Imperial College London

EPSRC
Pioneering research and skills

Common and Individual Feature Extraction using Tensor Decompositions: A Remedy for the Curse of Dimensionality?

Ilia Kisil*, Giuseppe G. Calvi*, Andrzej Cichocki†, Danilo P. Mandic*

*Electrical and Electronic Engineering Department, Imperial College London, SW7 2AZ, UK

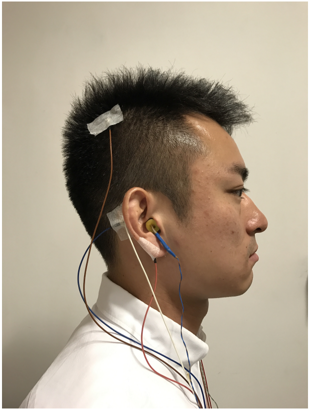
†Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, 351-0198, Japan

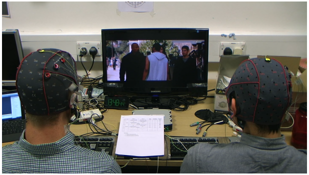
E-mails: {i.kisil15, giuseppe.calvi15, d.mandic}@imperial.ac.uk, a.cichocki@riken.jp

April 20, 2018

Outline

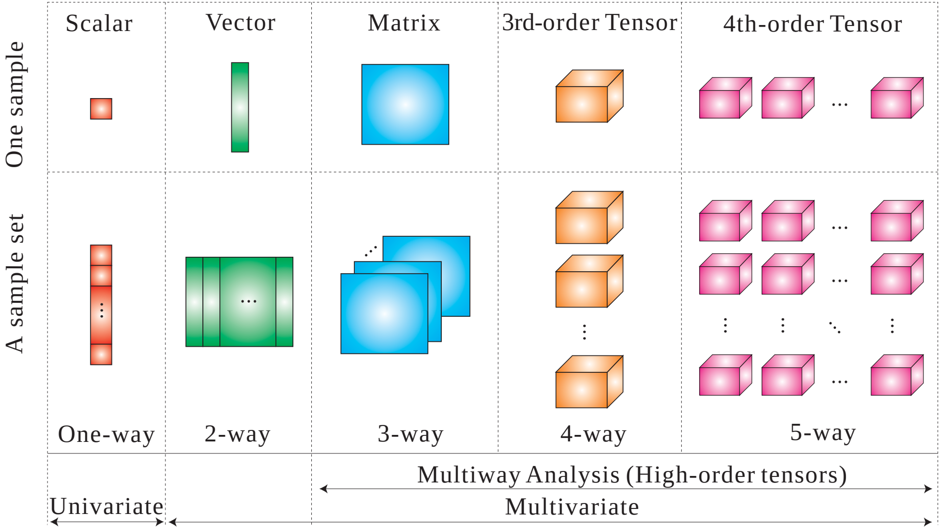
Naturally linked data





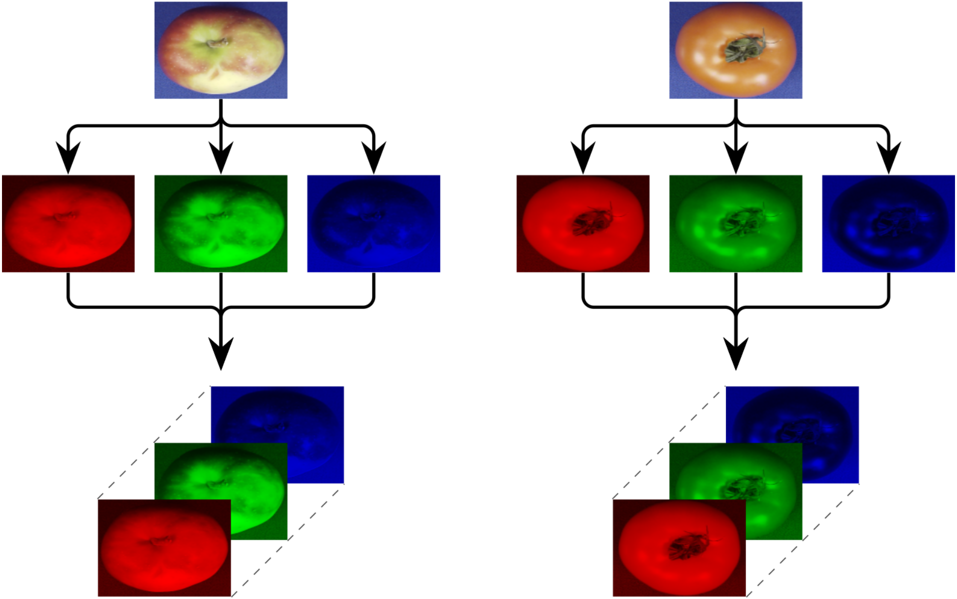
- Real-world data are often acquired as a collection of matrices $\looparrowright$ the same phenomenon is measured several times under various experimentation condition
- Such data blocks share some mutual components as well as individual information
- Common features reveal connections between members $\rightsquigarrow$ clustering
- Individual features characterise the members separately $\rightsquigarrow$ classification

