Dimensionality reduction for signal forecasting

Roman Isachenko

Skoltech advisor: Maxim Fedorov

MIPT advisor: Vadim Strijov

April 18, 2018.

General Problem

Task

To analyze the input signal.

- multicollinearity in input and target spaces;
- high correlation between signals;
- the covariance matrices are essentially non-diagonal.

There are two models for input and target spaces.

Goal

To adjust the models for high multicollinear spaces.

Problem

To select structurally simple, stable, and exact model in the case of data redundancy and multiextremality of an error function.

Solution

Feature selection method that takes into account the input and target spaces structures and gives self-consistent model.

Literature

- Katrutsa A., Strijov V. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. // *Expert Systems with Applications* 76, 2017.

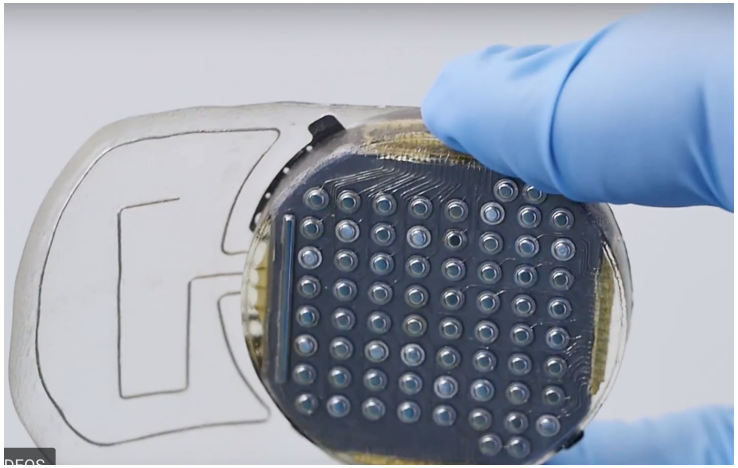
- Eliseyev A., Aksenova T. Stable and artifact-resistant decoding of 3D hand trajectories from ECoG signals using the generalized additive model //*Journal of neural engineering* 11(6), 2014.

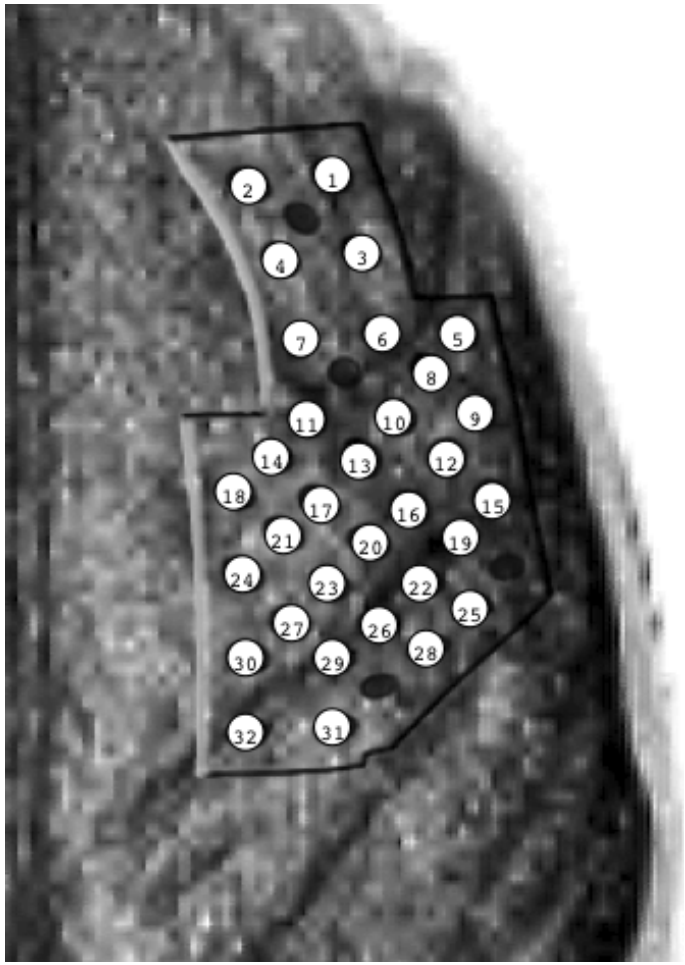
- Eliseyev A. et al. Iterative N-way partial least squares for a binary self-paced brain–computer interface in freely moving animals //*Journal of neural engineering* 4(8), 2011.

- Rodriguez-Lujan I. et al. Quadratic programming feature selection // *Journal of Machine Learning Research* 11(Apr), 2010.

- Motrenko A., Strijov V. Multi-way Feature Selection for ECoG-based Brain-Computer Interface // *Expert Systems with Applications* Submitted to the journal.

Application: Brain Computer Interface (BCI)





http://clinatec.fr; http://neurotycho.org

Problem Statement

Goal

Forecast a dependent variable $\mathbf{y} \in \mathbb{R}^r$ from an independent input object $\mathbf{x} \in \mathbb{R}^n$.

$$\mathbf{y} = \boldsymbol{\Theta}\mathbf{x} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\Theta} \in \mathbb{R}^{r \times n}$$

Given

Dataset $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} \in \mathbb{R}^{m \times n}$ is a design matrix, $\mathbf{Y} \in \mathbb{R}^{m \times r}$ is a target matrix

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^T = [\chi_1, \ldots, \chi_n]; \quad \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]^T = [\phi_1, \ldots, \phi_r].$$

Error function

$$S(\boldsymbol{\Theta}|\mathbf{X}, \mathbf{Y}) = \left\| \underset{m \times r}{\mathbf{Y}} - \underset{m \times n}{\mathbf{X}} \cdot \underset{r \times n}{\boldsymbol{\Theta}}^T \right\|_2^2 \to \min_{\boldsymbol{\Theta}}.$$

$$\boldsymbol{\Theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

The linear dependent columns of the matrix $\mathbf{X}$ leads to an instable solution. To avoid the strong linear dependence, feature selection and dimensionality reduction techniques are used.

Feature Selection

Goal

Find the index set $\mathcal{A} = \{1, \ldots, n\}$ of $\mathbf{X}$ columns.

Quality Criteria

To select the set $\mathcal{A}$ among all possible $2^n - 1$ subsets, introduce the feature selection quality criteria

$$\mathcal{A} = \underset{\mathcal{A}' \subseteq \{1,\ldots,n\}}{\arg\max} \; Q(\mathcal{A}' | \mathbf{X}, \mathbf{Y}).$$

Once the solution $\mathcal{A}$ is known:

$$S(\mathbf{\Theta}_{\mathcal{A}} | \mathbf{X}_{\mathcal{A}}, \mathbf{Y}) = \left\| \mathbf{Y} - \mathbf{X}_{\mathcal{A}} \mathbf{\Theta}_{\mathcal{A}}^T \right\|_2^2 \to \min_{\mathbf{\Theta}_{\mathcal{A}}},$$

where the subscript $\mathcal{A}$ indicates columns with indices from the set $\mathcal{A}$.

