Selection of Tensor Representations For Multiview Forecasting

Student: *Nadezhda Alsahanova*
Skoltech Advisor: *Maxim Panov*

2023

Skoltech

Area of research

Brain Computer Interfaces (BCI) help to restore communication and motor abilities. Data for BCI is acquired by electroencephalography (EEG) or electrocorticography (ECoG).

Problem:
Initial data acquired by EEG or ECoG systems is redundant and highly correlated. It leads to instability of models.

Source: GAO analysis (data). koya979/stock.adobe.com (images). | GAO-22-106118

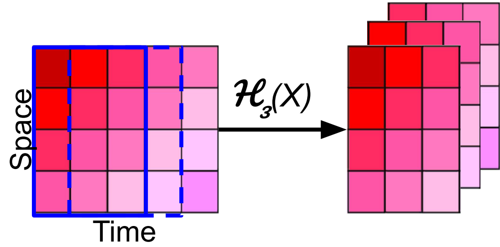
Skoltech

Tensorization

Tensorization can help to:

- find low-rank approximation with a high level of compression;
- reveal hidden correlations.

Hankelization is a natural data augmentation technique for time series to incorporate the intrinsic temporal correlation.
Hankelization connected with convolution:



$$x * h = \begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ x_2 & x_3 & \cdots & x_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{I-k+1} & x_{I-k+2} & \cdots & x_I \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_k \end{pmatrix} = \mathcal{H}_{I-k+1}(x)h$$

Hankelization has not been used in BCI area.
Data in BCI are spatially and temporally correlated. Local correlations can be revealed by hankelization.

Aim and objectives

The thesis goal is to find optimal representations of feature and target tensors in latent space by combining hankelization and dimensionality reduction methods. These tensor representations should be optimal in terms of the forecasting quality of the target variables and complexity of methods.

Objectives:

- determine whether hankelization along temporal mode improves quality of forecasting
- determine whether hankelization along spatial mode improves quality of forecasting
- determine whether hankelization along both modes improve quality of forecasting;

Skoltech

Task of multiview forecasting

Let $\mathbf{s}(t),\ \mathbf{y}(t)$ are time series, wher $\mathbf{y}(t)$ - target time series.

If there are several sources of initial time series $\mathbf{s}(t),\ \mathbf{y}(t)$, the dataset made from these time series presented as tensors:

$$\underline{\mathbf{X}} \in \mathbb{R}^{M \times I_1 \times \dots \times I_{D_x}}, \qquad \underline{\mathbf{Y}} \in \mathbb{R}^{M \times J_1 \times \dots \times J_{D_y}}$$

The task is to find an optimal model Φ for prediction $\underline{\mathbf{Y}}_m$ from an independent input object $\underline{\mathbf{X}}_m, m = 1, \dots, M$. The model is optimal, if it minimizes error functional $\mathcal{L}$:

$$\Phi^* = \arg\min_{\mathbf{\Theta}} \mathcal{L}\left(\Phi(\underline{\mathbf{X}}, \mathbf{\Theta}),\ \underline{\mathbf{Y}}\right)$$

where $\mathbf{\Theta}$ the parameters of model Φ .

Skoltech

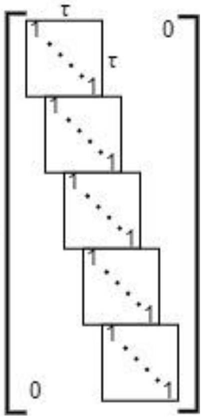
Hankelization

Hankelization is an effective way to transform lower-order data to higher-order tensors. It is due to
Hankelization of tensor $\underline{\mathbf{X}} \in \mathbb{R}^{T \times n_1 \times \ldots, n_D}$ can be done by multi-way delay embedding transform (MDT) with use of matrix $\mathbf{S}$:

$$\hat{\underline{\mathbf{X}}} = \mathcal{H}_\tau\left(\underline{\mathbf{X}}\right) = Fold_{(T,\tau)}\left(\underline{\mathbf{X}} \times_1 \mathbf{S}\right) \in \mathbb{R}^{(T-\tau+1) \times \tau \times n_1 \times \ldots \times n_D}$$

where $Fold_{(T,\tau)} : \mathbb{R}^{\tau(T-\tau+1) \times n_1 \times \ldots \times n_D} \rightarrow \mathbb{R}^{(T-\tau+1) \times \tau \times n_1 \times \ldots \times n_D}$.