Types of data: From a scalar to a tensor

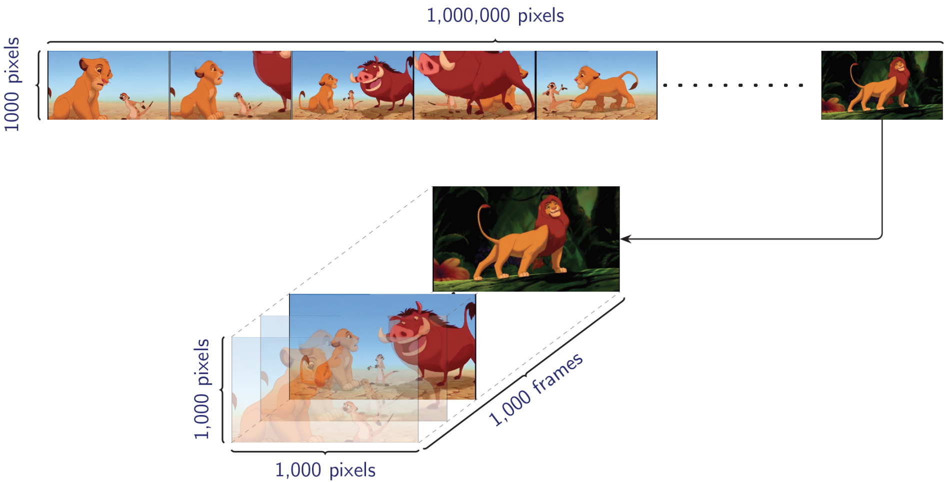


Source: "Tensor networks for dimensionality reduction and large-scale optimization. Part 1: Low-rank tensor decompositions"

Tensorisation: Image as base colors



Tensorisation: Video clip analysis



- A simple re-arrangement of frames (by stacking into a cube) transforms the matrix of 1,000 $\times$ 1,000,000 pixels into a 3-way tensor of size 1,000 $\times$ 1,000 $\times$ 1,000

Outer product: Efficient data representation

Consider the vectors $\mathbf{a} = \begin{bmatrix}1 & 1 & 1\end{bmatrix}^T, \mathbf{b} = \begin{bmatrix}1 & 2 & 3\end{bmatrix}^T, \mathbf{c} = \begin{bmatrix}1 & 10 & 100\end{bmatrix}^T$.

$$\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} = \textbf{?} \tag{1}$$

$$\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} = \begin{bmatrix}1\\1\\1\end{bmatrix} \circ \begin{bmatrix}1\\2\\3\end{bmatrix} \circ \begin{bmatrix}1\\10\\100\end{bmatrix} = \begin{bmatrix}1 & 2 & 3\\1 & 2 & 3\\1 & 2 & 3\end{bmatrix} \circ \begin{bmatrix}1\\10\\100\end{bmatrix}$$

Outer product: Efficient data representation

Consider the vectors $\mathbf{a} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T, \mathbf{b} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T, \mathbf{c} = \begin{bmatrix} 1 & 10 & 100 \end{bmatrix}^T.$

$$\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} = \mathbf{?} \tag{1}$$

$$\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 10 \\ 100 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 10 \\ 100 \end{bmatrix}$$

$$= \boxed{\begin{matrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{matrix}}$$