Quadratic Programming Feature Selection

$$\|\phi - \mathbf{X}\boldsymbol{\theta}\|_2^2 \to \min_{\boldsymbol{\theta}\in\mathbb{R}^n}.$$

Quadratic Programming Feature Selection

$$(1-\alpha)\cdot \underbrace{\mathbf{a}^T\mathbf{Q}\mathbf{a}}_{\text{Sim}(\mathbf{X})} - \alpha \cdot \underbrace{\mathbf{b}^T\mathbf{a}}_{\text{Rel}(\mathbf{X},\phi)} \to \min_{\substack{\mathbf{a}\in\mathbb{R}^n_+ \\ \|\mathbf{a}\|_1=1}}.$$

- $\mathbf{a} \in \mathbb{R}^n$ — feature importances;
- $\mathbf{Q} \in \mathbb{R}^{n\times n}$ - pairwise feature similarities;
- $\mathbf{b} \in \mathbb{R}^n$ - feature relevances to the target vector.

$$j \in \mathcal{A}^* \Leftrightarrow a_j > \tau$$

Quality Criteria

$$\mathcal{A} = \arg\max_{\mathcal{A}'\subseteq\{1,\ldots,n\}} Q(\mathcal{A}'|\mathbf{X},\phi) \Leftrightarrow \arg\min_{\mathbf{a}\in\mathbb{R}^n_+,\ \|\mathbf{a}\|_1=1} \left[\mathbf{a}^T\mathbf{Q}\mathbf{a} - \alpha\cdot\mathbf{b}^T\mathbf{a}\right].$$

Quadratic Programming Feature Selection

Quality Criteria

$$\mathcal{A} = \underset{\mathcal{A}' \subseteq \{1,\ldots,n\}}{\arg\max}\ Q(\mathcal{A}'|\mathbf{X}, \phi) \Leftrightarrow \underset{\mathbf{a} \in \mathbb{R}^n_+,\ \|\mathbf{a}\|_1 = 1}{\arg\min} \left[\mathbf{a}^T \mathbf{Q} \mathbf{a} - \alpha \cdot \mathbf{b}^T \mathbf{a}\right].$$

Similarity measure

- Correlation

$$|\mathrm{corr}(\mathbf{x}, \mathbf{y})| = \left| \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\mathrm{Var}(\mathbf{x})\mathrm{Var}(\mathbf{y})}} \right|$$

- Mutual information

$$I(\mathbf{x}, \mathbf{y}) = \int \int p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}) p(\mathbf{y})} d\mathbf{x} d\mathbf{y}.$$

$$\mathbf{Q} = \left\{ \left| \mathrm{corr}(\chi_i, \chi_j) \right| \right\}_{i,j=1}^{n}, \quad \mathbf{b} = \left\{ |\mathrm{corr}(\chi_i, \phi)| \right\}_{i=1}^{n}.$$

Statement

In the case of semidefinite matrix $\mathbf{Q}$ the QPFS problem is convex.

$$\mathbf{Q} \to \mathbf{Q} - \lambda_{\min} \mathbf{I}$$

Multivariate QPFS

Relevance aggregation

$$\mathbf{Q} = \left\{ \left| \mathrm{corr}(\chi_i, \chi_j) \right| \right\}_{i,j=1}^{n}, \quad \mathbf{b} = \left\{ \sum_{k=1}^{r} |\mathrm{corr}(\chi_i, \phi_k)| \right\}_{i=1}^{n}.$$

This approach does not use the dependencies in the columns of the matrix $\mathbf{Y}$.

Example:

$$\mathbf{X} = [\chi_1, \chi_2, \chi_3], \quad \mathbf{Y} = [\underbrace{\phi_1, \phi_1, \ldots, \phi_1}_{r-1}, \phi_2],$$

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \underbrace{0.4 & \ldots & 0.4}_{r-1} & 0 \\ 0.5 & \ldots & 0.5 & 0.8 \\ 0.8 & \ldots & 0.8 & 0.1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} (r-1) \cdot 0.4 + 0 \\ (r-1) \cdot 0.5 + 0.8 \\ (r-1) \cdot 0.8 + 0.1 \end{bmatrix}$$

Best subset: $[\chi_1, \chi_2]$.

QPFS ($r = 2$): $\mathbf{a} = [\mathbf{0.37}, \mathbf{0.61}, 0.02]$.
QPFS ($r = 5$): $\mathbf{a} = [\mathbf{0.40}, 0.17, \mathbf{0.43}]$.

Multivariate QPFS

Proposal

Penalize correlated targets by $\mathrm{Sim}(\mathbf{Y})$

$$\alpha_1 \cdot \underbrace{\mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x}_{\mathrm{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^T \mathbf{B} \mathbf{a}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} + \alpha_3 \cdot \underbrace{\mathbf{a}_y^T \mathbf{Q}_y \mathbf{a}_y}_{\mathrm{Sim}(\mathbf{Y})} \to \min_{\substack{\mathbf{a}_x \in \mathbb{R}_+^n \ \|\mathbf{a}_x\|_1 = 1 \\ \mathbf{a}_y \in \mathbb{R}_+^r \ \|\mathbf{a}_y\|_1 = 1}}.$$

$$\mathbf{Q}_x = \left\{ \left| \mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) \right| \right\}_{i,j=1}^{n}, \ \mathbf{Q}_y = \left\{ \left| \mathrm{corr}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) \right| \right\}_{i,j=1}^{r}, \ \mathbf{B} = \left\{ \left| \mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\phi}_j) \right| \right\}_{\substack{i=1,\ldots,n \\ j=1,\ldots,r}}.$$

$$\alpha_1 + \alpha_2 + \alpha_3 = 1 \quad \alpha_i \geq 0, \ i = 1, 2, 3.$$

Statement

For the case $r = 1$ the proposed functional coincides with the original QPFS algorithm.

Statement

Balance between the terms $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ for the problem is achieved by the following coefficients:

$$\alpha_1 = \frac{(1 - \alpha_3)\overline{\mathbf{B}}}{\overline{\mathbf{Q}}_x + \overline{\mathbf{B}}}; \quad \alpha_2 = \frac{(1 - \alpha_3)\overline{\mathbf{Q}}_x}{\overline{\mathbf{Q}}_x + \overline{\mathbf{B}}}; \quad \alpha_3 \in [0, 1],$$

where $\overline{\mathbf{Q}}_x$, $\overline{\mathbf{B}}$ are the mean values of $\mathbf{Q}_x$ and $\mathbf{B}$ respectively.

