Fusion Network

CAIT
Skolkowo Institute of Science and Technology (SKOLTECH), Moscow, Russia

March 21, 2022

Part I

Tensor normal estimation

Tensor normal I

- Consider real-valued tensors, $\mathcal{X}_t$, $t = 1, \ldots T$, of size $I_1 \times I_2 \times \cdots \times I_N$, drawn from a separable Gaussian distribution with mean, $\boldsymbol{\mu}_n \in \mathbb{R}^{I_n}$, and covariance, $\boldsymbol{\Theta}_n \in \mathbf{S}_{I_n}$ for each mode-n for $n = 1, \ldots, N$.
- Distribution of vectorization $\boldsymbol{x}_t = \operatorname{vec}(\mathcal{X}_t)$ is known to be normal with
    - mean $\boldsymbol{\mu} = \bigotimes_{n=N}^{1} \boldsymbol{\mu}_n$
    - and covariance matrix $\boldsymbol{\Theta} = \bigotimes_{n=N}^{1} \boldsymbol{\Theta}_n$ (7), that is,

$$\boldsymbol{x}_t = \operatorname{vec}(\mathcal{X}_t) \sim \mathcal{N}\left(\bigotimes_{n=N}^{1} \boldsymbol{\mu}_n, \bigotimes_{n=N}^{1} \boldsymbol{\Theta}_n\right)$$

- Probability density function given by

$$p(\mathcal{X}_t) = \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{x}_t - \boldsymbol{\mu}\right)^T \left(\bigotimes_{n=N}^{1} \boldsymbol{\Theta}_n\right)^{-1} \left(\boldsymbol{x}_t - \boldsymbol{\mu}\right)\right)}{(2\pi)^{\frac{K}{2}} \det^{\frac{1}{2}}\left(\bigotimes_{n=N}^{1} \boldsymbol{\Theta}^{(n)}\right)} \qquad (1)$$

  where $K = I_1 I_2 \cdots I_N$ is the total number of elements of the tensor $\mathcal{X}_t$.

Problem: Estimation of the means μ_n and covariance matrices $\boldsymbol{\Theta}_n$

Maximizing the log-likelihood of T samples, $\mathcal{X}(t)$, (4; 7; 3)

$$\begin{aligned}\mathcal{L}(\{\boldsymbol{\mu}_n\}_{n=1}^N, \{\boldsymbol{\Theta}_n\}_{n=1}^N) &= \sum_{t=1}^{T} \ln p\left(\mathcal{X}(t) \middle| \{\boldsymbol{\mu}_n\}_{n=1}^N, \{\boldsymbol{\Theta}_n\}_{n=1}^N\right) \\ &= -\frac{TK}{2}\ln(2\pi) - \frac{T}{2}\ln\det\left(\bigotimes_{n=N}^{1} \boldsymbol{\Theta}^{(n)}\right) \\ &\quad - \frac{1}{2}\sum_{t=1}^{T} (\boldsymbol{x}_t - \boldsymbol{\mu})^T \left(\bigotimes_{n=N}^{1} \boldsymbol{\Theta}_n\right)^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}) \,. \qquad (2)\end{aligned}$$

Tensor normal IV

- The mean tensor or its vectorization, $\boldsymbol{\mu}$, is often estimated either as the sample mean or its best rank-1 or low-rank tensor approximation (4; 6; 3).

  For example, (10; 16) model the mean tensor, $\boldsymbol{\mu}$, in the Tucker format and estimate it from the sample mean.

$$\boldsymbol{\mu} = \sum_{r_1 \ldots r_N} g_{r_1 \ldots r_N} \boldsymbol{\mu}_{B,r_N} \otimes \cdots \otimes \boldsymbol{\mu}_{1,r_1} \tag{3}$$

  This estimation method works when the noise is small or homogeneous uncorrelated or when the number of samples, T, is sufficiently large.

Tensor normal V

How to estimate the parameters Update in an alternating fashion:

1. fix the mean $\boldsymbol{\mu}_n$ and optimize $\boldsymbol{\Theta}_n$
2. fix the covariance matrices $\boldsymbol{\Theta}_n$ and optimize the mean $\boldsymbol{\mu}_n$ using the best rank-1 tensor approximation.
3. Repeat until convergence is achieved. The first step can be performed through the Flip-Flop algorithm(4; 15).