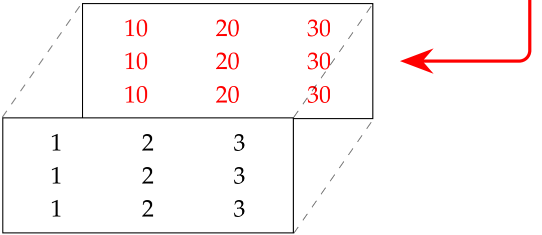
Outer product: Efficient data representation

Consider the vectors $\mathbf{a} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$, $\mathbf{b} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T$, $\mathbf{c} = \begin{bmatrix} 1 & 10 & 100 \end{bmatrix}^T$.

$$\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} = \mathbf{?} \tag{1}$$

$$\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 10 \\ 100 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 10 \\ 100 \end{bmatrix}$$

=



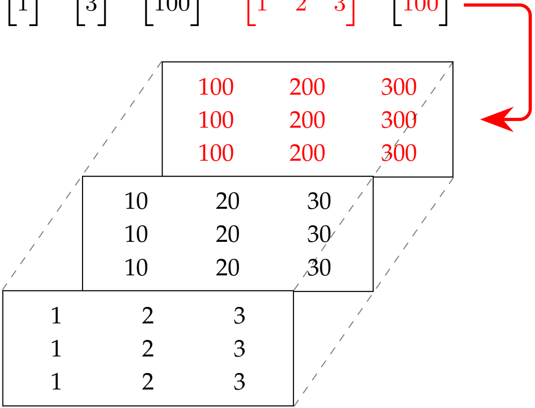
Outer product: Efficient data representation

Consider the vectors $\mathbf{a} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$, $\mathbf{b} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T$, $\mathbf{c} = \begin{bmatrix} 1 & 10 & 100 \end{bmatrix}^T$.

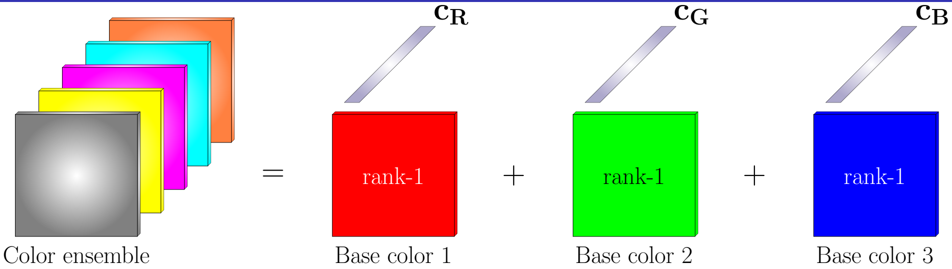
$$\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} = \textbf{?} \tag{1}$$

$$\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 10 \\ 100 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 10 \\ 100 \end{bmatrix}$$

$=$



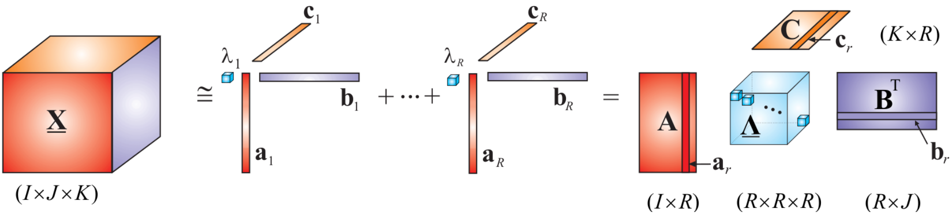
Outer product: Colorful example



- All colors are just combination of three base colors: red, green and blue
- We can represent this ensemble as a linear combination of outer products of base colors (red, green and blue) with the corresponding intensity vectors $\mathbf{c}_R, \mathbf{c}_G, \mathbf{c}_B$
- Their values characterise how much of the base color there is in the respective sample

$$\mathbf{c}_R = \begin{bmatrix} 0.5 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \quad \mathbf{c}_G = \begin{bmatrix} 0.5 \\ 1 \\ 0 \\ 1 \\ 0.5 \end{bmatrix} \quad \mathbf{c}_B = \begin{bmatrix} 0.5 \\ 0 \\ 1 \\ 1 \\ 0.125 \end{bmatrix} \tag{2}$$

The canonical polyadic decomposition (CPD)