S =



The inverse MDT:

$$\underline{\mathbf{X}} = \mathcal{H}_\tau^{-1}\left(\hat{\underline{\mathbf{X}}}\right) = Unfold_{(T,\tau)}\left(\hat{\underline{\mathbf{X}}}\right) \times_1 \mathbf{S}^\dagger$$

It was a hankelization along the temporal mode. Similarly, it can be done for any spatial modes.

Skoltech

Multilinear Principal Component Analysis

MPCA objective is to define a multilinear transformation that maps the original tensor space $\mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \ldots \otimes \mathbb{R}^{I_D}$ into a tensor subspace $\mathbb{R}^{L_1} \otimes \mathbb{R}^{L_2} \otimes \ldots \otimes \mathbb{R}^{L_D}$ with $L_d < I_d$

$$\hat{\underline{\mathbf{X}}}_m \approx \underline{\mathbf{X}}_m \times_1 \tilde{\mathbf{U}}^{(1)\top} \times_2 \tilde{\mathbf{U}}^{(2)\top} \ldots \times_D \tilde{\mathbf{U}}^{(D)\top}, \;\; m = 1, \ldots, M,$$

such $\hat{\underline{\mathbf{X}}}_m$ captures most of the variations observed in the original tensor objects. So, the D projection matrices $\tilde{\mathbf{U}}^{(d)}$ should maximize the total tensor scatter $\Upsilon_{\underline{\mathbf{X}}}$:

$$\left\{\tilde{\mathbf{U}}^{(d)} \in \mathbb{R}^{I_d \times L_d}, \; d = 1, \ldots, D\right\} = \arg \max_{\tilde{\mathbf{U}}^{(1)}, \ldots, \tilde{\mathbf{U}}^{(d)}} \Upsilon_{\underline{\mathbf{X}}}.$$

$$\Upsilon_{\underline{\mathbf{X}}} = \sum_{m=1}^{M} \|\underline{\mathbf{X}}_m - \overline{\underline{\mathbf{X}}}\|_F^2$$

Skoltech

High-order partial least squares

HOPLS performs simultaneously Tucker decompositions for an independent tensor $\underline{\mathbf{X}} \in \mathbb{R}^{M \times I_1 \times \ldots \times I_D}$ and a dependent tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{M \times J_1 \times \ldots \times J_D}$ which have the same size in the first mode. The HOPLS model:

$$\underline{\mathbf{X}} = \underline{\mathbf{G}}_x \times_1 \mathbf{T} \times_2 \overline{\mathbf{P}}^{(1)} \ldots \times_{D+1} \overline{\mathbf{P}}^{(D)} + \underline{\mathbf{E}}_R,$$

$$\underline{\mathbf{Y}} = \underline{\mathbf{G}}_y \times_1 \mathbf{T} \times_2 \overline{\mathbf{Q}}^{(1)} \ldots \times_{D+1} \overline{\mathbf{Q}}^{(D)} + \underline{\mathbf{F}}_R,$$

where $\underline{\mathbf{E}}_R, \underline{\mathbf{F}}_R$ are the residuals, $\overline{\mathbf{P}}^{(d)}, \overline{\mathbf{Q}}^{(d)}$ are the mode-d loading matrices, $\mathbf{T}$ is the latent matrix, and $\underline{\mathbf{G}}_x, \underline{\mathbf{G}}_y$ are the core tensors. The cross-covariance tensor is defined by