Estimation of mean components.

- Denote Cholesky decomposition of $\boldsymbol{\Theta}_n^{-1} = \mathbf{F}_n^T \mathbf{F}_n$ and define $\boldsymbol{v}_n = \mathbf{F}_n \boldsymbol{\mu}_n$, for $n = 1, \ldots, N$
- and define by $\tilde{\boldsymbol{x}}_t$ vectorization of the samples, $\mathcal{X}_t$, transformed by the covariance matrices,

$$\tilde{\boldsymbol{x}}_t = \mathrm{vec}(\mathcal{X}_t \times_1 \mathbf{F}_1 \times_2 \mathbf{F}_2 \cdots \times_N \mathbf{F}_N)$$

The sample mean tensor, $\bar{\mathcal{X}} = \frac{1}{T} \sum_t \mathcal{X}_t$, and its transformed tensor by $\mathbf{F}_n$

$$\bar{\boldsymbol{x}}_\theta = \frac{1}{T} \sum_t \tilde{\boldsymbol{x}}_t = \mathrm{vec}\left(\bar{\mathcal{X}} \times_1 \mathbf{F}_1 \times_2 \mathbf{F}_2 \cdots \times_N \mathbf{F}_N\right) . \qquad (4)$$

Tensor normal VII

- While keeping the covariance matrices, $\mathbf{\Theta}_n$, fixed, maximizing the log-lihood function $\mathcal{L}()$ in (2) is equivalent to minimizing the last term in (2)

$$
\begin{aligned}
\min \; f(\{\boldsymbol{\mu}_n\}) &= \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{x}_t - \boldsymbol{\mu})^T \left(\bigotimes_{n=N}^{1} \mathbf{F}_n^T \mathbf{F}_n\right)(\boldsymbol{x}_t - \boldsymbol{\mu}) \\
&= \frac{1}{2}\sum_{t=1}^{T} \|\tilde{\boldsymbol{x}}_t - \bigotimes_{n=N}^{1} \mathbf{F}_n \boldsymbol{\mu}_n\|_F^2 \\
&= \frac{T}{2}\|\bar{\boldsymbol{x}}_\theta - \bigotimes_{n=N}^{1} \boldsymbol{v}_n\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}\|\tilde{\boldsymbol{x}}_t\|_2^2 - \frac{T}{2}\|\bar{\boldsymbol{x}}_\theta\|_2^2 \qquad (5)
\end{aligned}
$$

- Seeking $\{\boldsymbol{\mu}_n\}$ in the minimization problem (5) can be simplified into finding the best rank-1 tensor, $\boldsymbol{v}_1 \circ \boldsymbol{v}_2 \circ \cdots \circ \boldsymbol{v}_N$, of the transformed sample mean tensor,
$\mathcal{X}_\theta = \bar{\mathcal{X}} \times_1 \mathbf{F}_1 \times_2 \mathbf{F}_2 \cdots \times_N \mathbf{F}_N$.

Estimation of Covariance matrices.

- Note that

$$\begin{aligned}
\ln(\det(\boldsymbol{\Theta}_2 \otimes \boldsymbol{\Theta}_1)) &= \ln(\det(\boldsymbol{\Theta}_2)^{I_1} \ \det(\boldsymbol{\Theta}_1)^{I_2}) \\
&= I_1 \ln(\det(\boldsymbol{\Theta}_2)) + I_2 \ln(\det(\boldsymbol{\Theta}_1))
\end{aligned} \quad (6)$$

$$\frac{\partial(\lg(\det(\mathbf{X})))}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T \quad (7)$$

- We can constraint $\mathrm{tr}(\boldsymbol{\Theta}_n) = 1$ or $\det(\boldsymbol{\Theta}_n) = 1$, and introduce an addition standard deviation σ

$$\ln(\det(\sigma\,(\boldsymbol{\Theta}_2 \otimes \boldsymbol{\Theta}_1))) = I_1 I_2 \ln(\sigma) + \ln(\det(\boldsymbol{\Theta}_2 \otimes \boldsymbol{\Theta}_1)) \quad (8)$$

