{A}_{cp}$, $\boldsymbol{B}_{cp}$ and $\boldsymbol{C}_{cp}$ extracted from the CP model. Too many components will result in a strong degree of cross talk across the loadings of the modes and will yield a low value of the CCD. Too few components will not have any cross-talk at all. Thus, the "correct" number of components is taken to be just before a major drop-off in the curve of (d, CCD). As explained by Bro and Kiers

"As a rule of thumb, a core consistency above 90% can be interpreted as 'very trilinear', whereas a core consistency in the neighborhood of 50% would mean a problematic model with signs of both trilinear variation and variation which is not trilinear. A core consistency close to zero or even negative implies an invalid model, because the space covered by the component matrices is then not primarily describing trilinear variation."

Other tensor factorization methods



CP

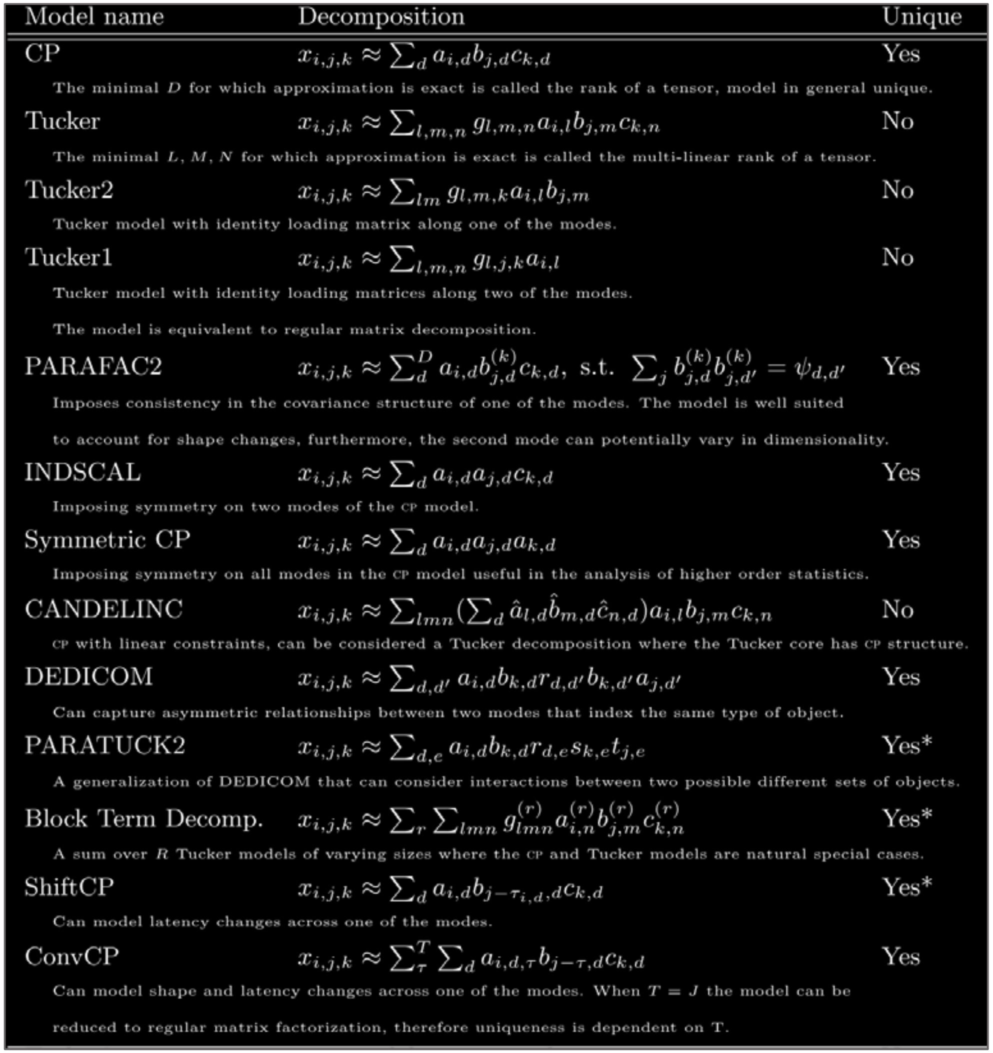
Tucker

PARAFAC2

DEDICOM

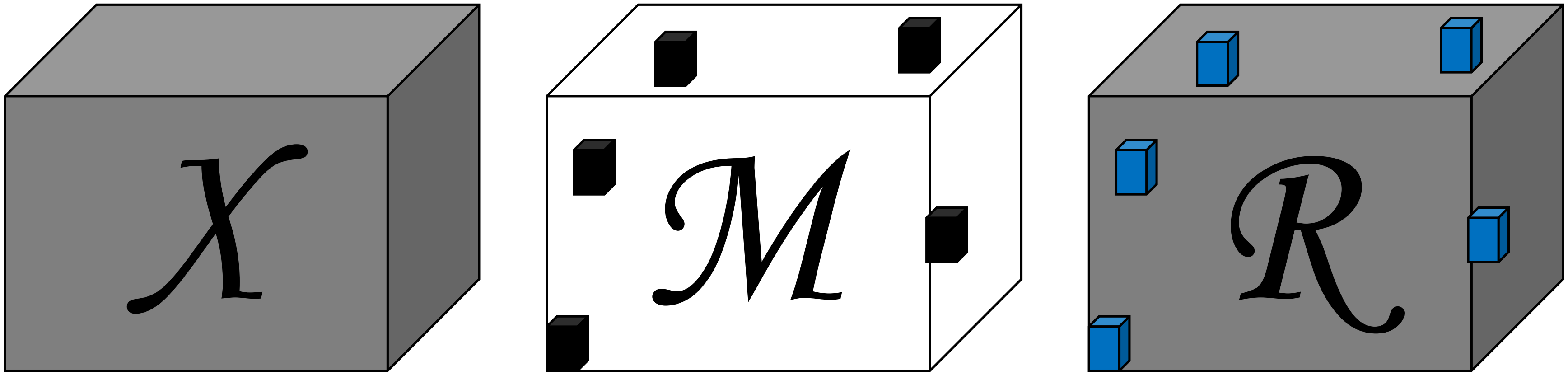
Many More ...