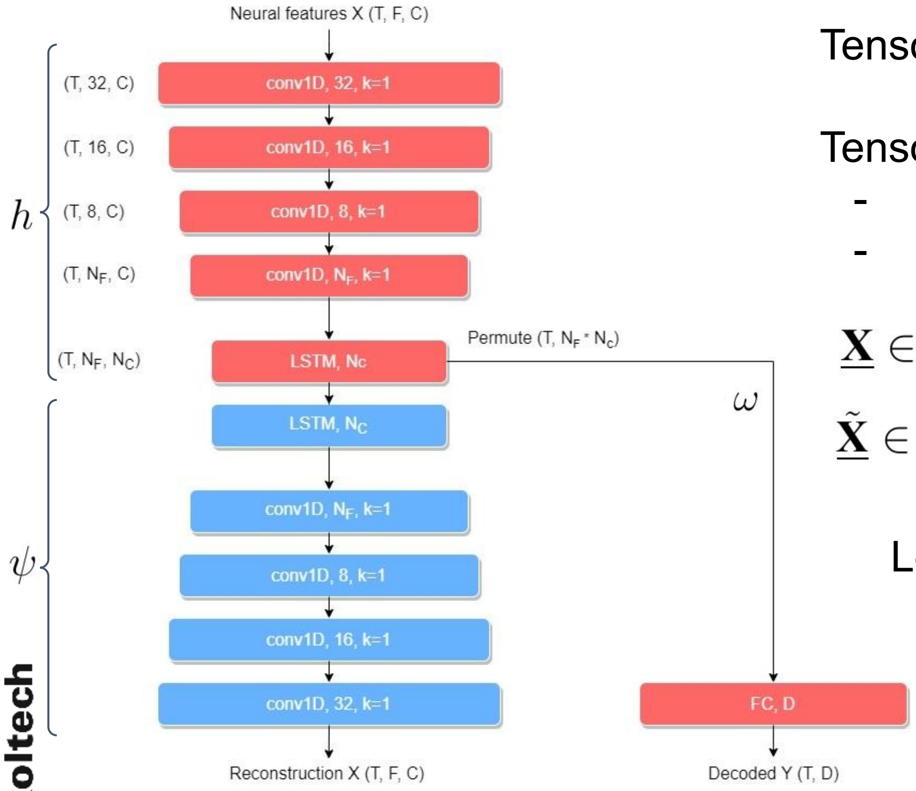
$$\underline{\mathbf{C}} = COV_{\{1,1\}}(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) \in \mathbb{R}^{I_1 \times \ldots \times I_D \times J_1 \times \ldots \times J_D}$$

The optimization problem can be formulated as

$$\left\| [\![\underline{\mathbf{C}}; \mathbf{P}^{(1)\top}, \ldots, \mathbf{P}^{(D)\top}, \mathbf{Q}^{(1)\top}, \ldots, \mathbf{Q}^{(D)\top}]\!] \right\|_F^2 \rightarrow \max_{\substack{\left\{ \mathbf{P}^{(d)},\ \mathbf{Q}^{(d)} \right\}, \\ \text{s.t.}\ \ \mathbf{P}^{(d)\top}\mathbf{P}^{(d)} = \mathbf{I}_{L_d}, \\ \mathbf{Q}^{(d)\top}\mathbf{Q}^{(d)} = \mathbf{I}_{K_d}}},$$

Skoltech

TensorReducedNet



Example of 3D autoencoder

TensorReducedNet is a SOTA model.

TensorReducedNet helps to
- keep tensor structure of initial data;
- align features with target data.

$$\underline{\mathbf{X}} \in \mathbb{R}^{M \times I_1 \times \ldots \times I_G} \xrightarrow{\text{conv1D blocks}} \tilde{\underline{\mathbf{X}}} \in \mathbb{R}^{M \times I_1 \times L_2 \times \ldots \times L_G}$$

$$\tilde{\underline{\mathbf{X}}} \in \mathbb{R}^{M \times I_1 \times L_2 \times \ldots \times L_G} \xrightarrow{\text{LSTM block}} \hat{\underline{\mathbf{X}}} \in \mathbb{R}^{M \times L_1 \times L_2 \times \ldots \times L_G}$$

Loss: $\mathcal{L}_1 = \mathcal{L}_{rec} + \alpha \cdot \mathcal{L}_{dec},$

$$\mathcal{L}_{rec} = \frac{1}{M} \sum_{m=1}^{M} \|\underline{\mathbf{X}}_m - h \circ \psi(\underline{\mathbf{X}}_m)\|^2$$

$$\mathcal{L}_{dec} = \frac{1}{M} \sum_{m=1}^{M} \|\mathbf{Y}_m - \omega \circ h(\underline{\mathbf{X}}_m)\|^2$$

Tensor Regression

We are trying to avoid matricization on every step of decoding target variables. So, tensor regression can be defined as:

$$\mathbf{y}_m = \langle \underline{\mathbf{X}}_m | \underline{\mathbf{W}} \rangle + \varepsilon$$