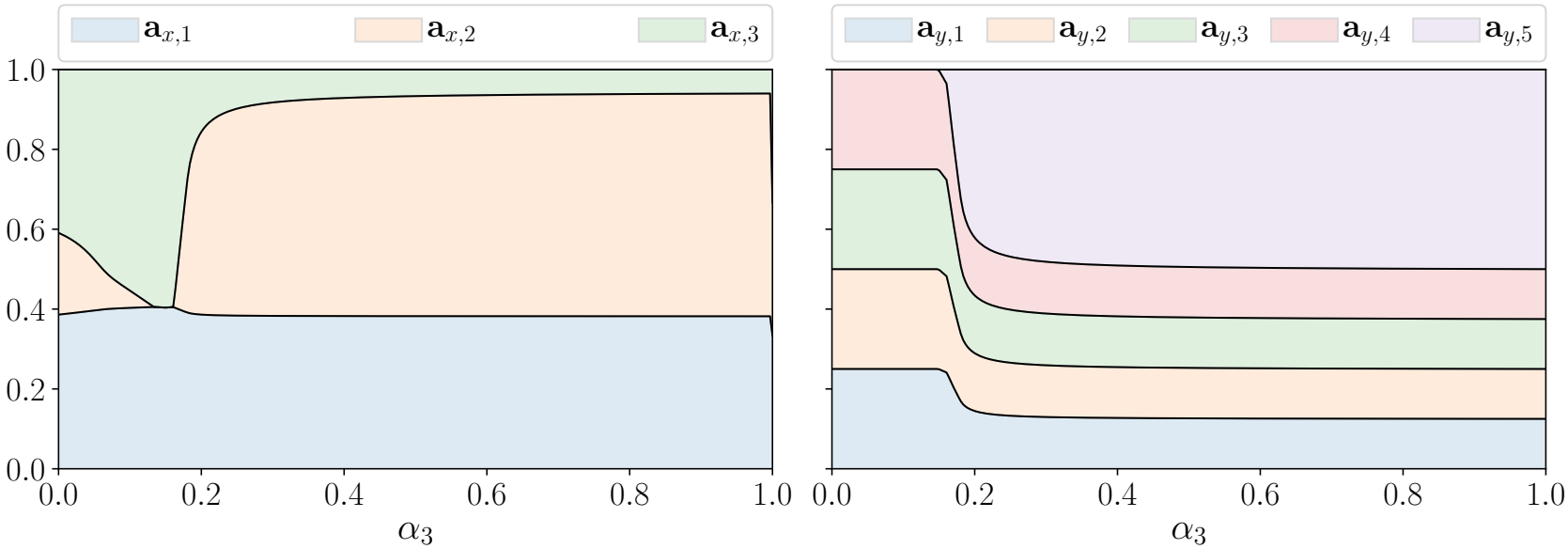
Multivariate QPFS

Example:

$$\mathbf{X} = [\chi_1, \chi_2, \chi_3], \quad \mathbf{Y} = [\underbrace{\phi_1, \phi_1, \ldots, \phi_1}_{r-1}, \phi_2],$$

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \underbrace{0.4 & \ldots & 0.4}_{r-1} & 0 \\ 0.5 & \ldots & 0.5 & 0.8 \\ 0.8 & \ldots & 0.8 & 0.1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} (r-1) \cdot 0.4 + 0 \\ (r-1) \cdot 0.5 + 0.8 \\ (r-1) \cdot 0.8 + 0.1 \end{bmatrix}$$

$\mathbf{Q}_x = \mathbf{Q}$; $\mathbf{Q}_y$: $\text{corr}(\phi_1, \phi_2) = 0.2$ all others entries $= 1$.



Feature categorization

1. non-relevant features
$$\left\{j : \text{corr}(\boldsymbol{\chi}_j, \boldsymbol{\phi}_k) = 0,\ \forall k \in \{1, \ldots, r\}\right\};$$
2. non-**X**-correlated features, which are relevant to non-**Y**-correlated targets
$$\left\{j : \left(\text{VIF}(\boldsymbol{\chi}_j) < 10\right) \text{ and } \left(\text{VIF}(\boldsymbol{\phi}_k) < 10,\ \forall k \in \{1, \ldots, r\} : \text{corr}(\boldsymbol{\chi}_j, \boldsymbol{\phi}_k) \neq 0\right)\right\};$$
3. non-**X**-correlated features, which are relevant to **Y**-correlated targets
$$\left\{j : \left(\text{VIF}(\boldsymbol{\chi}_j) < 10\right) \text{ and } \left(\exists k \in \{1, \ldots, r\} : \text{VIF}(\boldsymbol{\phi}_k) > 10 \ \&\ \text{corr}(\boldsymbol{\chi}_j, \boldsymbol{\phi}_k) \neq 0\right)\right\};$$
4. **X**-correlated features, which are relevant to non-**Y**-correlated targets
$$\left\{j : \left(\text{VIF}(\boldsymbol{\chi}_j) > 10\right) \text{ and } \left(\text{VIF}(\boldsymbol{\phi}_k) < 10,\ \forall k \in \{1, \ldots, r\} : \text{corr}(\boldsymbol{\chi}_j, \boldsymbol{\phi}_k) \neq 0\right)\right\};$$
5. **X**-correlated features, which are relevant to **Y**-correlated targets
$$\left\{j : \left(\text{VIF}(\boldsymbol{\chi}_j) > 10\right) \text{ and } \left(\exists k \in \{1, \ldots, r\} : \text{VIF}(\boldsymbol{\phi}_k) > 10 \ \&\ \text{corr}(\boldsymbol{\chi}_j, \boldsymbol{\phi}_k) \neq 0\right)\right\}.$$

$$r = \text{corr}(\boldsymbol{\chi}, \boldsymbol{\phi}), \quad t = \frac{r\sqrt{m-2}}{1-r^2} \sim \text{St}(m-2).$$

$$\text{VIF}(\boldsymbol{\chi}_j) = \frac{1}{1-R_j^2}, \quad \text{VIF}(\boldsymbol{\phi}_k) = \frac{1}{1-R_k^2},$$

where $R_j^2 (R_k^2)$ are coefficients of determination for the regression of $\boldsymbol{\chi}_j (\boldsymbol{\phi}_k)$ on the other features(targets).

Computational experiment

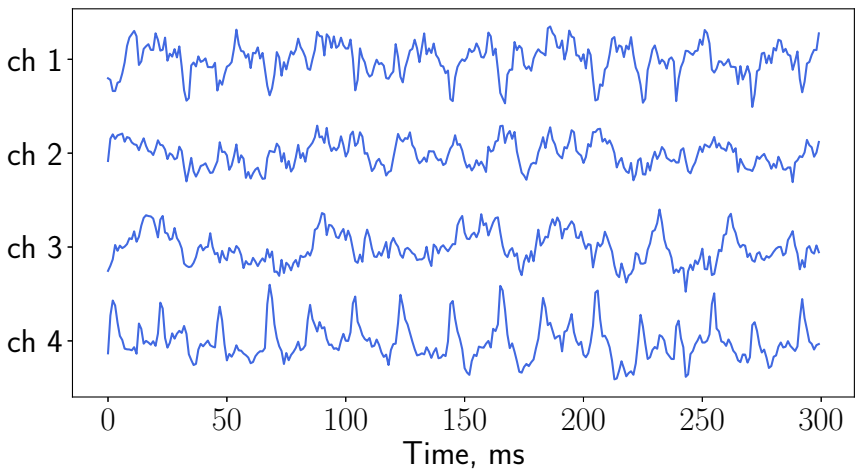
Datasets

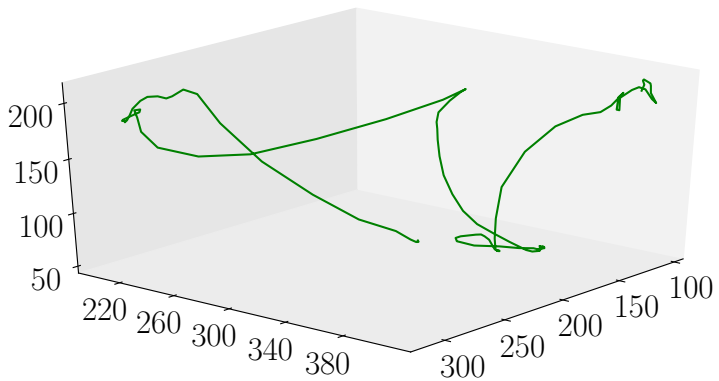
- energy consumption
- electrocorticogram signals (ECoG)