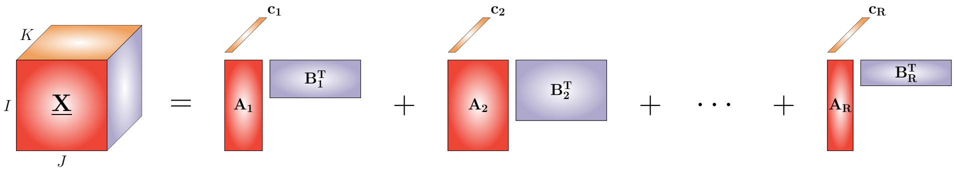


- Any tensor with arbitrarily many dimensions can be represented through the CPD

$$\underline{\mathbf{X}} \cong \sum_{r=1}^{R} \underline{\mathbf{X}}_r \cong \sum_{r=1}^{R} \lambda_r \cdot \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \tag{3}$$

- Mode-n vectors $\mathbf{a_r}, \mathbf{b_r}, \mathbf{c_r}$ are grouped into factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$
- Each factor matrix efficiently represents only one specific characteristic in accordance with corresponding mode of original data
- Real data are corrupted by noise $\looparrowright$ CPD is rarely exact and is estimated by solving

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\underline{\mathbf{X}} - \hat{\underline{\mathbf{X}}}\|_F^2 \qquad \text{with } \hat{\underline{\mathbf{X}}} = [\![\underline{\mathbf{\Lambda}}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \tag{4}$$



- LL1 is a linear combination of tensors with different multi-linear rank

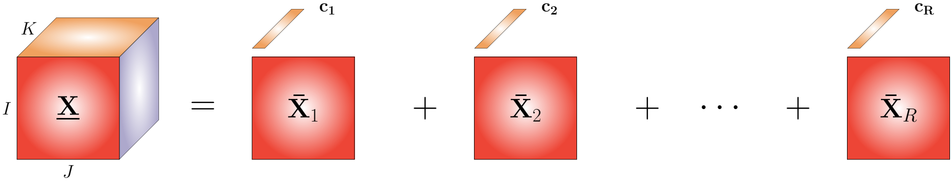
$$\underline{\mathbf{X}} \cong \sum_{r=1}^{R} \underline{\mathbf{X}}_r \cong \sum_{r=1}^{R} \mathbf{A}_r \circ \mathbf{B}_r \circ \mathbf{c}_r \tag{5}$$

- The outer product of matrices $\mathbf{A}_r \in \mathbb{R}^{I \times L_r}$ and $\mathbf{B}_r \in \mathbb{R}^{J \times L_r}$ is capable of representing of complex structure

$$\begin{gathered} \mathbf{X}_r = \mathbf{A}_r \circ \mathbf{B}_r = \mathbf{A}_r \mathbf{B}_r^T \\ \text{rank}(\mathbf{X}_r) > 1 \end{gathered} \tag{6}$$

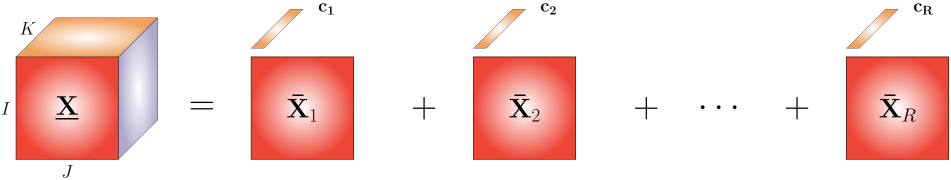
- More flexible representation of data, but computationally more expensive

Extraction of common features



- Interpretation of the factor matrices requires imposing constraints
- By introducing non-negativity constraint on $\mathbf{C}$ in Eq. (3) and on $\mathbf{c}_r$ in Eq. (5) the base matrices are considered to be common information $\bar{\mathbf{X}}_r$
- Common components are computed as:
  1. For the CPD
$$\bar{\mathbf{X}}_r = \mathbf{a}_r \circ \mathbf{b}_r = \mathbf{a}\mathbf{b}^T \tag{7}$$
  2. For the LL1
$$\bar{\mathbf{X}}_r = \mathbf{A}_r \circ \mathbf{B}_r = \mathbf{A}_r\mathbf{B}_r^T \tag{8}$$