Block Term Decomposition



Model name	Decomposition	Unique
CP	$x_{i,j,k} \approx \sum_d a_{i,d} b_{j,d} c_{k,d}$	Yes
The minimal D for which approximation is exact is called the rank of a tensor, model in general unique.		
Tucker	$x_{i,j,k} \approx \sum_{l,m,n} g_{l,m,n} a_{i,l} b_{j,m} c_{k,n}$	No
The minimal L, M, N for which approximation is exact is called the multi-linear rank of a tensor.		
Tucker2	$x_{i,j,k} \approx \sum_{lm} g_{l,m,k} a_{i,l} b_{j,m}$	No
Tucker model with identity loading matrix along one of the modes.		
Tucker1	$x_{i,j,k} \approx \sum_{l,m,n} g_{l,j,k} a_{i,l}$	No
Tucker model with identity loading matrices along two of the modes. The model is equivalent to regular matrix decomposition.		
PARAFAC2	$x_{i,j,k} \approx \sum_d^D a_{i,d} b_{j,d}^{(k)} c_{k,d}$, s.t. $\sum_j b_{j,d}^{(k)} b_{j,d'}^{(k)} = \psi_{d,d'}$	Yes
Imposes consistency in the covariance structure of one of the modes. The model is well suited to account for shape changes, furthermore, the second mode can potentially vary in dimensionality.		
INDSCAL	$x_{i,j,k} \approx \sum_d a_{i,d} a_{j,d} c_{k,d}$	Yes
Imposing symmetry on two modes of the CP model.		
Symmetric CP	$x_{i,j,k} \approx \sum_d a_{i,d} a_{j,d} a_{k,d}$	Yes
Imposing symmetry on all modes in the CP model useful in the analysis of higher order statistics.		
CANDELINC	$x_{i,j,k} \approx \sum_{lmn} (\sum_d \hat{a}_{l,d} \hat{b}_{m,d} \hat{c}_{n,d}) a_{i,l} b_{j,m} c_{k,n}$	No
CP with linear constraints, can be considered a Tucker decomposition where the Tucker core has CP structure.		
DEDICOM	$x_{i,j,k} \approx \sum_{d,d'} a_{i,d} b_{k,d} r_{d,d'} b_{k,d'} a_{j,d'}$	Yes
Can capture asymmetric relationships between two modes that index the same type of object.		
PARATUCK2	$x_{i,j,k} \approx \sum_{d,e} a_{i,d} b_{k,d} r_{d,e} s_{k,e} t_{j,e}$	Yes*
A generalization of DEDICOM that can consider interactions between two possible different sets of objects.		
Block Term Decomp.	$x_{i,j,k} \approx \sum_r \sum_{lmn} g_{lmn}^{(r)} a_{i,n}^{(r)} b_{j,m}^{(r)} c_{k,n}^{(r)}$	Yes*
A sum over R Tucker models of varying sizes where the CP and Tucker models are natural special cases.		
ShiftCP	$x_{i,j,k} \approx \sum_d a_{i,d} b_{j-\tau_{i,d},d} c_{k,d}$	Yes*
Can model latency changes across one of the modes.		
ConvCP	$x_{i,j,k} \approx \sum_\tau^T \sum_d a_{i,d,\tau} b_{j-\tau,d} c_{k,d}$	Yes
Can model shape and latency changes across one of the modes. When $T = J$ the model can be reduced to regular matrix factorization, therefore uniqueness is dependent on T.		

Missing Values



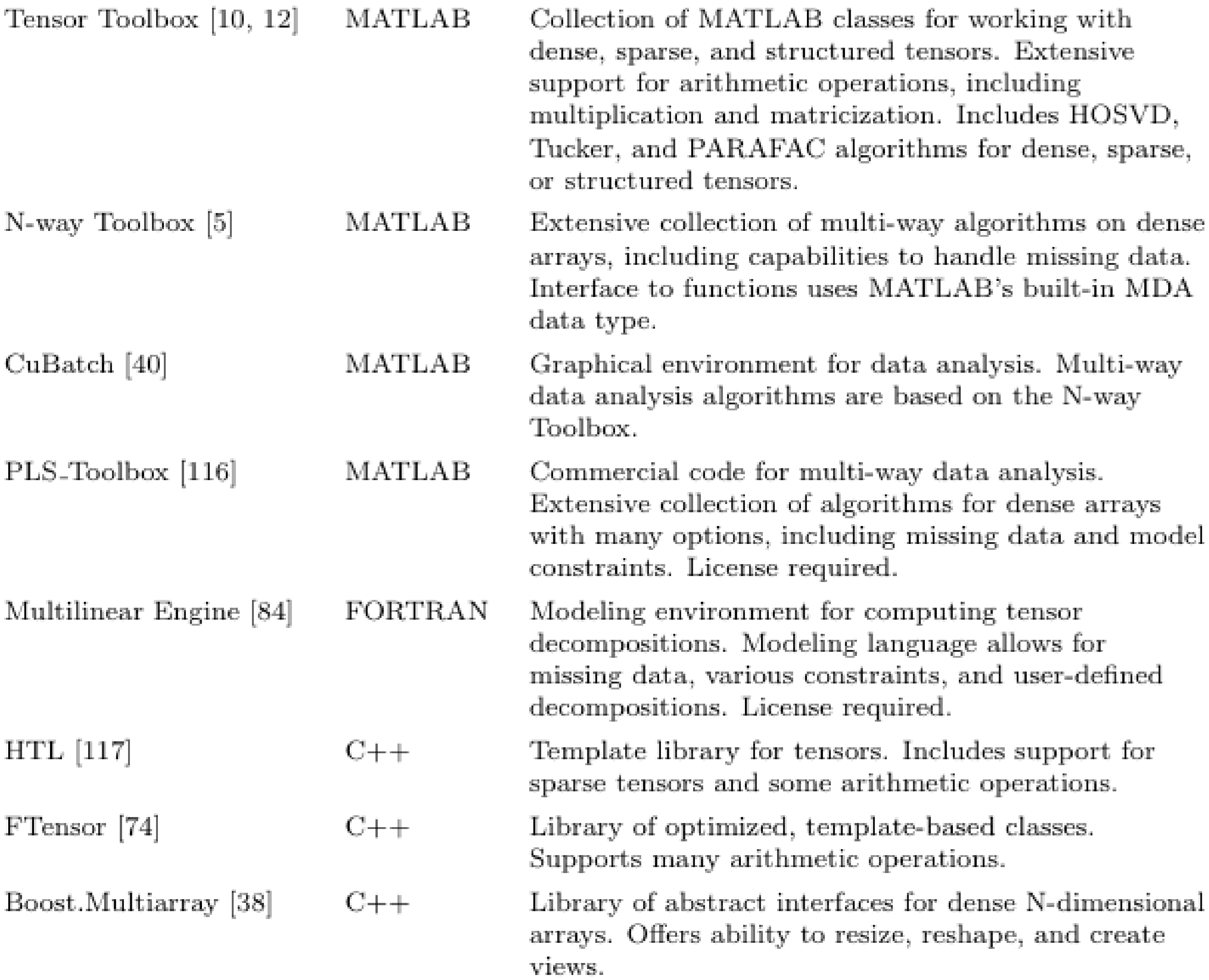
Least squares objective function:

$$\text{Marginalization:} \quad \tfrac{1}{2}\|(1-\mathcal{M})*(\mathcal{X}-\mathcal{R})\|_F^2 \quad = \frac{1}{2}\|(1-\mathcal{M})*(\mathcal{X}-\mathcal{R})\|_F^2$$

$$\text{Imputation:} \quad \tfrac{1}{2}\|(1-\mathcal{M})*(\mathcal{X}-\mathcal{R})\|_F^2 \quad = \frac{1}{2}\|(\mathcal{X}-\mathcal{R}\|_F^2, \text{ where } \|\mathcal{M}*(\mathcal{X}-\mathcal{R})\|_F^2 = 0$$

For a reasonable sized third order tensor of size I x J x K it turns out we can treat up to 99% of the data as missing and still recover the true CP-components. Why do you think this is possible?
(Hint what are the number of observations to free variables in the CP model)

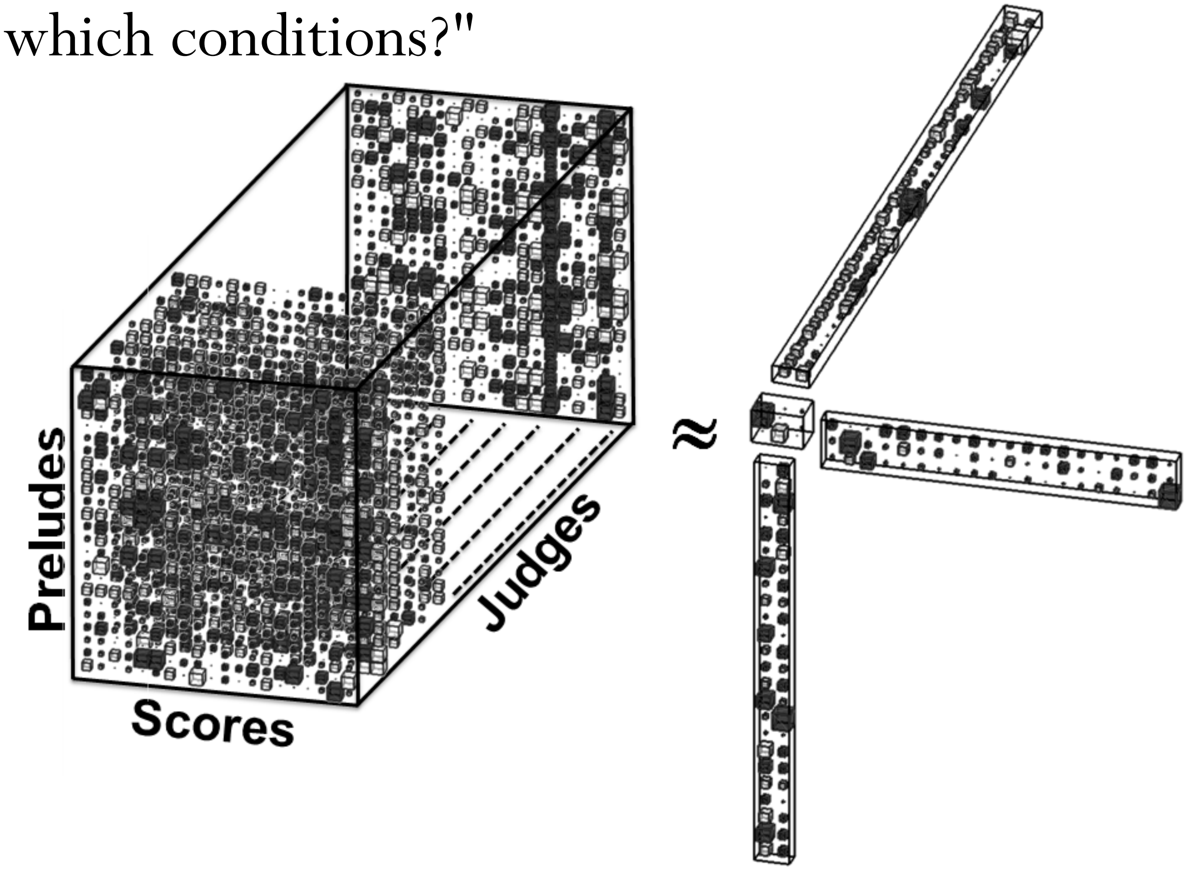
Software



Package	Language	Description
Tensor Toolbox [10, 12]	MATLAB	Collection of MATLAB classes for working with dense, sparse, and structured tensors. Extensive support for arithmetic operations, including multiplication and matricization. Includes HOSVD, Tucker, and PARAFAC algorithms for dense, sparse, or structured tensors.
N-way Toolbox [5]	MATLAB	Extensive collection of multi-way algorithms on dense arrays, including capabilities to handle missing data. Interface to functions uses MATLAB's built-in MDA data type.
CuBatch [40]	MATLAB	Graphical environment for data analysis. Multi-way data analysis algorithms are based on the N-way Toolbox.
PLS_Toolbox [116]	MATLAB	Commercial code for multi-way data analysis. Extensive collection of algorithms for dense arrays with many options, including missing data and model constraints. License required.
Multilinear Engine [84]	FORTRAN	Modeling environment for computing tensor decompositions. Modeling language allows for missing data, various constraints, and user-defined decompositions. License required.
HTL [117]	C++	Template library for tensors. Includes support for sparse tensors and some arithmetic operations.
FTensor [74]	C++	Library of optimized, template-based classes. Supports many arithmetic operations.
Boost.Multiarray [38]	C++	Library of abstract interfaces for dense N-dimensional arrays. Offers ability to resize, reshape, and create views.