Tensor normal IX

- Upon introducing the centred tensor variable, based on the rank-1 tensor mean model,

$$\mathcal{S}_t = \mathcal{X}_t - \left(\circ_{n=1}^{N} \boldsymbol{\mu}^{(n)}\right) \tag{9}$$

  the stationary point of $\mathcal{L}$ in (2) with respect to $\boldsymbol{\Theta}^{(n)}$ yields the ML estimator

$$\boldsymbol{\Theta}^{(n)} = \frac{I_n}{\sigma^2 TK} \mathbf{Q}^{(n)} \tag{10}$$

  where

$$\mathbf{Q}^{(n)} = \sum_{t=0}^{T-1} \mathbf{S}_{(n)}(t) \left(\bigotimes_{i \neq n} {\boldsymbol{\Theta}^{(i)}}^{-1} \right) \mathbf{S}_{(n)}^T(t) \tag{11}$$

  A variant of the flip-flop algorithm in (4; 7; 9; 10).

- σ

$$\sigma^2 = \frac{1}{KT} \sum_{t=0}^{T-1} \mathbf{s}^T(t) \left(\bigotimes_{n=N}^{1} {\boldsymbol{\Theta}^{(n)}}^{-1} \right) \mathbf{s}(t) \tag{12}$$

Tensor normal X

Higher rank for the mean tensor. e.g., $\boldsymbol{\mu} = \sum_{r=1}^{R} \bigotimes_{n=N}^{1} \boldsymbol{\mu}_r^{(n)}$

$$\boldsymbol{x}_t = \mathrm{vec}(\mathcal{X}_t) \sim \mathcal{N}\left(\boldsymbol{\mu}, \sigma \bigotimes_{n=N}^{1} \boldsymbol{\Theta}_n\right) \tag{13}$$

By applying the same transform in (4), the estimation of the mean tensor becomes decomposition of the transformed sample mean

$$\min \quad \frac{1}{2}\|\mathcal{X}_\theta - \mathcal{V}\|_2^2 \tag{14}$$

where $\mathcal{V}$ can be in any low-rank tensor format.

Analysis of EEG motor imagery I

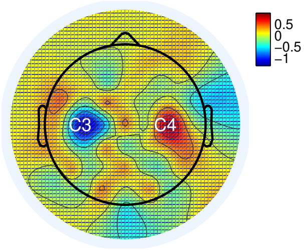
- We demonstrate an application in analysis of EEG data involving left/right motor imagery (MI) movements (for Brain-Like Computing and Intelligence). Instead of estimating the mean tensor as low-rank approximation of the sample mean, we consider the tensor model with in-homogeneous Gaussian noise.
- The EEG signals were recorded from 62 channels at a sampling frequency of 500 Hz for duration of 2 seconds with a 4 second break between the trials.
- The signals were preprocessed by a bandpass filter with cutoff frequencies of 8 Hz and 30 Hz, then transformed into the time-frequency domain using the complex Morlet wavelets with the bandwidth parameter $f_b = 1$ Hz, and the wavelet center frequency $f_c = 1$ Hz.

Analysis of EEG motor imagery II

- There are 200 trials for one subject, 100 per class, each represented as an order-3 tensor, $\mathcal{X}_t$, of size 62 *channels* $\times$ 23 *frequency bins* (8-30 Hz) $\times$ 50 *time frames*. See (14; 13) for detailed processing steps for the EEG signals.

Single trial Recognition

- Event-Related Desynchronization (ERD) (11): mu and beta rhythms over the contralateral primary sensorimotor
- Event-Related Synchronization (ERS)
- By convention, an ERD corresponds to a power decrease and an ERS to a power increase.



Right hand imagery movement

An ERD distributes over the left hemisphere and an ERS over the right hemisphere

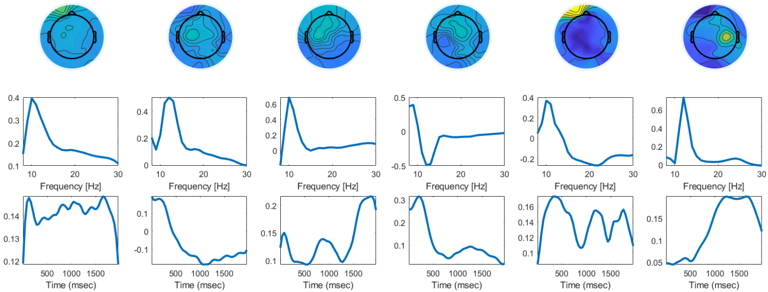
Left hand imagery movement

ERS/ERD phenomena occur on the left and right hemisphere, respectively.