Extraction of individual features



- Interpretation of the factor matrices requires imposing constraints

- By introducing non-negativity constraint on $\mathbf{C}$ in Eq. (3) and on $\mathbf{c}_r$ in Eq. (5) the base matrices are considered to be common information $\bar{\mathbf{X}}_r$

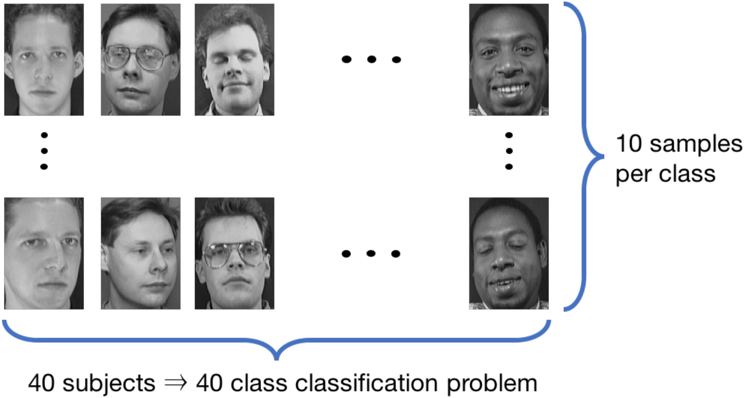
- For a sample $\mathbf{X}_k$, its common $\bar{\mathbf{X}}_k$ and individual $\check{\mathbf{X}}_k$ components are separable

$$\mathbf{X}_k = \bar{\mathbf{X}}_k + \check{\mathbf{X}}_k \quad \text{where } \mathbf{X}_k = \underline{\mathbf{X}}_{(:,:,k)} \tag{9}$$

- Values $\mathbf{C}_{(k,r)}$ indicate whether $\bar{\mathbf{X}}_r$ contributes to the k-th slice of $\underline{\mathbf{X}}$

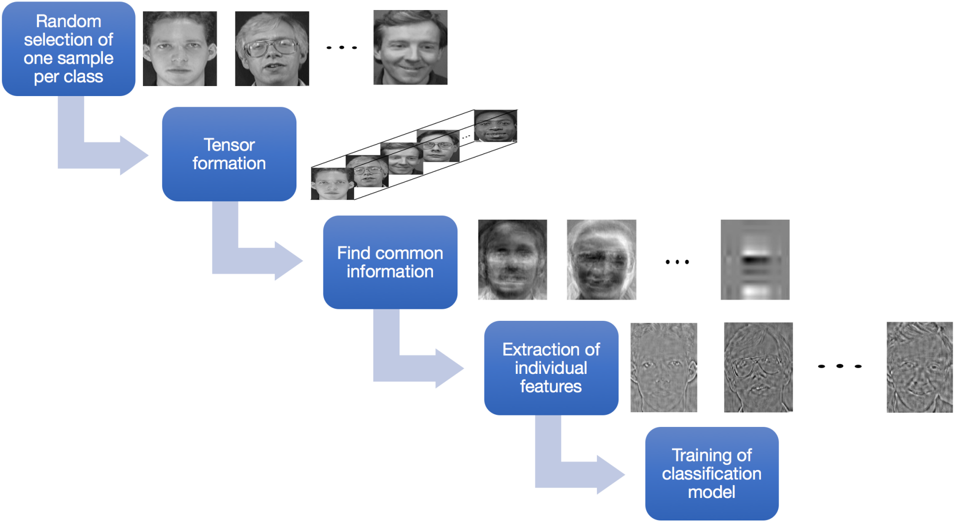
$$\begin{aligned} \check{\mathbf{X}}_k &= \mathbf{X}_k - \bar{\mathbf{X}}_k \\ &= \underline{\mathbf{X}}_{(:,:,k)} - \sum_{i \in I_k} \alpha_i \underline{\mathbf{Y}}_{(:,:,i)} \end{aligned} \tag{10}$$

ORL faces dataset

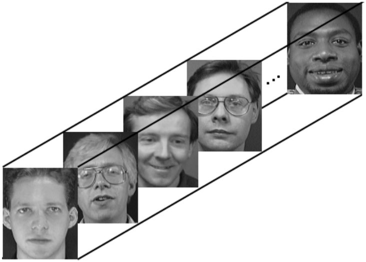


- We employed the benchmark **ORL faces dataset** for the classification of face images
- **400** samples = **40** (subjects) × **10** (different lighting conditions and facial expressions)
- Train test split for each class is **70%** and **30%** of samples respectively
- All samples from this dataset share **a lot of common information**