Application in Psychology

"which group of subjects behave differently on which variables under which conditions?"

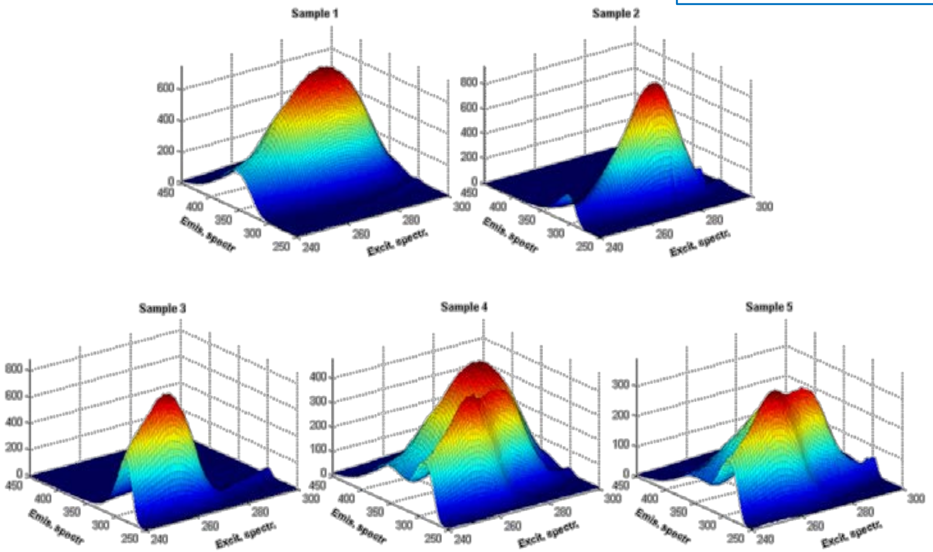


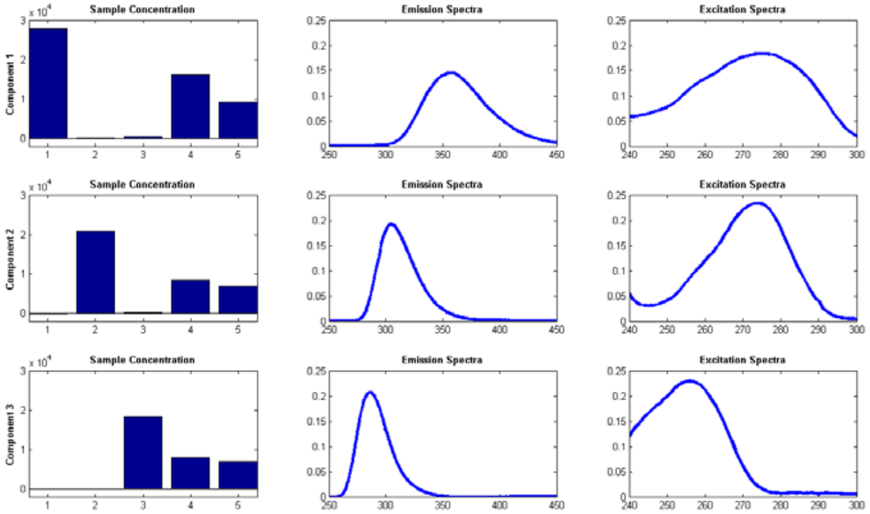
$$\mathcal{X}^{Predude \times Score \times Subject} \approx \sum_{lmn} g_{lmn} \boldsymbol{a}_l^{Prelude} \circ \boldsymbol{b}_m^{Score} \circ \boldsymbol{c}_n^{Subject}$$

Application in Chemistry

“Second order advantage, making it possible to do quantitative chemical analytes even in the presence of un-calibrated interferents"

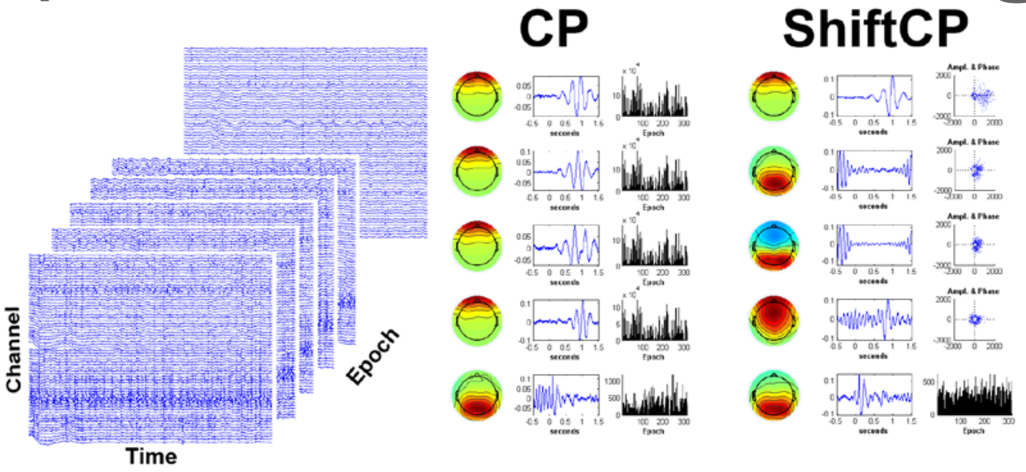
Beer-Lambert’s law stating a linear relation between absorbance of light and the concentration of a compound

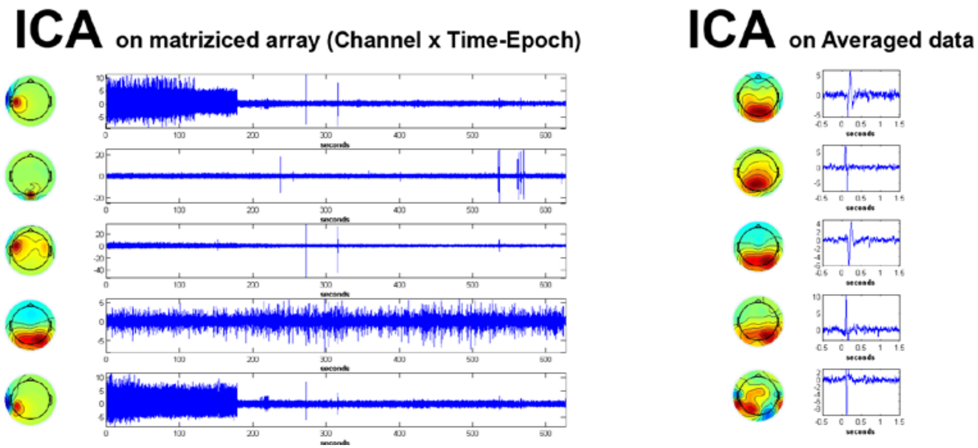




$$\mathcal{X}^{Exication \times Emission \times Samples} \approx \sum_{d=1}^{D_{components}} \boldsymbol{a}_d^{Excitation} \circ \boldsymbol{b}_d^{Emmision} \circ \boldsymbol{c}_d^{Samples}$$