Autoregressive approach

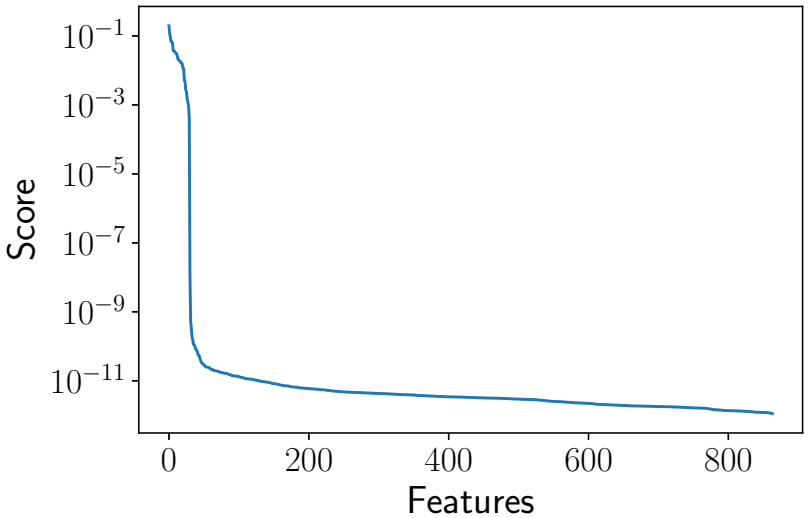
$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ x_2 & x_3 & \dots & x_{n+1} \\ \dots & \dots & \dots & \dots \\ x_{T-n+1} & x_{T-n+2} & \dots & x_T \end{pmatrix}$$

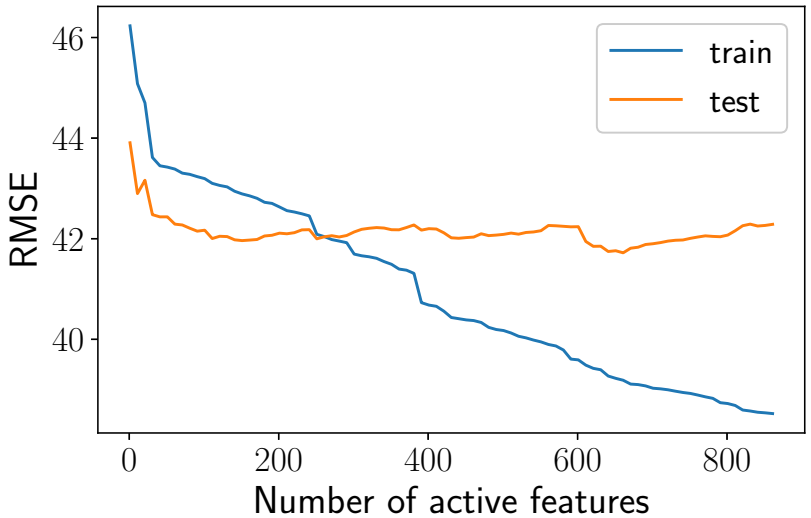
ECoG data

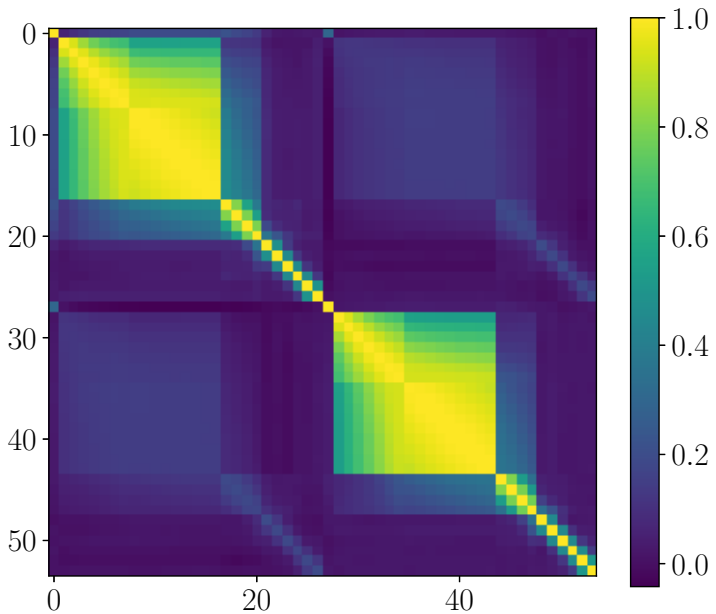


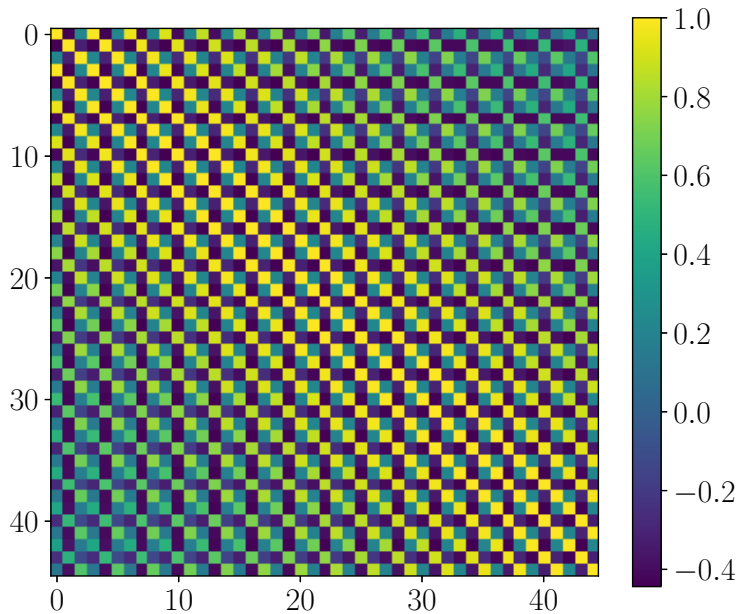


Experiment

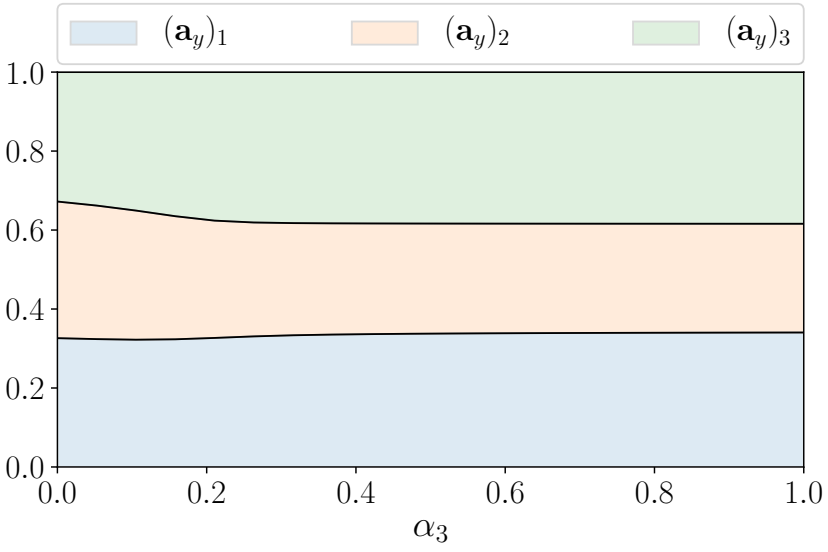

Score
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
0
200
400
600
800
Features