In preparation and imagination of movement the mu and beta rhythms are desynchronized over the contralateral primary sensorimotor area, i.e., Event-Related Desynchronization (ERD), and an ipsilateral Event-Related Synchronization (ERS) or a contralateral beta ERS following the beta ERD (11; 12).

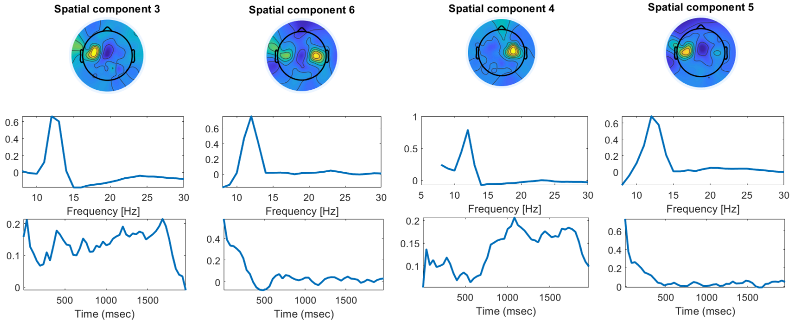
Analysis of EEG motor imagery I

A mean tensor of rank-6 and three covariance matrices were estimated from $T = 100$ tensors, $\mathcal{X}_t$, in the same imagery movement group.



(a) CPD of the sample mean tensor of the right hand MI group

Analysis of EEG motor imagery II



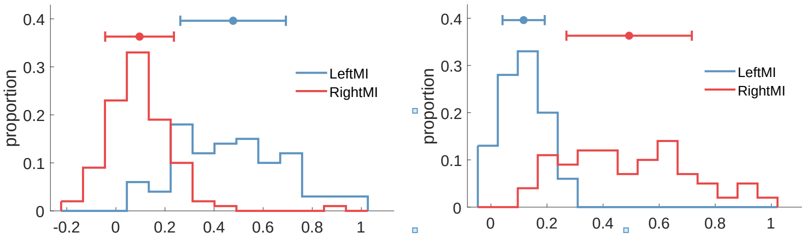
(b) Relevant components for EEG trials in the left hand MI group

(c) Relevant components for EEG trials in the right hand MI group

Figure: (a) Visualization of components in a rank-6 CPD of the sample mean tensor from 100 right hand MI trials. Components of one rank-1 tensor are shown in the same column. (b)-(c) Relevant components of the mean tensor of rank-6 estimated for each group of MI.

For the right hand MI group, the two spatial components show the ERS/ERD cover strongly the motor cortex area indicated by bright yellow and dark blue regions. The 4th temporal component indicates the power at the mu band (8-13 Hz) increasing after 500ms.

Similarly, for the left hand MI group, the EEG activities corresponding to ERD/ERS are manifested in the 6 and 5 components, respectively.



(a) Features learnt from the mean tensor for left hand MI group

(b) Features learnt from the mean tensor for right hand MI group

Figure: Comparison of features extracted for each group of MI trials using the mean tensor learnt from the same or different MI group.

Analysis of EEG motor imagery V

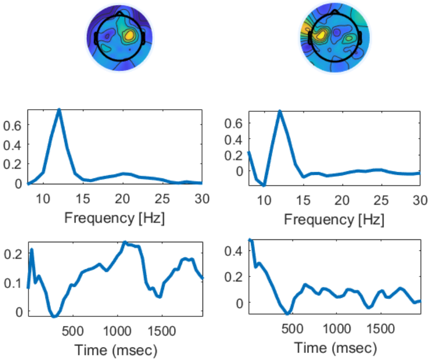


Figure: Relevant components of the mean tensor estimated from 15 samples in the right hand MI group.

Similar components were still able to be retrieved from only $T = 15$ samples, i.e. trials.