Application in Neuroimaging



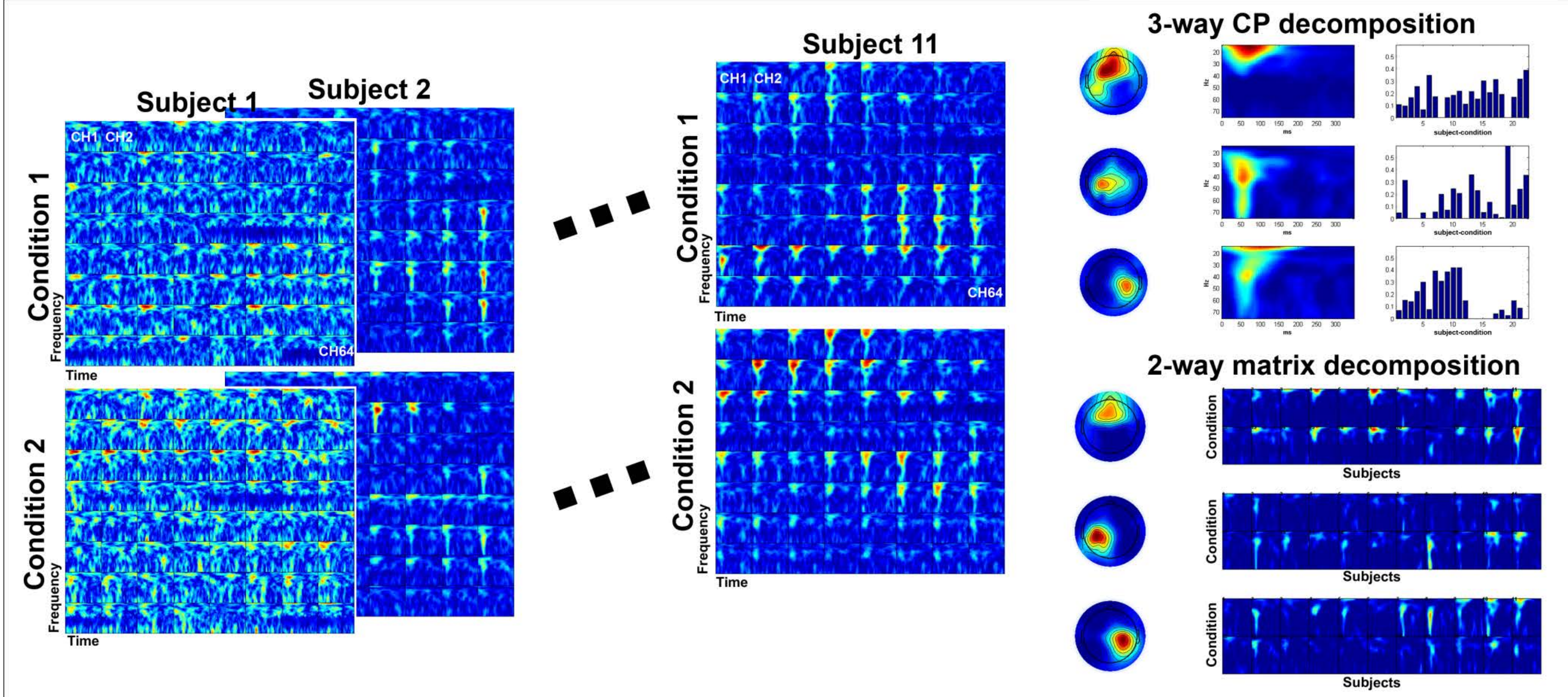


$$\chi^{Channel\times Time\times Epoch} \approx \sum_{d=1}^{D} \boldsymbol{a}_d^{Channel} \circ \boldsymbol{b}_d^{Time} \circ \boldsymbol{c}_d^{Epoch}$$

What are the difference in assumptions between ICA on matriziced array of Channel x Time-Epoch and CP on Channel x Time x Epoch?

Neuroimaging cont.