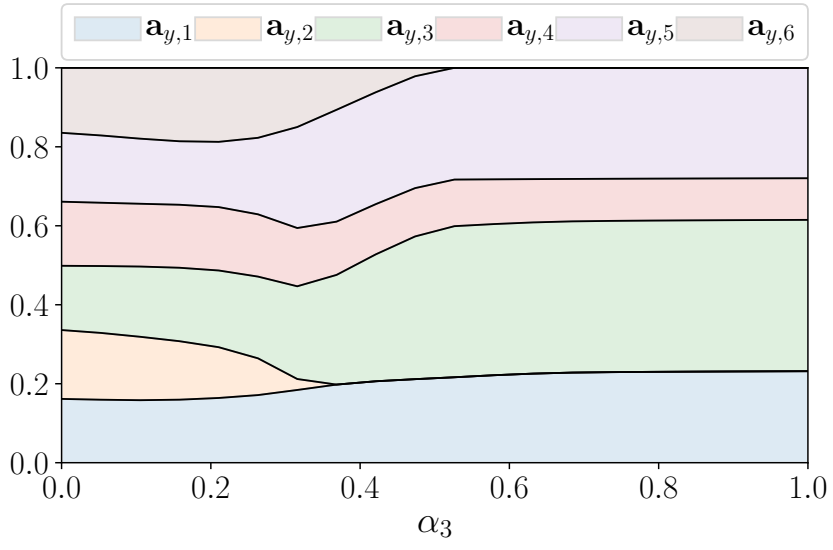

RMSE
46
44
42
40
train
test
0
200
400
600
800
Number of active features

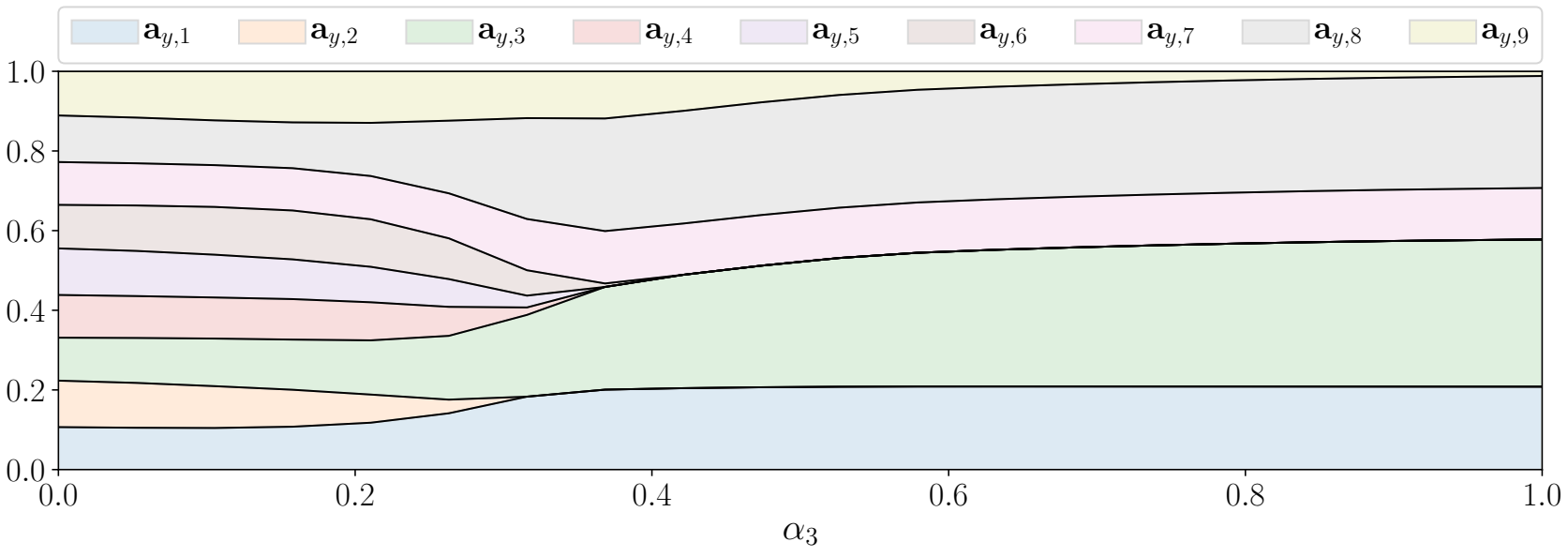

0
10
20
30
40
50
0
20
40
1.0
0.8
0.6
0.4
0.2
0.0


0
10
20
30
40
0
10
20
30
40
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