Where $\langle \underline{\mathbf{X}}_m | \underline{\mathbf{W}} \rangle$ denotes a tensor contraction along the first D modes:

$$\langle \underline{\mathbf{X}}_m | \underline{\mathbf{W}} \rangle_k = \sum_{i_1=1}^{n_1} \cdots \sum_{i_D=1}^{n_D} x_{i_1,\ldots,i_D} w_{i_1,\ldots,i_D,k}$$

In practice, for very large scale problems, tensors are expressed approximately in tensor network formats. For example, with the application of Tucker multilinear rank tensor representation:

$$\underline{\mathbf{W}} \approx \underline{\mathbf{G}} \times_1 \mathbf{U}^{(1)} \ldots \times_D \mathbf{U}^{(D)} \times_{D+1} \mathbf{U}^{(D+1)}$$

Skoltech

Algorithm

$$
\begin{tikzcd}
\underline{\mathbf{X}} \arrow[r, "\Phi"] \arrow[d, "\Psi_x"'] & \mathbf{Y} \arrow[d, "\Psi_y"] \\
\underline{\hat{\mathbf{X}}} \arrow[r, "\underline{\Phi}"] \arrow[d, "h_x"'] & \underline{\mathbf{Y}} \arrow[d, "h_y"] \\
\mathbf{T} \arrow[r, "g"'] & \mathbf{U}
\end{tikzcd}
$$

$$\Phi = \Psi_x \circ h_x \circ g \circ h_y^{-1} \circ \Psi_y^{-1}$$

Ψ_x, Ψ_y are tensorization methods:

- without tensorization
- hankelization along time
- hankelization along space
- hankelization along both dimensions

h_x, h_y are dimensionality reduction models

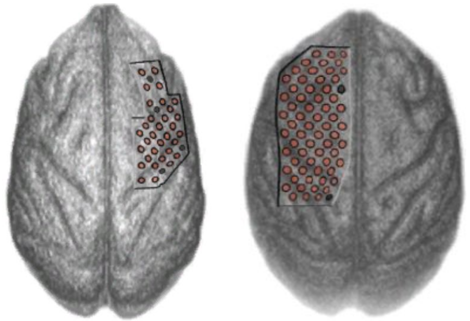
g is regression model in latent space

$\underline{\Phi}$ can be presented by PLS, HOPLS

$\mathbf{T}, \mathbf{U}$ are latent tensors

NeuroTycho food-tracking dataset

ECoG signals were obtained from 32 channels. Moreover, frequency-domain features were obtained with wavelet transform with 27 frequencies.

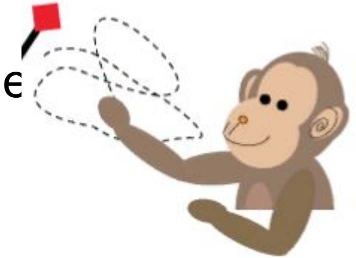


$$\underline{\mathbf{X}} \in \mathbb{R}^{T\times 32\times 27}, \mathbf{Y} \in \mathbb{R}^{T\times 3}$$

After hankelization along temporal and spatial modes:

$$\hat{\underline{\mathbf{X}}} \in \mathbb{R}^{\hat{T}\times 10\times 27\times 31\times 2}, \underline{\mathbf{Y}} \in \mathbb{R}^{\hat{T}\times 10\times 3}$$

Additionally, we considered the data from othe



with 64 channels ECoG.

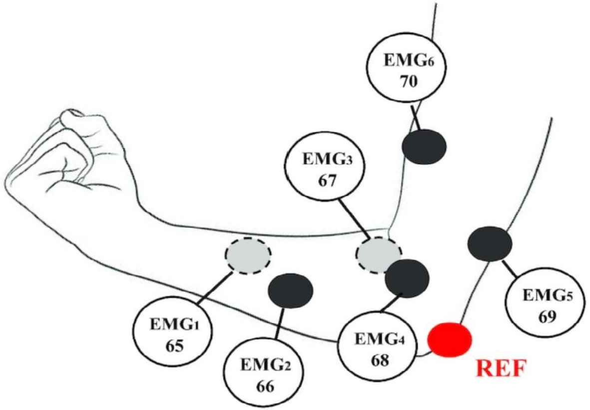
Skoltech

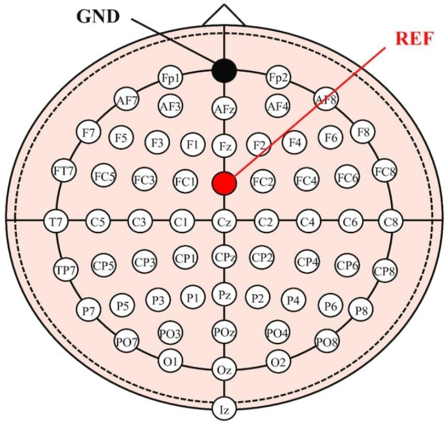
Multimodal signal human EEG dataset

EEG signals were obtained from 60 channels. Moreover, frequency-domain features were obtained with wavelet transform with 24 frequencies. EMG data consists of 6 signals.