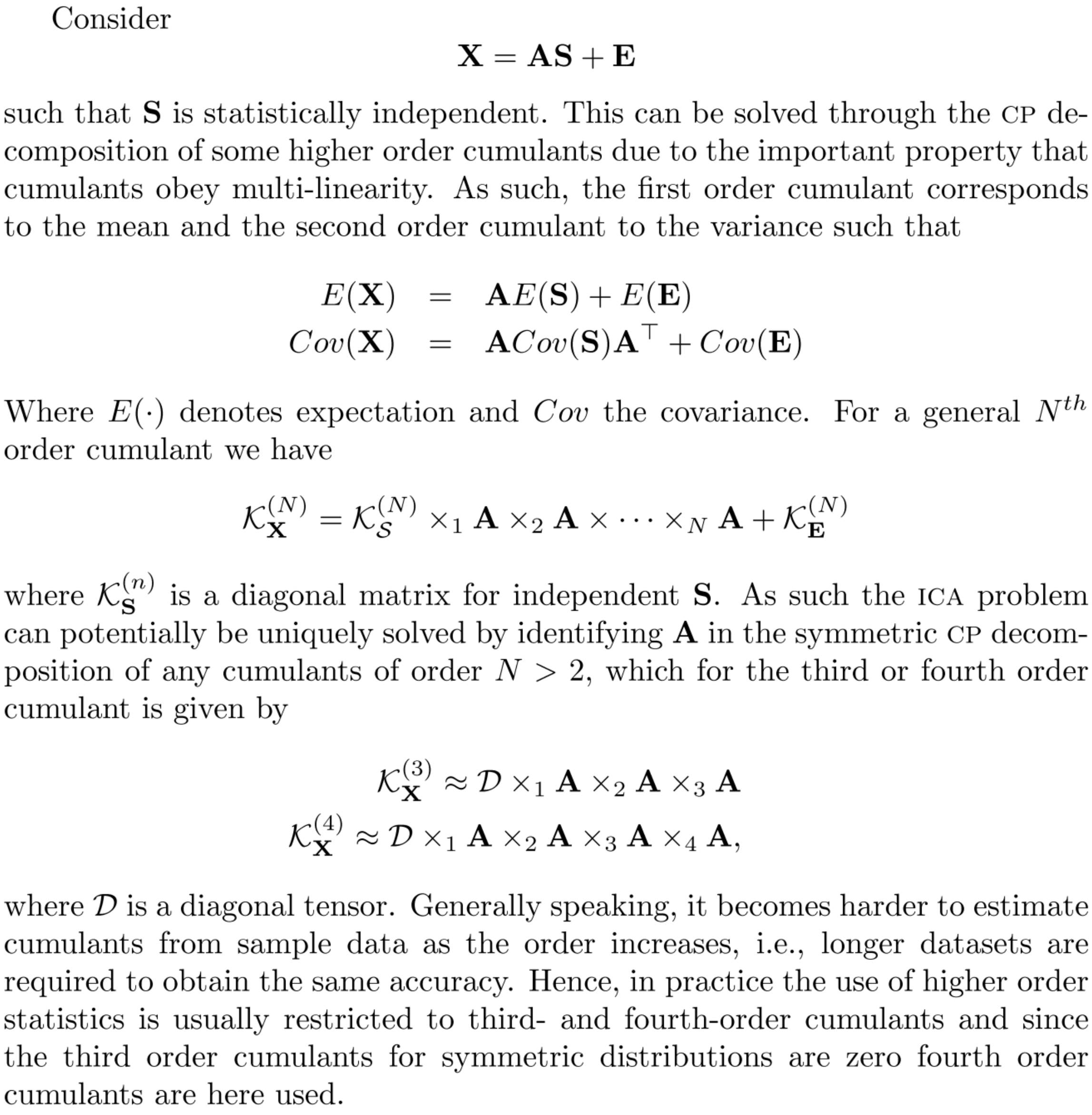
$$\mathcal{X}^{channel\times frequency-time\times subject-conditions}$$



$$\mathcal{X}^{channel\times frequency-time-subject-conditions}$$

$$\mathcal{X}^{Channel\times Time-Frequency\times Subject-Condition} \approx \sum_{d=1}^{D} \boldsymbol{a}_d^{Channel}\circ \boldsymbol{b}_d^{Time-Frequency}\circ \boldsymbol{c}_d^{Subject-Condition}.$$

Signal Processing (ICA)



Consider

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{E}$$

such that $\mathbf{S}$ is statistically independent. This can be solved through the CP decomposition of some higher order cumulants due to the important property that cumulants obey multi-linearity. As such, the first order cumulant corresponds to the mean and the second order cumulant to the variance such that

$$\begin{aligned} E(\mathbf{X}) &= \mathbf{A}E(\mathbf{S}) + E(\mathbf{E}) \\ Cov(\mathbf{X}) &= \mathbf{A}Cov(\mathbf{S})\mathbf{A}^\top + Cov(\mathbf{E}) \end{aligned}$$

Where $E(\cdot)$ denotes expectation and Cov the covariance. For a general N^{th} order cumulant we have

$$\mathcal{K}_{\mathbf{X}}^{(N)} = \mathcal{K}_{\mathcal{S}}^{(N)} \times_1 \mathbf{A} \times_2 \mathbf{A} \times \cdots \times_N \mathbf{A} + \mathcal{K}_{\mathbf{E}}^{(N)}$$

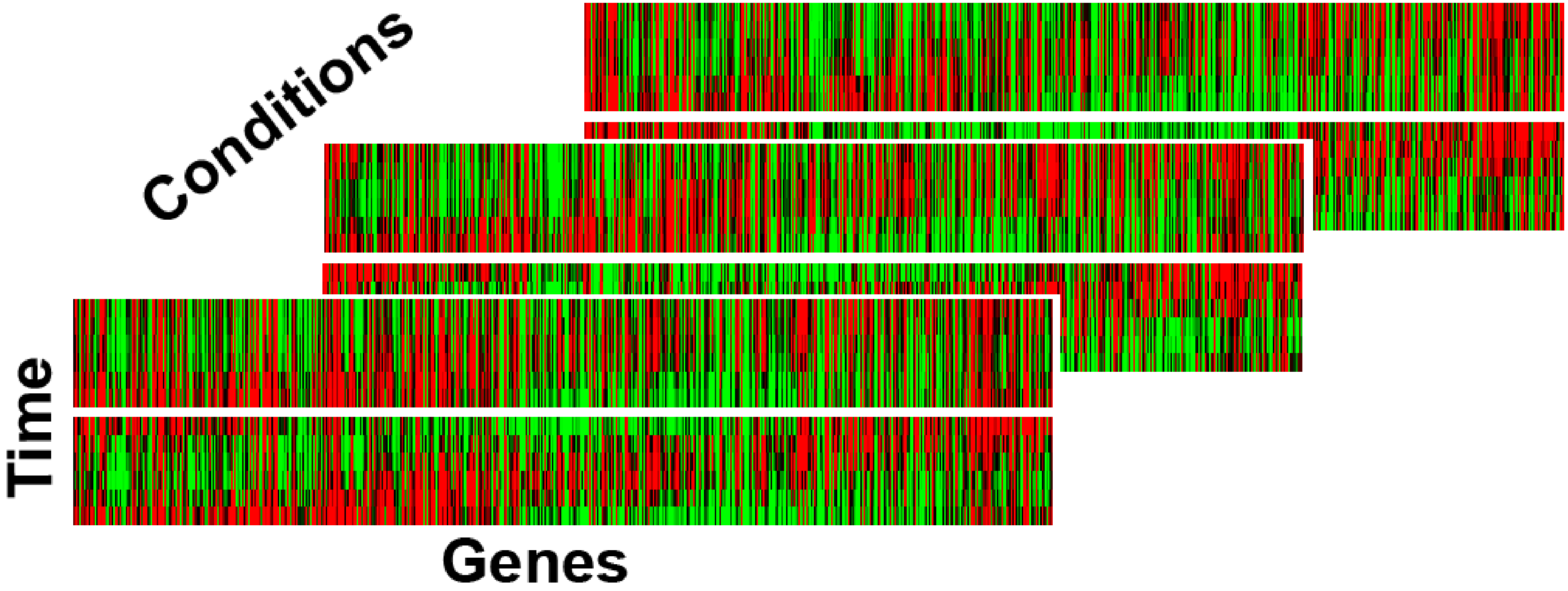
where $\mathcal{K}_{\mathbf{S}}^{(n)}$ is a diagonal matrix for independent $\mathbf{S}$. As such the ICA problem can potentially be uniquely solved by identifying $\mathbf{A}$ in the symmetric CP decomposition of any cumulants of order $N > 2$, which for the third or fourth order cumulant is given by