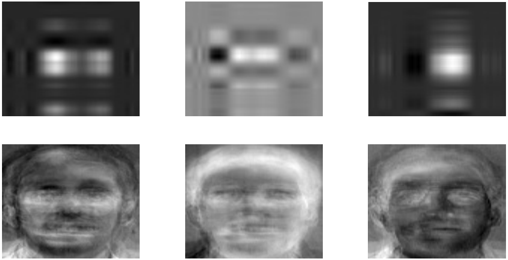
Pipeline for training a classification model

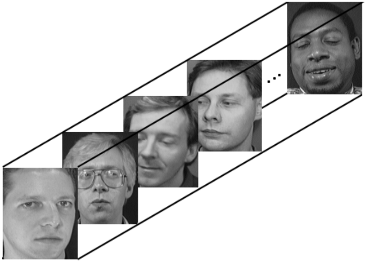


Results: Common information

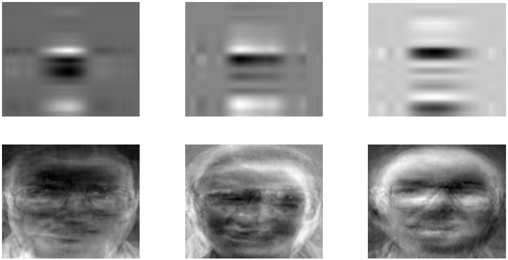


Group 1. Top – CPD, bottom - LL1





Group 2. Top – CPD, bottom - LL1



Results: Individual information

Subject 1, Group 1
LL1 approx error = 0.09

Subject 2, Group 1
LL1 approx error = 0.09

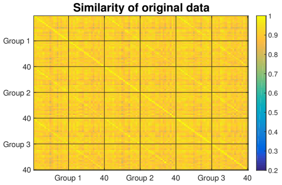
Subject 3, Group 2
LL1 approx error = 0.09

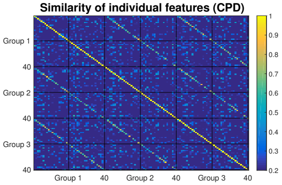
Subject 1, Group 2
LL1 approx error = 0.09

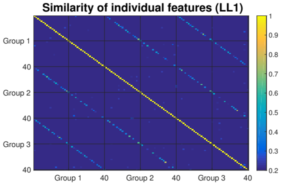
Subject 2, Group 2
LL1 approx error = 0.09

Subject 3, Group 2
LL1 approx error = 0.09

Results: Classification rates and analysis







- Similarity is estimated through the cosine distance
- Degree of similarity is mostly affected by the common information in data
- Individual components exhibit much less similar patterns across different classes
- This significantly reduces the searching space for decision boundaries
- Results were obtained by averaging rates of 100 independent simulations

Table 1: Classification Performance in %

	SVM	NN	QD	cKNN
Original	83.9	4.35	**91.5**	79.0
CPD	91.5	81.8	89.8	**85.5**
LL1	**94.7**	**92.2**	86.8	84.3

Conclusions: Key points to take home

1. The constraints imposed on different modes of a tensor decomposition should have physical meaning
2. The outer product plays a key role in separation of common and individual information
3. The dimensionality of search spaces can be dramatically reduced
4. There is a finite number of common features for a given data
5. The individual features can tackle overfitting of the classification model and enhance its performance

Conclusions: Key points to take home

1. The constraints imposed on different modes of a tensor decomposition should have physical meaning
2. The outer product plays a key role in separation of common and individual information
3. The dimensionality of search spaces can be dramatically reduced
4. There is a finite number of common features for a given data
5. The individual features can tackle overfitting of the classification model and enhance its performance

New Software: Higher Order Tensors ToolBOX (HOTTBOX)