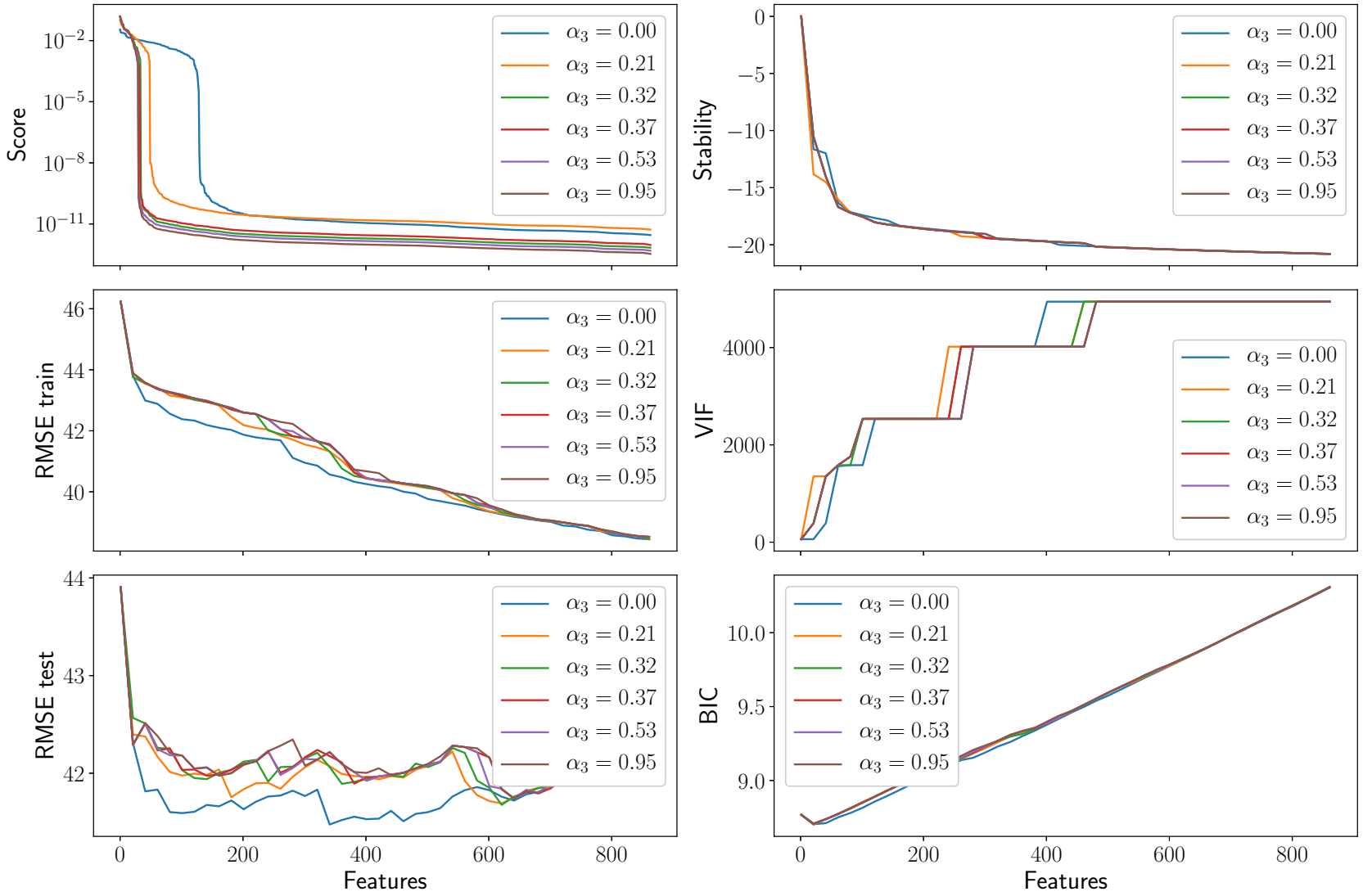
Experiment


$(\mathbf{a}_y)_1$
$(\mathbf{a}_y)_2$
$(\mathbf{a}_y)_3$
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
α_3

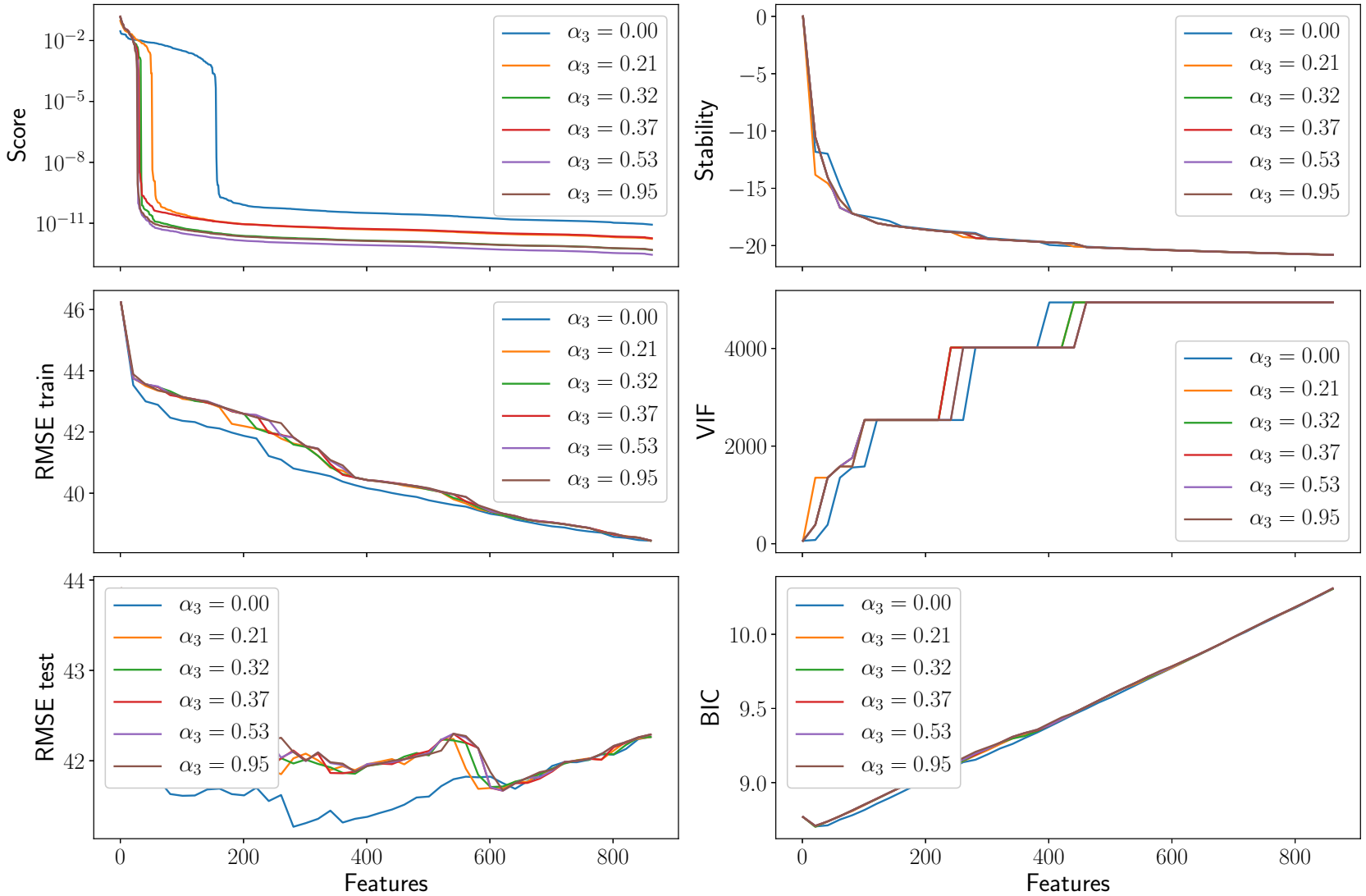

$\mathbf{a}_{y,1}$
$\mathbf{a}_{y,2}$
$\mathbf{a}_{y,3}$
$\mathbf{a}_{y,4}$
$\mathbf{a}_{y,5}$
$\mathbf{a}_{y,6}$
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
α_3


$\mathbf{a}_{y,1}$
$\mathbf{a}_{y,2}$
$\mathbf{a}_{y,3}$
$\mathbf{a}_{y,4}$
$\mathbf{a}_{y,5}$
$\mathbf{a}_{y,6}$
$\mathbf{a}_{y,7}$
$\mathbf{a}_{y,8}$
$\mathbf{a}_{y,9}$
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
α_3