$$\begin{gathered} \mathcal{K}_{\mathbf{X}}^{(3)} \approx \mathcal{D} \times_1 \mathbf{A} \times_2 \mathbf{A} \times_3 \mathbf{A} \\ \mathcal{K}_{\mathbf{X}}^{(4)} \approx \mathcal{D} \times_1 \mathbf{A} \times_2 \mathbf{A} \times_3 \mathbf{A} \times_4 \mathbf{A}, \end{gathered}$$

where $\mathcal{D}$ is a diagonal tensor. Generally speaking, it becomes harder to estimate cumulants from sample data as the order increases, i.e., longer datasets are required to obtain the same accuracy. Hence, in practice the use of higher order statistics is usually restricted to third- and fourth-order cumulants and since the third order cumulants for symmetric distributions are zero fourth order cumulants are here used.

Application in Bio-informatics

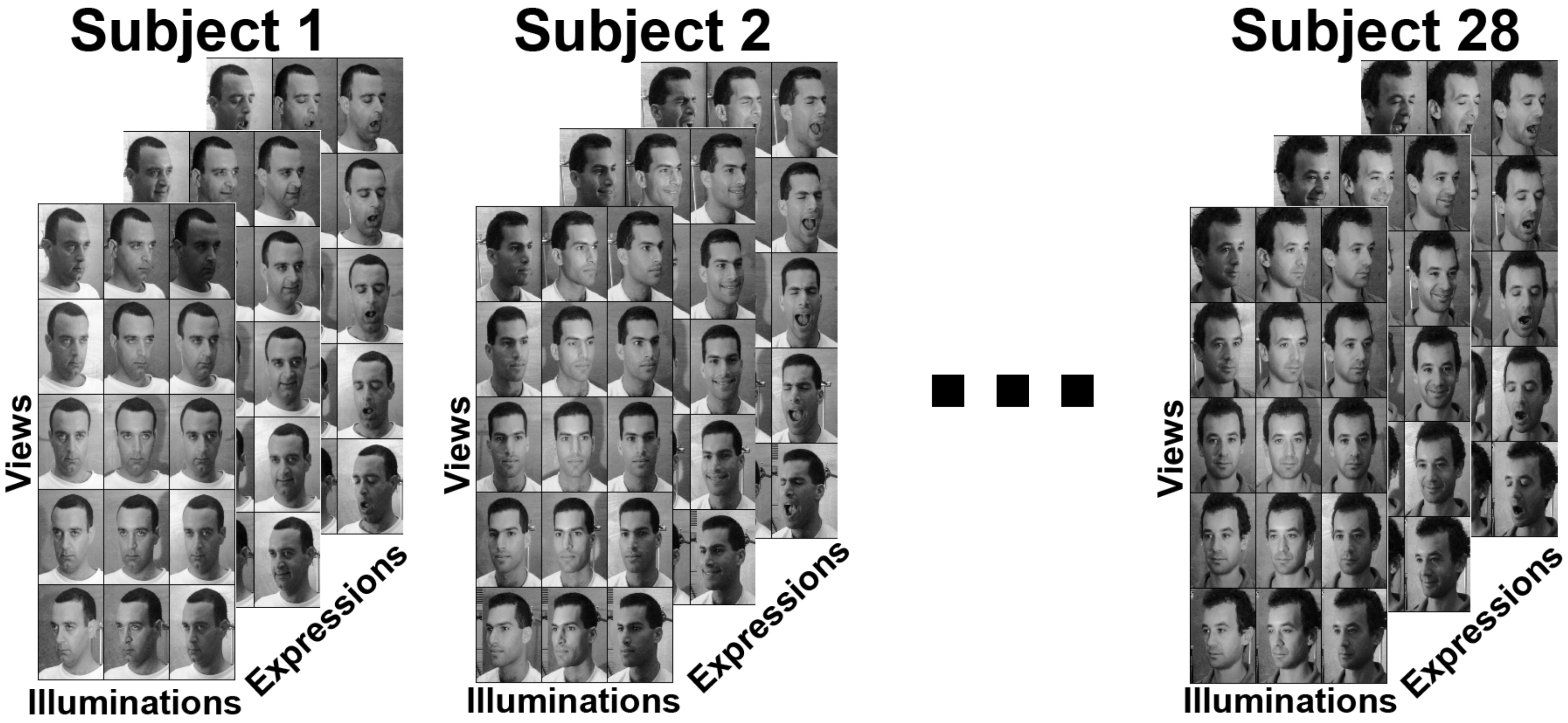
In micro-array analysis tensor decomposition useful in order to remove experimental artifacts and detect distinct stimulus dependent sets of functionally related genes



$$\mathcal{X}^{Gene \times Time \times Condition} \approx \sum_{lmn} g_{lmn} \boldsymbol{a}_l^{Gene} \circ \boldsymbol{b}_m^{Time} \circ \boldsymbol{c}_n^{Condition}$$

Application in Computer Vision

Extraction of patterns that generalize well across modes of variation ”TensorFaces” demonstrated to be significantly more accurate than PCA as multiple interrelated subspaces can collaborate to discriminate different classes



$$\mathcal{X}^{People \times Views \times Illum. \times Expres. \times Pixels} \approx \mathcal{G} \times_1 \mathbf{A}^{People} \times_2 \mathbf{B}^{Views} \times_3 \mathbf{C}^{Illum.} \times_4 \mathbf{D}^{Expres.} \times \mathbf{E}^{Pixels}$$

Application in Web-mining

Acar, Camtepe, Krishnamoorthy, Yener, Intelligence and Security Informatics 2005

Chatroom: Tucker, $\mathcal{X}^{users \times keywords \times time}$

Captured well underlying group structure

Kolda, Bader, Kenny, ICDM 2005

Hyperlink graphs: CP, $\mathcal{X}^{webpages \times webpages \times anchor\ text}$