Analysis of EEG motor imagery VI

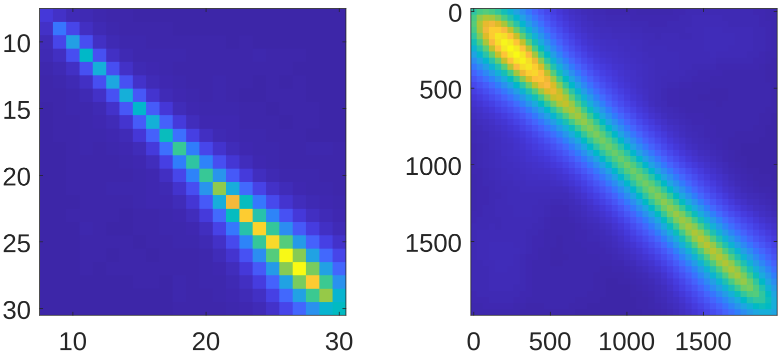


Figure: (left) Spectral covariance matrix and (right) temporal covariance matrix estimated for the left hand MI group.

Part II

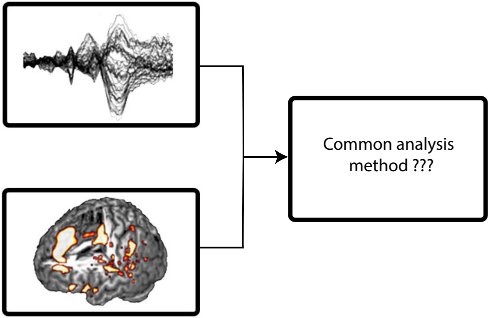
Data Fusion

Multi-Modal Analysis I

Examples of industrial applications taken from NEC - AIST AI Cooperative Research Laboratory.

- In interdisciplinary learning, data are often taken from different sensors, different devices.
  e.g., EEG signals, MEG, MRI, image sequence and audio, ...
  or due to different time-frequency representations, Fourier transform, wavelets
  or personal data including daily activities, ...
- **Data fusion approach.** The multi-modal data in general cannot be combined and processed by a common method.

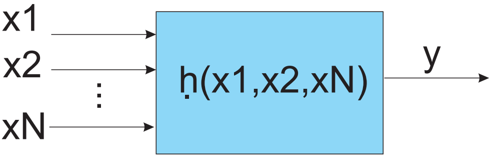
Multi-Modal Analysis II



Multi-Modal Analysis III

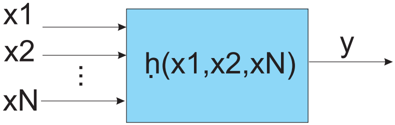
Our method

Identify a nonlinear system whose each input corresponds to each kind of dataset.
Tensor Network is used in the Multivariate Polynomial Regression to learn the nonlinear system ((1; 2)).

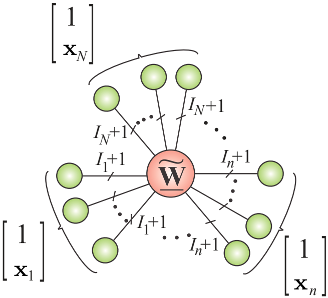


$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$: multi-modal data for each subject
y : output, e.g., label for each subject

Multi-Modal Analysis IV



(a) Nonlinear system identification



(b) Multivariate polynomial regression

Figure: (a) Identification of a nonlinear system from the observed system inputs and outputs can be implemented (b) using MPR with a weight tensor of N^2th-order.

Tensor Networks

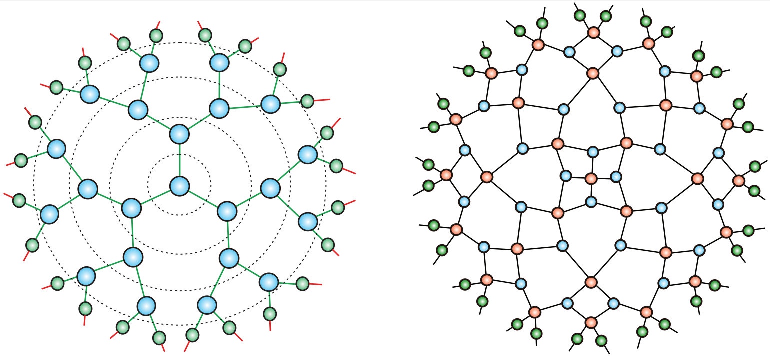


Figure: The Tree Tensor Network State with 3rd-order cores for the representation of 24th-order data tensors.