$$\underline{\mathbf{X}} \in \mathbb{R}^{T \times 64 \times 24}, \quad \mathbf{Y} \in \mathbb{R}^{T \times 6}$$

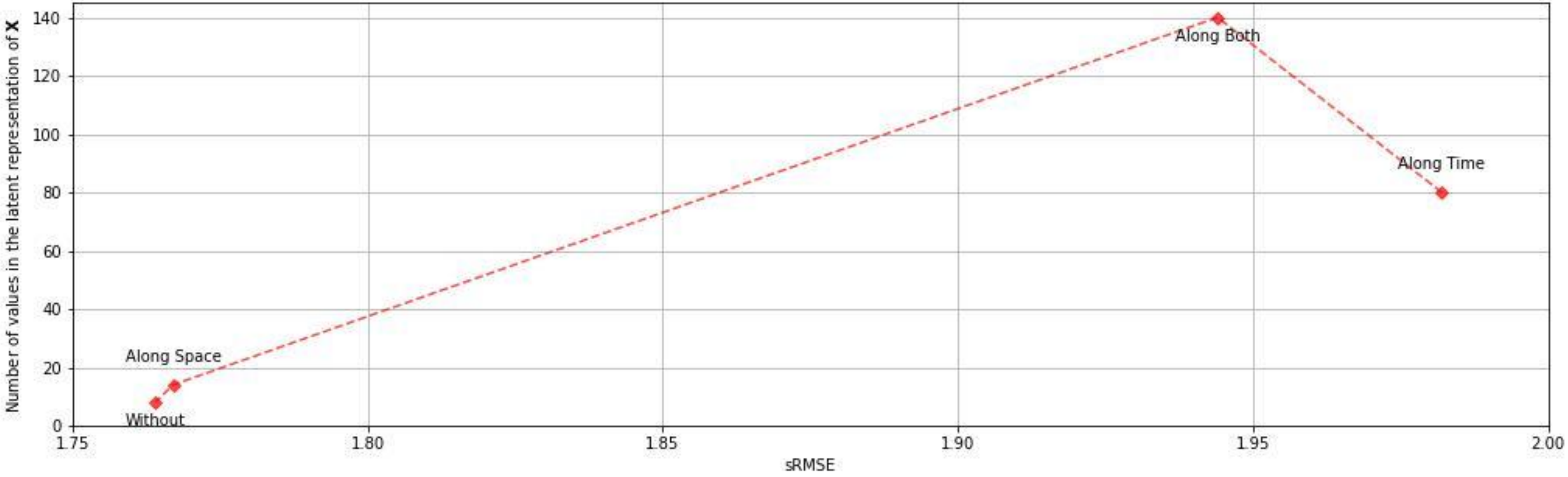
Hankelization was made the same way as for previous dataset.