Decomposition automatically identifed topics along with associated authoritative web-pages

Sun, Zeng, Liu, Lu, Chen,WWW 2005

Click-through data: TUCKER, $\mathcal{X}^{users \times queries \times webpages}$

”CubeSVD” significantly improve Web search performance over LSI and Collaborative Filtering approaches

Sun, Tao, Faloutsos, KDD 2006

Network traffic data: Tucker, $\mathcal{X}^{Source \times Destination \times Port}$

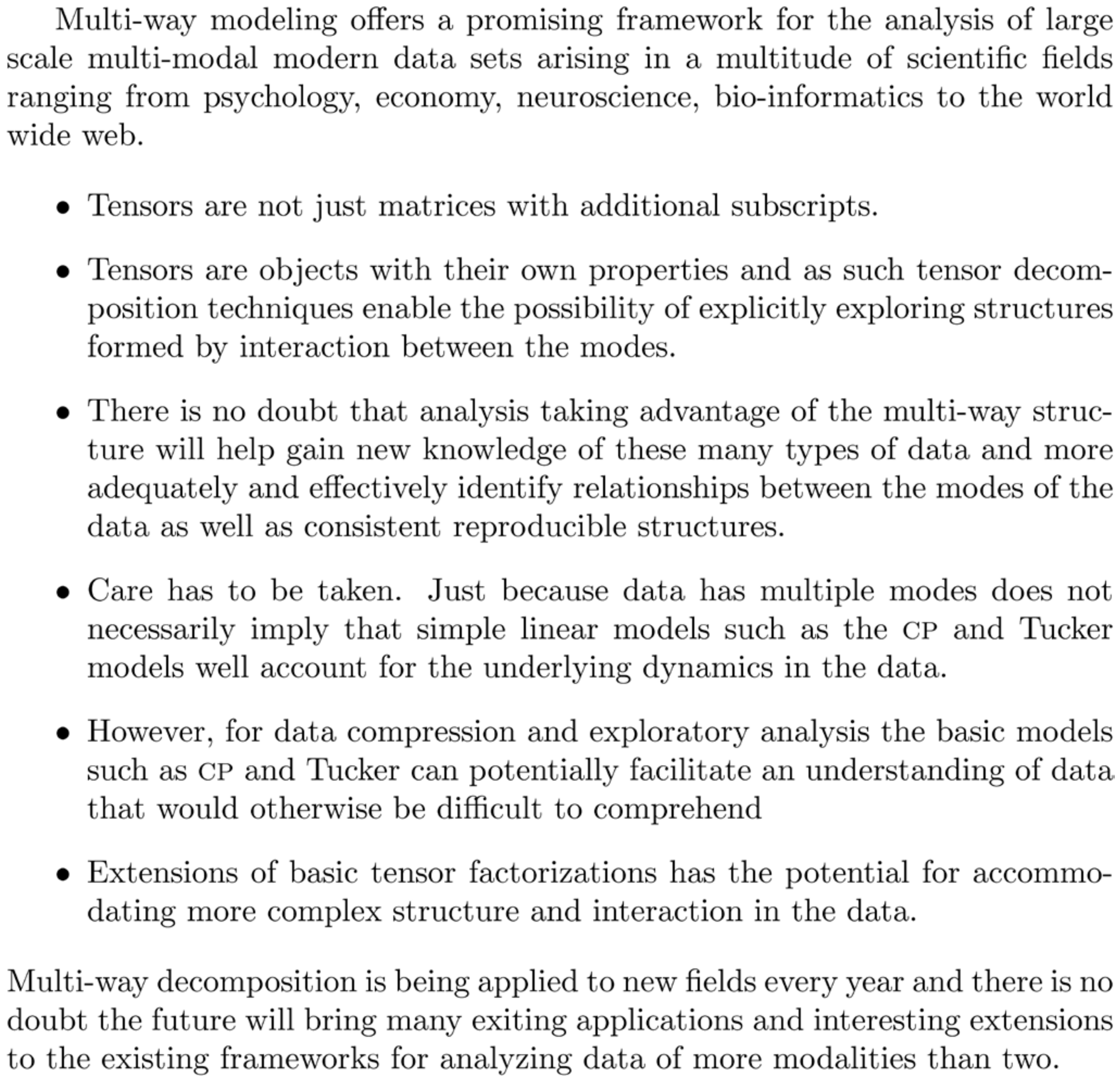
”Dynamic/Streaming Tensor Analysis” useful for anomaly detection and multi-way LSI

Bader, Harshman, Kolda, ICDM2007

Email communiciation: DEDICOM, $\mathcal{X}^{User \times User \times Month}$

Decomposition had strong correspondence with known job classifications while revealing the patterns of communication between these roles and the changes in communication pattern over time

Outlook



Multi-way modeling offers a promising framework for the analysis of large scale multi-modal modern data sets arising in a multitude of scientific fields ranging from psychology, economy, neuroscience, bio-informatics to the world wide web.

- Tensors are not just matrices with additional subscripts.
- Tensors are objects with their own properties and as such tensor decomposition techniques enable the possibility of explicitly exploring structures formed by interaction between the modes.
- There is no doubt that analysis taking advantage of the multi-way structure will help gain new knowledge of these many types of data and more adequately and effectively identify relationships between the modes of the data as well as consistent reproducible structures.
- Care has to be taken. Just because data has multiple modes does not necessarily imply that simple linear models such as the CP and Tucker models well account for the underlying dynamics in the data.
- However, for data compression and exploratory analysis the basic models such as CP and Tucker can potentially facilitate an understanding of data that would otherwise be difficult to comprehend
- Extensions of basic tensor factorizations has the potential for accommodating more complex structure and interaction in the data.

Multi-way decomposition is being applied to new fields every year and there is no doubt the future will bring many exiting applications and interesting extensions to the existing frameworks for analyzing data of more modalities than two.

Acknowledgements

- Tammy Kolda, Brett Bader, Rasmus Bro, Evrim Acar, Lek-Heng Lim, Pierre Comon, Lieven de Lathauwer.

Todays exercise

- Implement CP, Tucker and Core Consistency Diagnostics
- Analyse fluorescence spectroscopy data
- (Extra !) Generalize models to arbitrary N-way arrays