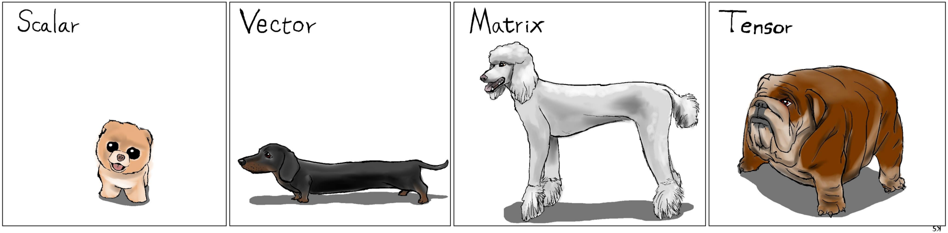
- Our python package for multilinear algebra: `github.com/hottbox/hottbox`
- Documentation: `hottbox.github.io`
- Tutorials: `github.com/hottbox/hottbox-tutorials`

Selected references

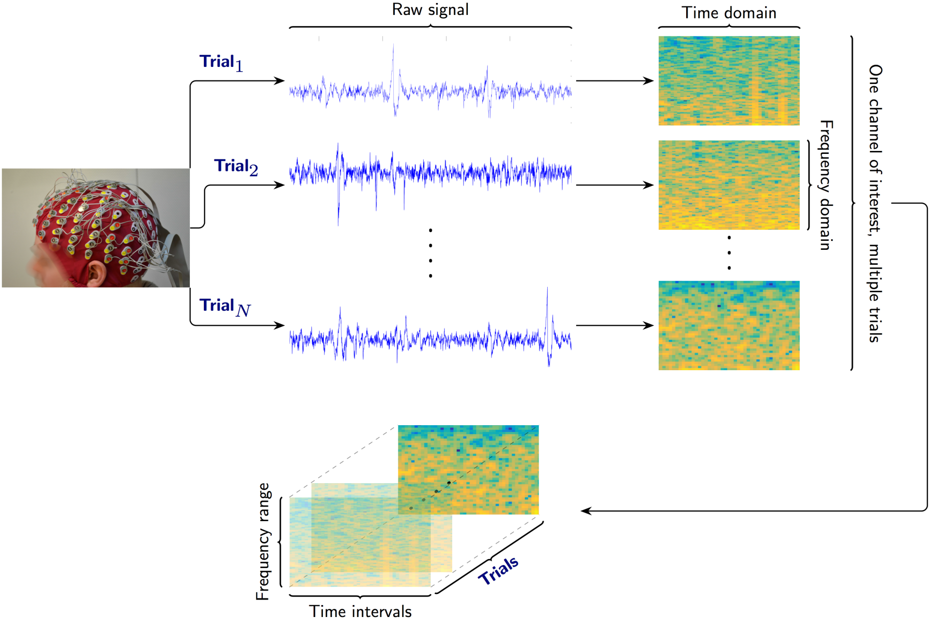
- A. Cichocki, D. P. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 145-163, 2015.
- A. Cichocki, N. Lee, I. Oseledets, A. H. Phan, Q. Zhao, and D. P. Mandic, "Tensor networks for dimensionality reduction and large-scale optimization. Part 1: Low-rank tensor decompositions," *Foundations and Trends® in Machine Learning*, vol. 9, no. 4-5, pp. 249-429, 2016.
- T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," SIAM Review, vol. 51, no. 3, pp. 455-500, 2009.
- L. Sorber, M. Van Barel, and L. De Lathauwer, "Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-$(L_r, L_r, 1)$ terms, and a new generalization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 695-720, 2013.
- G. Zhou, A. Cichocki, Y. Zhang, and D. P. Mandic, "Group component analysis for multiblock data: Common and individual feature extraction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2426-2439, 2016.