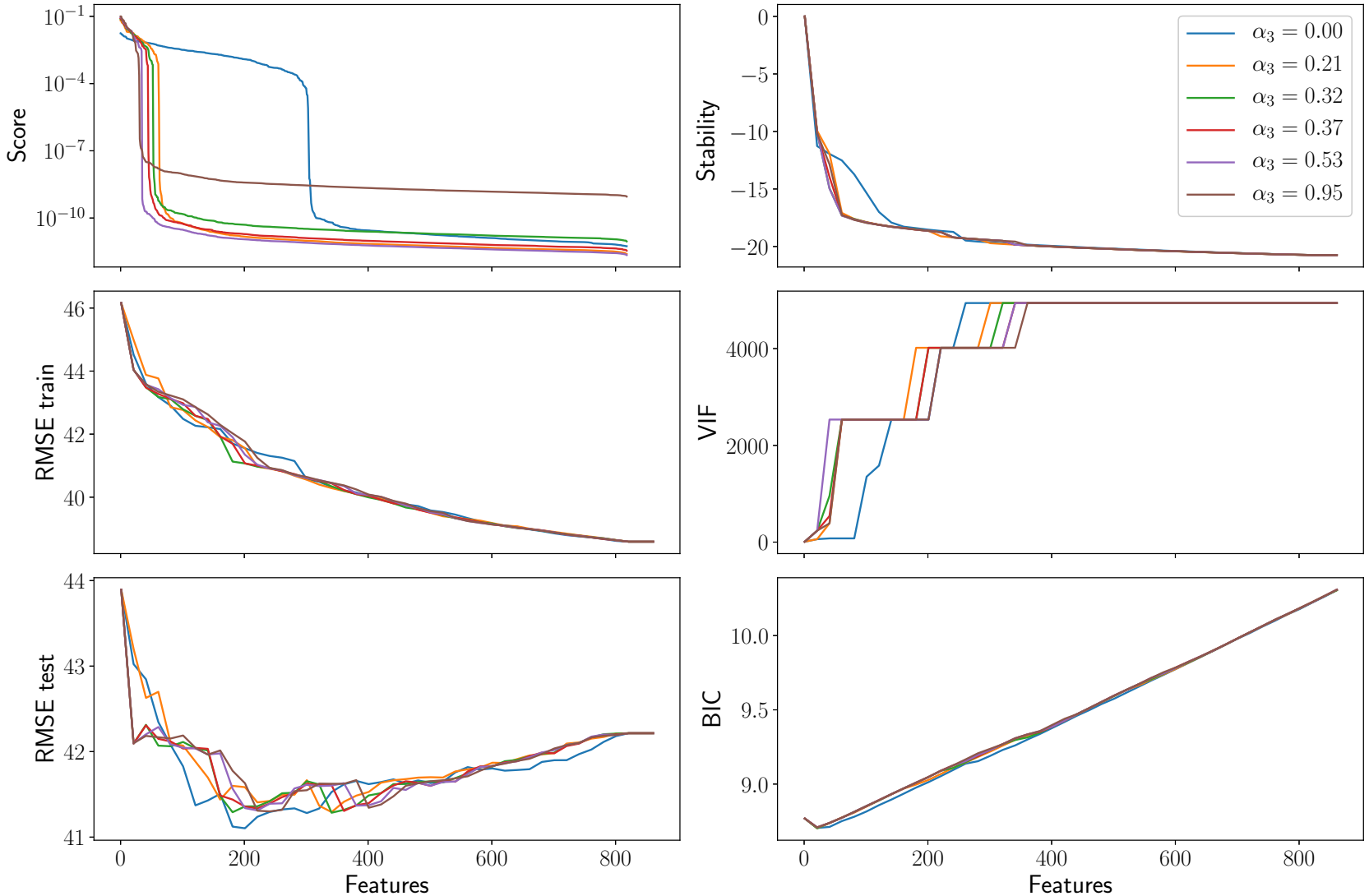
Autoregression step = 1



Autoregression step = 3



Autoregression step = 30



Experiment

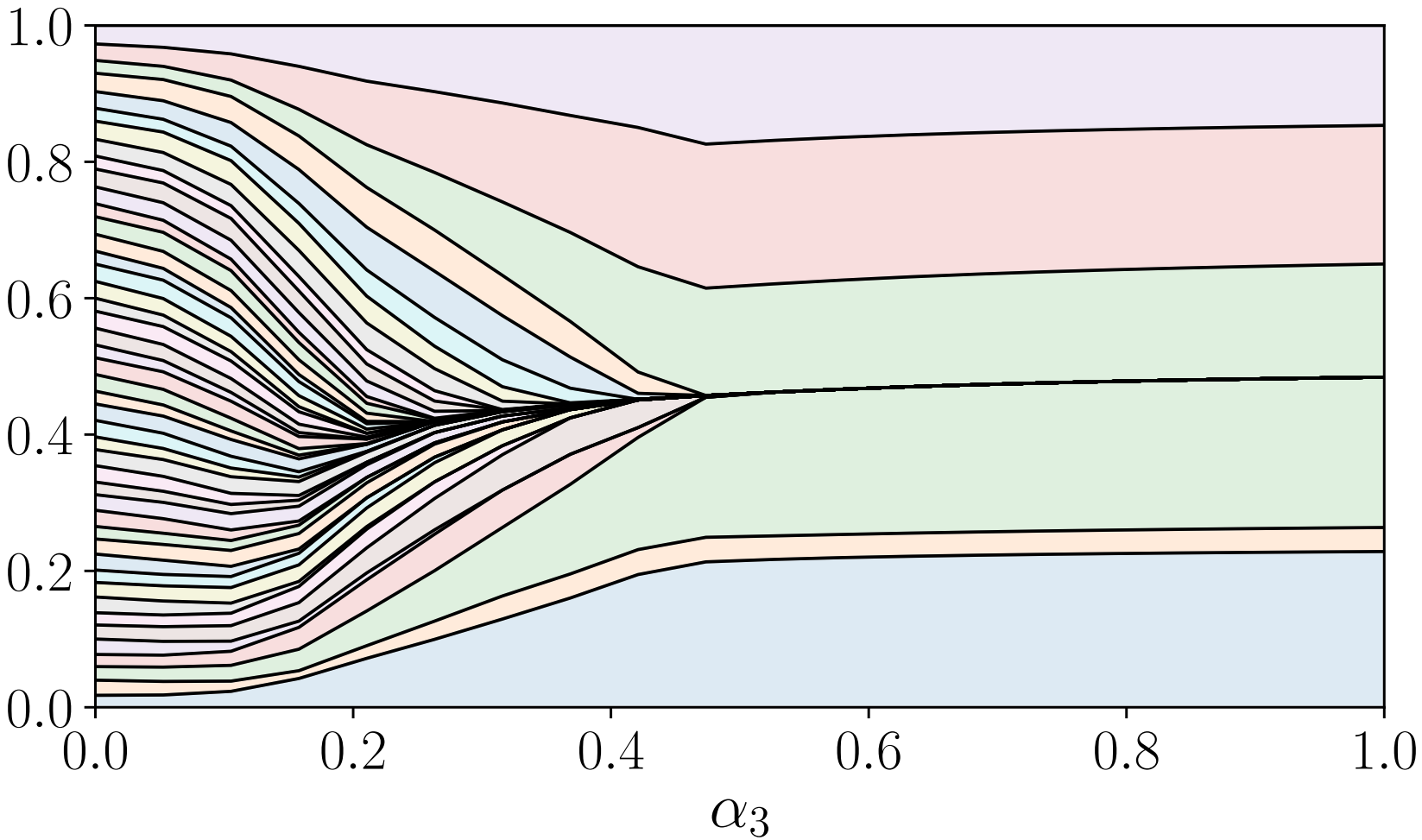
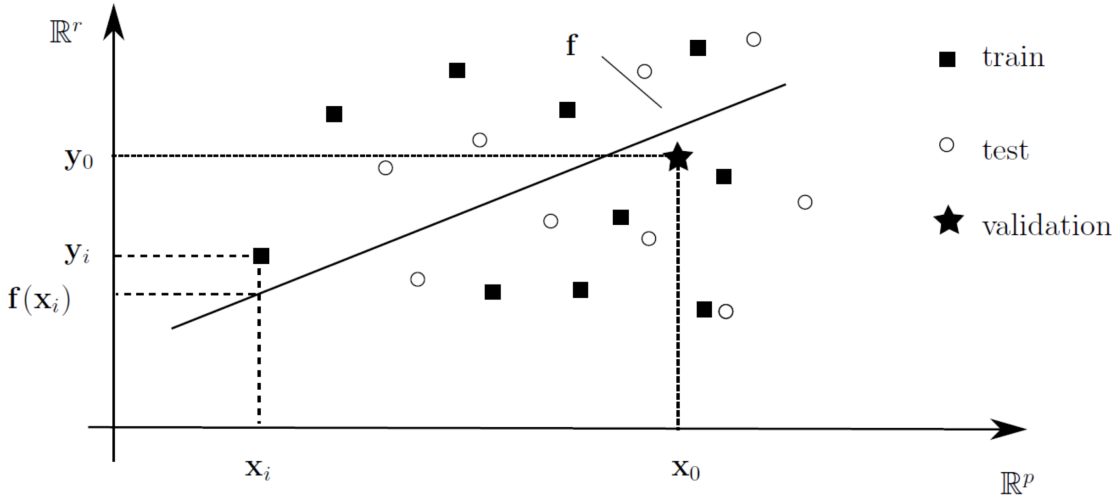