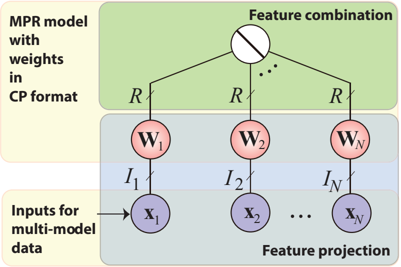
- *Tensor network* (TN) is constructed from small core tensors which are interconnected.
- Quasi-ranks can be determined in a stable way.

Shallow Network for Single-level Data Fusion

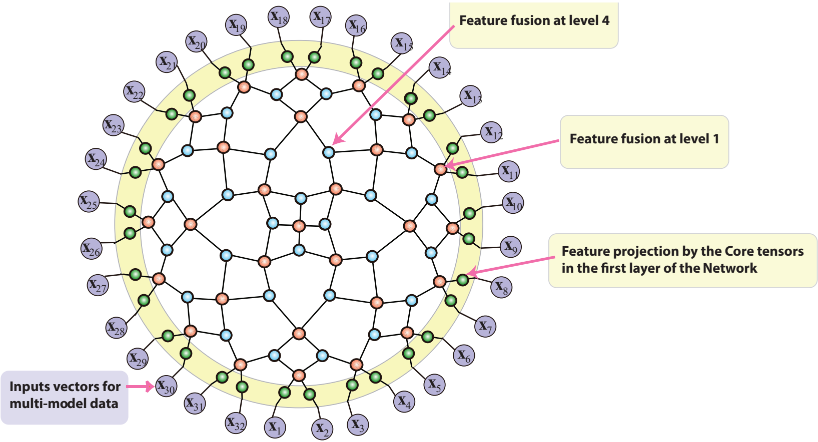
When the weight tensor $\mathcal{W}$ in a nonlinear system is represented as a shallow network, Output for a single sample is given by

$$y_k = (\boldsymbol{v}_{N,k} \odot \cdots \odot \boldsymbol{v}_{1,k})^T \operatorname{vec}(\mathcal{W})$$
$$= ((\boldsymbol{v}_{N,k}^T \mathbf{W}_N) \circledast \cdots \circledast (\boldsymbol{v}_{1,k}^T \mathbf{W}_1))\mathbf{1}$$

where $\mathbf{W}_n$ plays as a linear filter for each data type $\boldsymbol{v}_{n,k}$.



Shallow tensor network works as a single-layer feature combination method.



When $\mathcal{W}$ is given in a generalized TN form

- core tensors at the 1st layer play as projected filters
- core tensors at higher levels combine features.

We demonstrate an application of best rank-1 tensor approximation in multiview Tensor Canonical Correlation Analysis (TCCA).

- Given a dataset of K instances, each consists of N views, $\{\boldsymbol{x}_{k1}, \boldsymbol{x}_{k2}, \ldots, \boldsymbol{x}_{kN}\}$, where $k = 1, 2, \ldots, K$, $\boldsymbol{x}_{kn}$ of length I_n.
- The TCCA (8), an extension of CCA, seeks N canonical vectors, $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N$, which maximize the higher order canonical correlation

$$\begin{aligned} \max \quad & \mathcal{C} \,\bar{\times}_1 \boldsymbol{h}_1 \,\bar{\times}_2 \boldsymbol{h}_2 \cdots \bar{\times}_N \boldsymbol{h}_N \\ \text{s.t} \quad & \boldsymbol{h}_n^T \mathbf{C}_n \boldsymbol{h}_n = 1, \quad n = 1, \ldots, N. \end{aligned} \tag{15}$$

where $\mathbf{C}_n = \sum_k \boldsymbol{x}_{k,n} \boldsymbol{x}_{k,n}^T$ are correlation matrices for the view-n and

$$\mathcal{C} = \frac{1}{K} \sum_k \boldsymbol{x}_{k1} \circ \boldsymbol{x}_{k2} \circ \cdots \circ \boldsymbol{x}_{kN}$$

is an N-order tensor.