Thank you for your attention

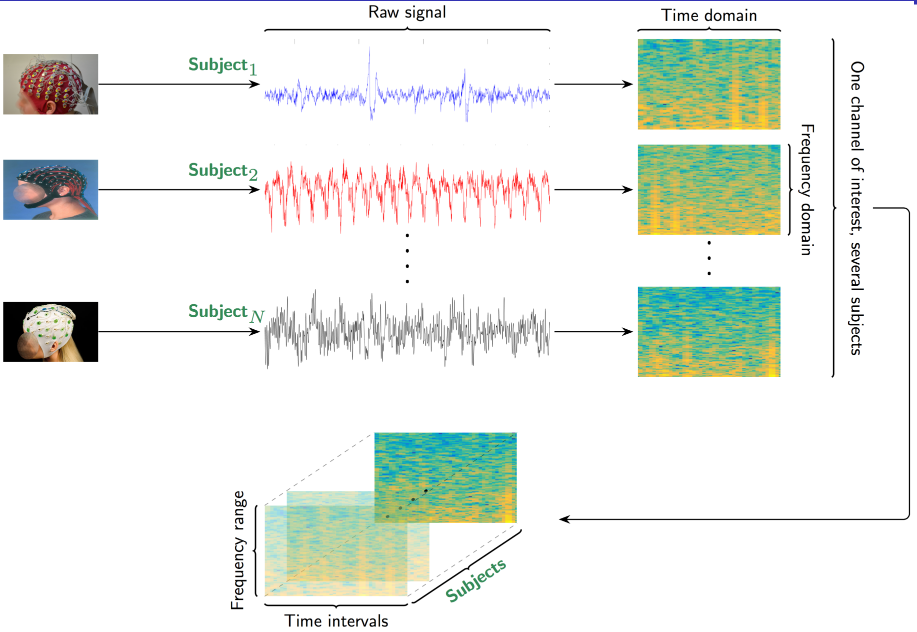
Questions?



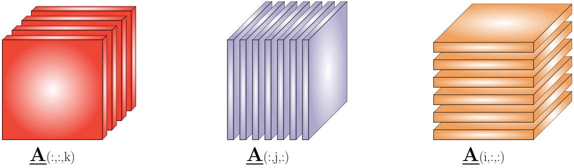
Appendix: Tensorisation for multiple trials

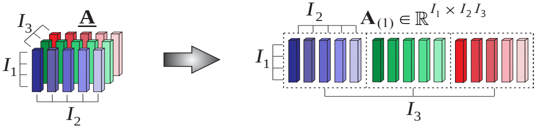


Appendix: Tensorisation for multiple subjects

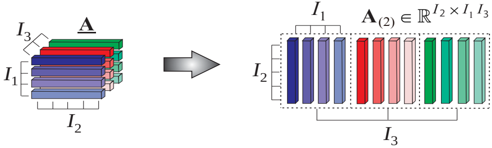


Appendix: Sub-structures within a tensor

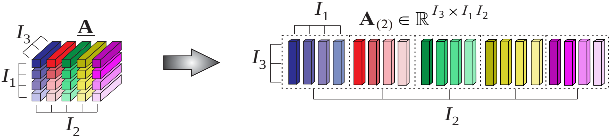




→ mode-1 unfolding

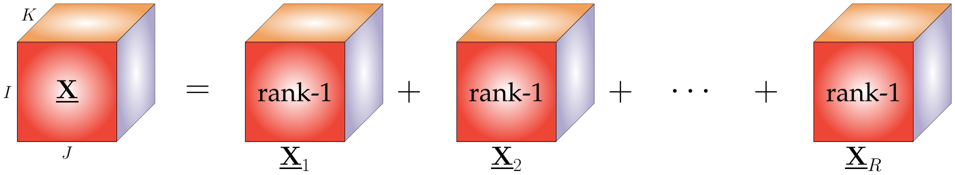


→ mode-2 unfolding

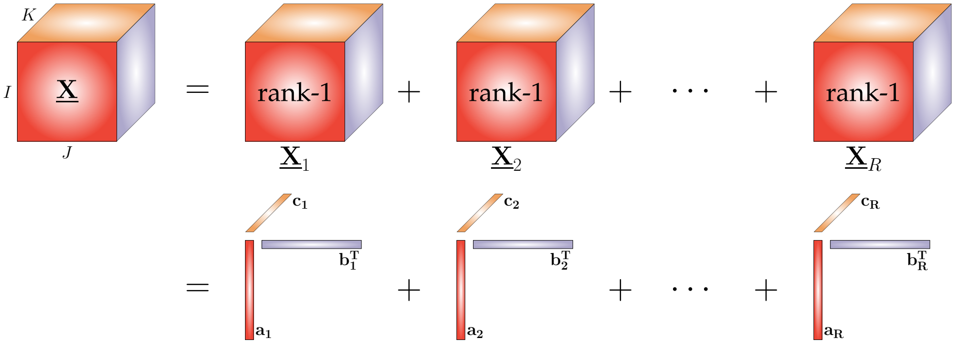


→ mode-3 unfolding

Appendix: CPD as a sum of common components



Appendix: CPD as a sum of common components



Appendix: CPD as a sum of common components