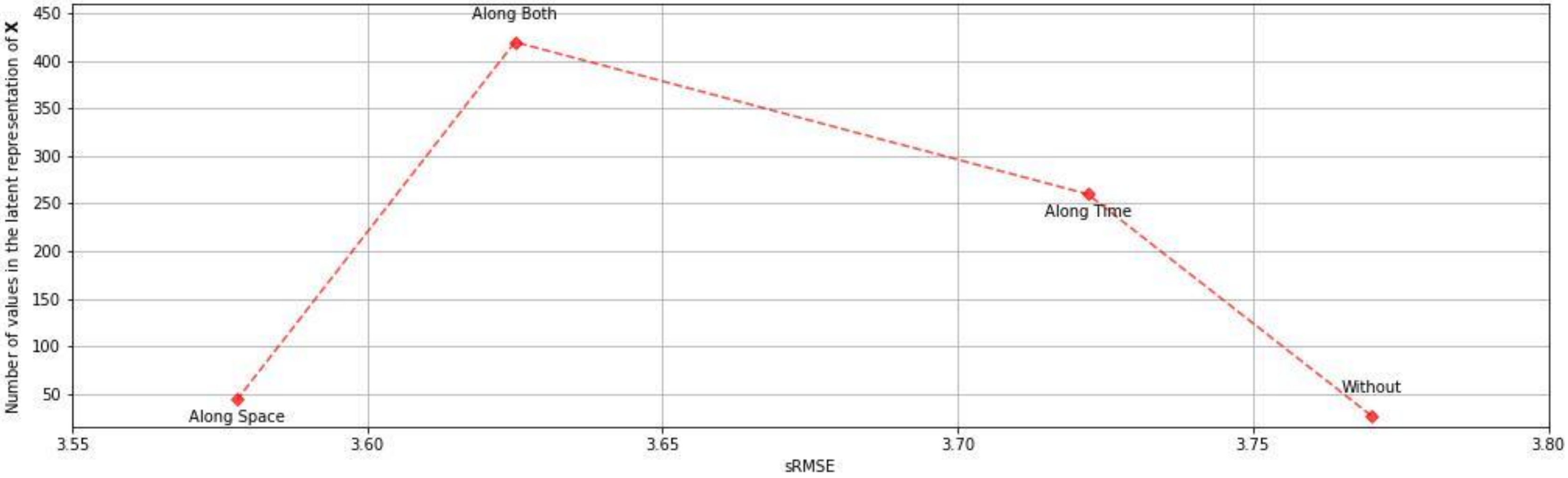
Skoltech

Results for NeuroTycho (ch=32) with MPCA



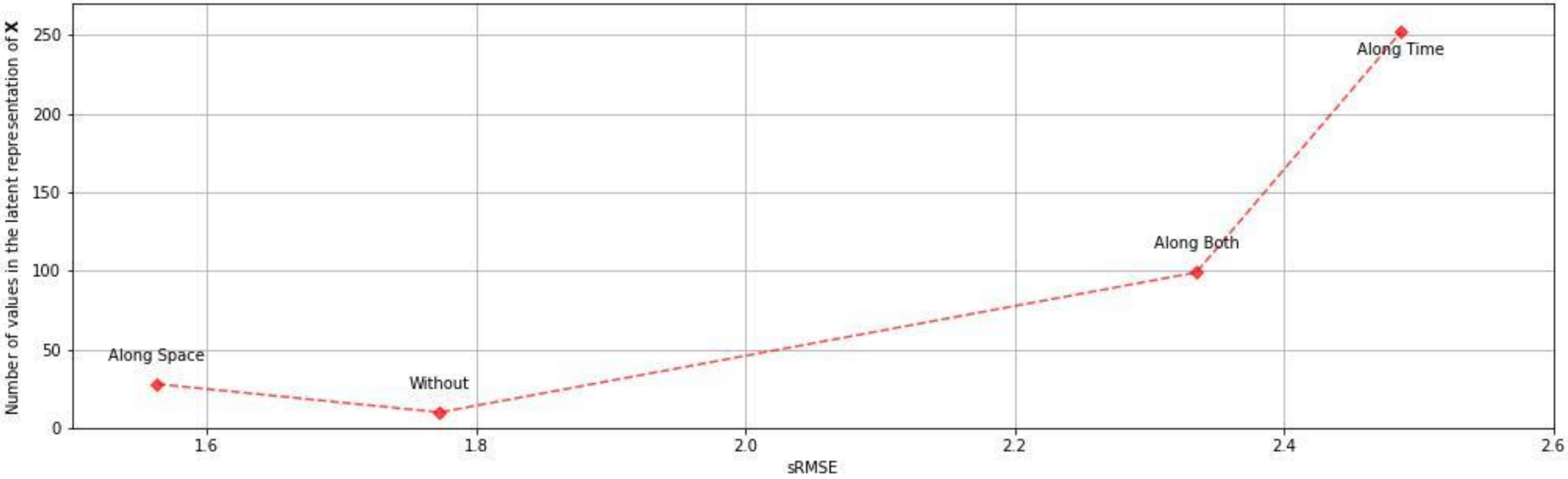
Skoltech

Results for NeuroTycho (ch=64) with MPCA



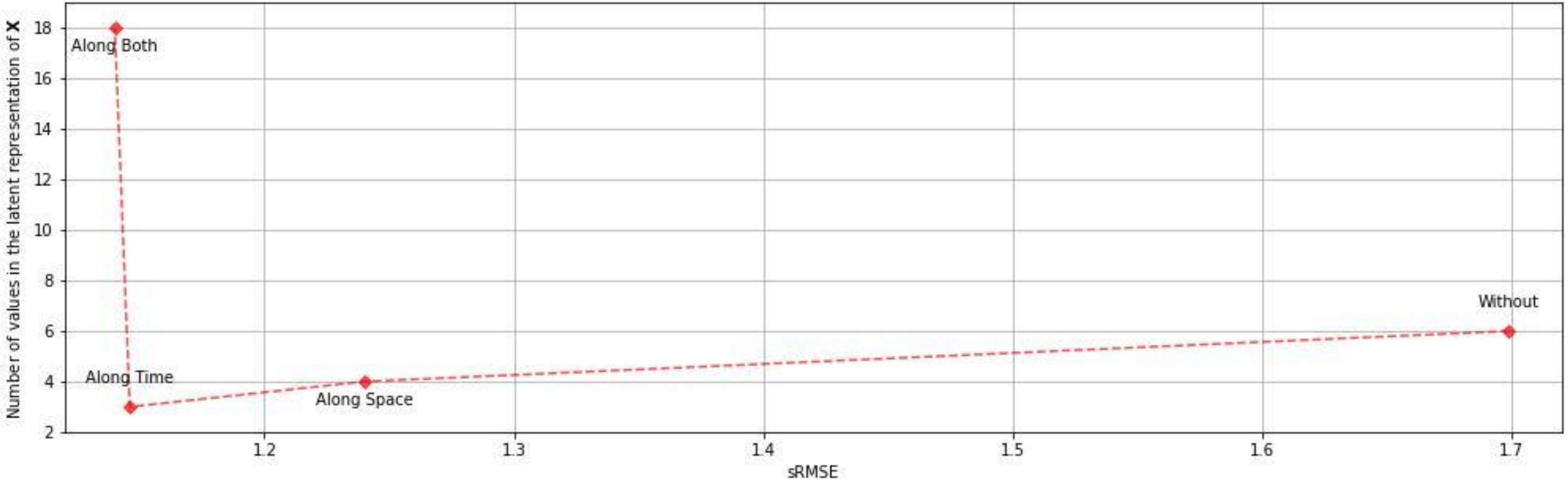
Skoltech

Results for EEG (ch=60) with MPCA



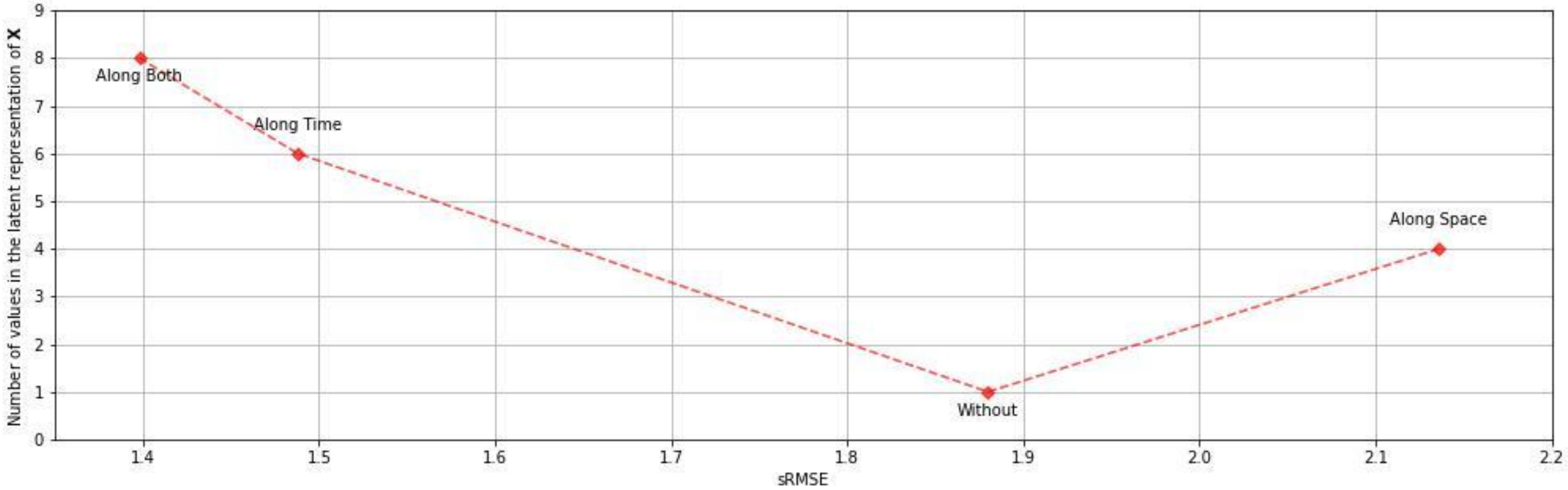
Skoltech

Results for NeuroTycho (ch=32) with HOPLS



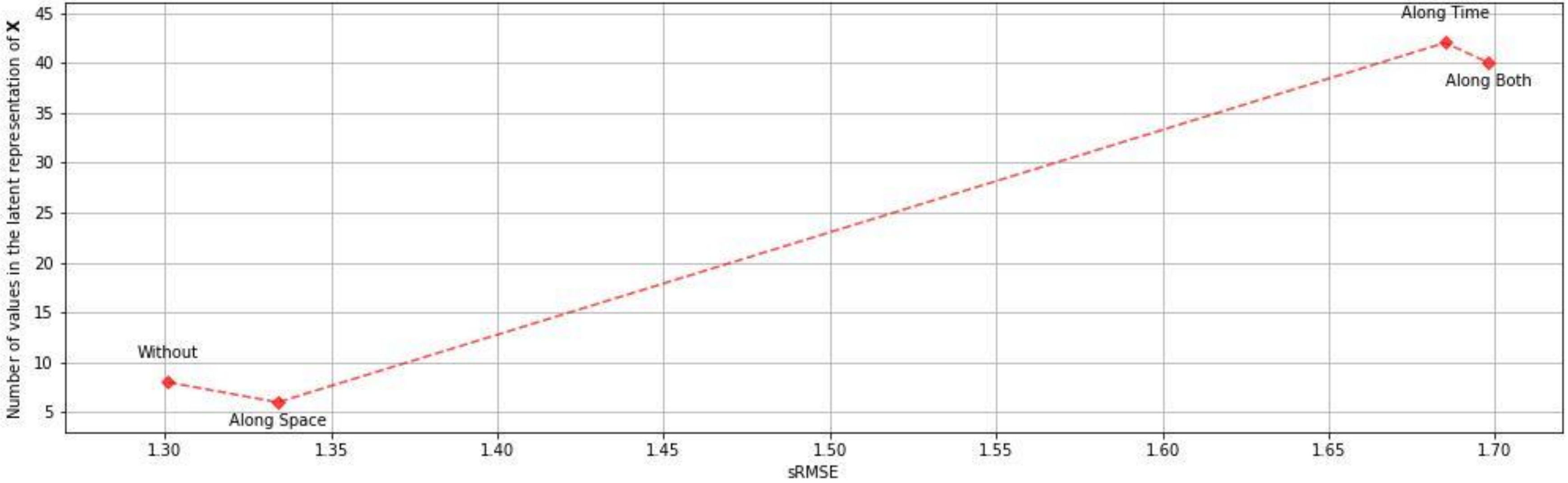
Skoltech

Results for NeuroTycho (ch=64) with HOPLS



Skoltech

Results for EEG (ch=60) with HOPLS



Skoltech

Summary of the results

MPCA:

- the optimal type of tensorization mostly is hankelization along space dimension;
- no tensorization gives fewer number of values in latent representation

HOPLS:

- the optimal type of tensorization is hankelization along both dimensions for two datasets;
- for the Human EEG dataset, the smallest metric was observed without hankelization, but the fewest number of the values of latent representations was obtained by hankelization along space dimension.

Skoltech

Discussion

- Previously it was shown that the forecasting of time series not for BCI is better with hankelization only along temporal dimension.
- For BCI:
  - hankelization along spatial dimension works the best way for MPCA in many cases.
  - hankelization along temporal and spatial dimensions works better than only along temporal dimension for HOPLS.
- It can be because of high correlation between data from different electrodes.

Skoltech

Plans

- Finish experiments for autoencoder;
- Finish writing text of the thesis.

Skoltech