- By reparameterization $\boldsymbol{u}_n = \mathbf{C}_n^{1/2}\boldsymbol{h}_n$, the TCCA can be rewritten as best rank-1 tensor approximation of the tensor $\tilde{\mathcal{C}} = \mathcal{C} \times_1 \mathbf{C}_1^{-1/2} \times_2 \mathbf{C}_2^{-1/2} \cdots \times_N \mathbf{C}_N^{-1/2}$

$$\begin{array}{ll} \max & \tilde{\mathcal{C}}\, \bar{\times}_1 \boldsymbol{u}_1\, \bar{\times}_2 \boldsymbol{u}_2 \cdots \bar{\times}_N \boldsymbol{u}_N \\ \text{s.t.} & \boldsymbol{u}_n^T \boldsymbol{u}_n = 1, n = 1, 2, \ldots, N. \end{array} \tag{16}$$

- For illustration and comparison purpose, we consider the TCCA for imbalanced classification of 100 samples.
  - The first class has 33 samples for the dataset with three views, and 25 samples for the four views case. Other samples belong to the 2nd class.
  - Feature matrices of size 100×200, i.e., $K = 100$ and $I_n = 200$, for each view-n are drawn from the normal distribution with means $\mu_n = n$ and variance $\sigma_n^2 = 0.5n$, $n = 1, 2, 3, 4$.

- Best rank-1 tensor approximation problem are applied to decompose the covariance tensors. The performance measure is reported by 100×5-fold cross-validation and shown below

Table: Performance comparison of three algorithms, ALS, RORO and R1LM for the TCCA.

	3 views			4 views		
	ALS	**RORO**	**R1LM**	**ALS**	**RORO**	**R1LM**
Sensitivity	73.767	74.798	**74.893**	68.638	**69.865**	69.739
Specificity	28.738	27.605	**27.477**	33.438	**32.296**	32.491
F-measure	74.448	74.940	**74.980**	71.64	**72.267**	72.096
Relative Error	0.969	**0.957**	**0.957**	0.975	**0.955**	**0.955**

[1] Cichocki, A., Lee, N., Oseledets, I., Phan, A.-H., Zhao, Q., and Mandic, D. P. (2016). Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. *Foundations and Trends® in Machine Learning*, 9(4-5):249–429.

[2] Cichocki, A., Phan, A.-H., Zhao, Q., Lee, M., Oseledets, I., Sugiyama, M., and Mandic, D. P. (2017). Tensor networks for dimensionality reduction and large-scale optimization: Part 2 applications and future perspectives. *Foundations and Trends® in Machine Learning*, 9(6):431–673.

[3] Dees, B. S., Phan, A.-H., and Mandic, D. P. (2019). A statistically identifiable model for tensor-valued gaussian random variables.

[4] Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64:105–123.

[for Brain-Like Computing and Intelligence] for Brain-Like Computing, C. and Intelligence, M. Data set for single trial EEG classification in BCI. http://bcmi.sjtu.edu.cn/data1/.

[6] Gerards, D. and Hoff, P. D. (2015). Equivariant minimax dominators of the MLE in the array normal model. *Journal of Multivariate Analysis*, 137:32–49.

[7] Hoff, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196.

[8] Luo, Y., Tao, D., Ramamohanarao, K., Xu, C., and Wen, Y. (2015). Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3111–3124.

References III

[9] Manceur, A. M. and Dutilleul, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239:37–49.

[10] Nzabanita, J., von Rosen, D., and Singull, M. (2015). Maximum likelihood estimation in the tensor normal model with a structured mean. *Linköping University Electronic Press, LiTH-MAT-R-2015/08-SE*.

[11] Pfurtscheller, G. and da Silva, L. F. H. (1997). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin Neurophysiol*, 110:1842–1857.

[12] Pfurtscheller, G. and da Silva, L. F. H. (2005). EEG event-related desynchronization (ERD) and event-related synchronization (ERS). In Niedermeyer, E. and da Silva, F. L., editors, *Electroencephalography: Basic Principles , Clinical Applications, and Related Fields*, volume 5.

References IV

[13] Phan, A.-H. (2011). NFEA: Tensor toolbox for feature extraction and applications. http://www.bsp.brain.riken.jp/~phan/nfea.html.

[14] Phan, A.-H. and Cichocki, A. (2010). Tensor decompositions for feature extraction and classification of high dimensional datasets. *Nonlinear Theory and Its Applications, IEICE*, 1:37–68 (invited paper).

[15] Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494.

[16] Yue, X., Park, J. G., Liang, Z., and Shi, J. (2020). Tensor mixed effects model with application to nanomanufacturing inspection. *Technometrics*, 62(1):116–129.