Figure: autoregression step=45

Problem Statement



$$\begin{array}{ccc} \mathbf{X} \in \mathbb{X} & \longrightarrow & \mathbf{Y} \in \mathbb{Y} \\ \updownarrow & & \updownarrow \\ \mathbf{T} \in \mathbb{Z} & \longleftrightarrow & \mathbf{U} \in \mathbb{Z} \end{array}$$

Partial Least Squares (PLS)

$$\underset{m\times n}{\mathbf{X}} = \underset{m\times l}{\mathbf{T}} \cdot \underset{l\times n}{\mathbf{P}^T} + \underset{m\times n}{\mathbf{F}} = \sum_{k=1}^{l} \underset{m\times 1}{\mathbf{t}_k} \cdot \underset{1\times n}{\mathbf{p}_k^T} + \underset{m\times n}{\mathbf{F}}$$

$$\underset{m\times r}{\mathbf{Y}} = \underset{m\times l}{\mathbf{U}} \cdot \underset{l\times r}{\mathbf{Q}^T} + \underset{m\times r}{\mathbf{E}} = \sum_{k=1}^{l} \underset{m\times 1}{\mathbf{t}_k} \cdot \underset{1\times r}{\mathbf{q}_k^T} + \underset{m\times r}{\mathbf{E}}$$

$$\hat{\mathbf{Y}} = \mathbf{T}\mathrm{diag}(\beta)\mathbf{Q}^T = \mathbf{X}\Theta.$$

- map $\mathbf{X}$ into low-dimensional $\mathbf{T}$;
- map $\mathbf{Y}$ into low-dimensional $\mathbf{U}$;
- maximize correlation between $\mathbf{t}_k$ and $\mathbf{u}_k$.

PLS Example

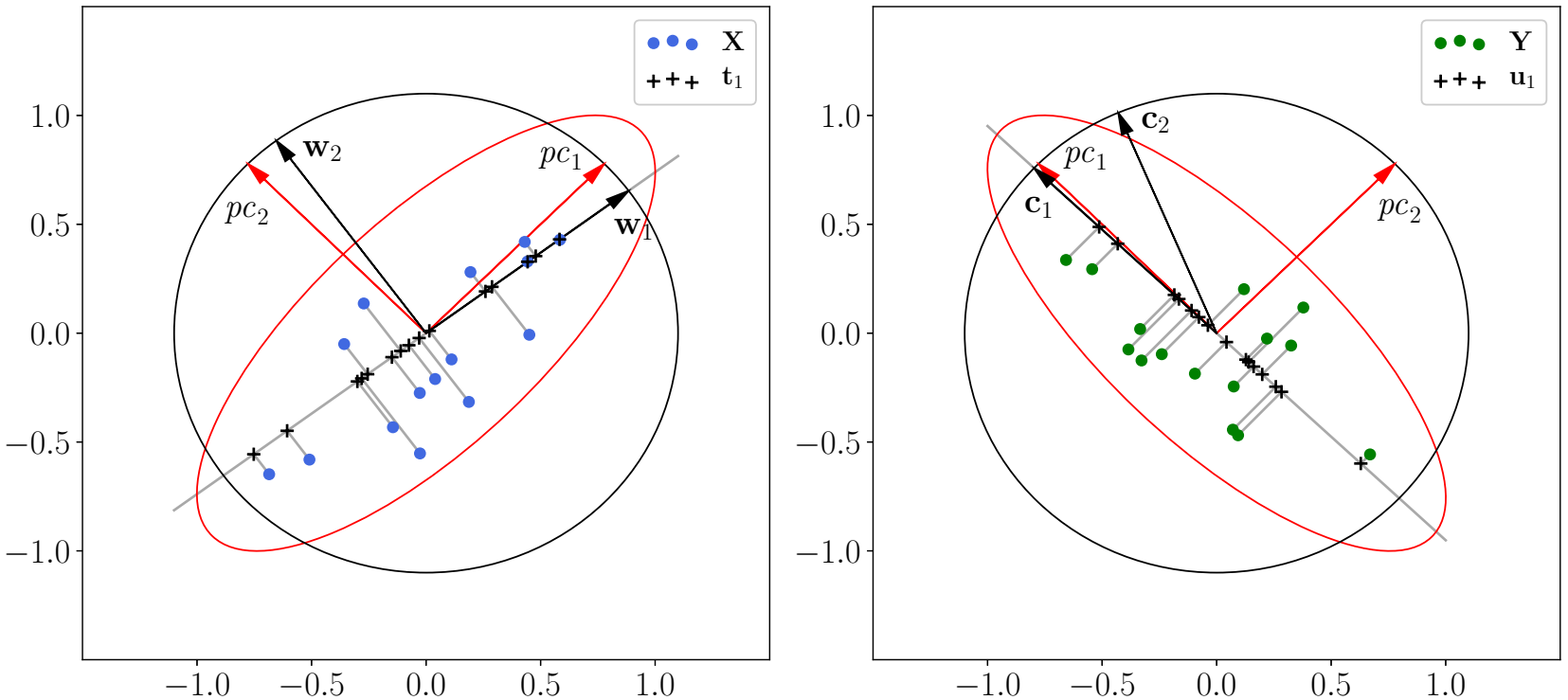


Figure: The result of the PLS algorithm for the case $n = r = l = 2$.

Computational experiment

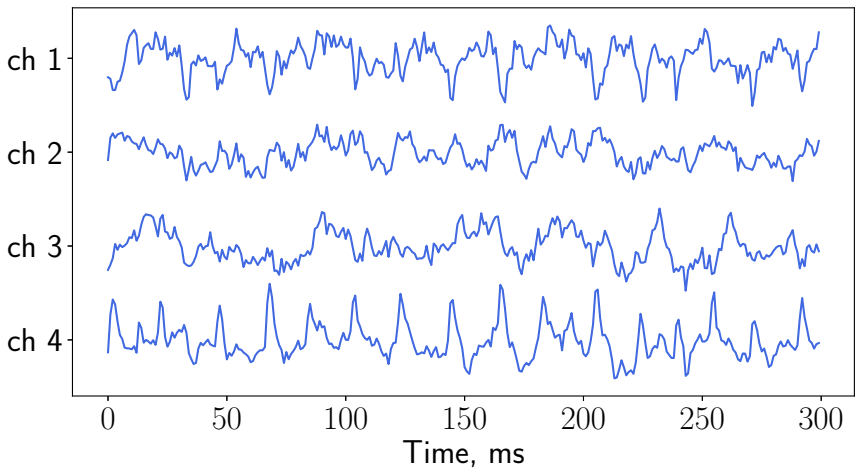
Datasets

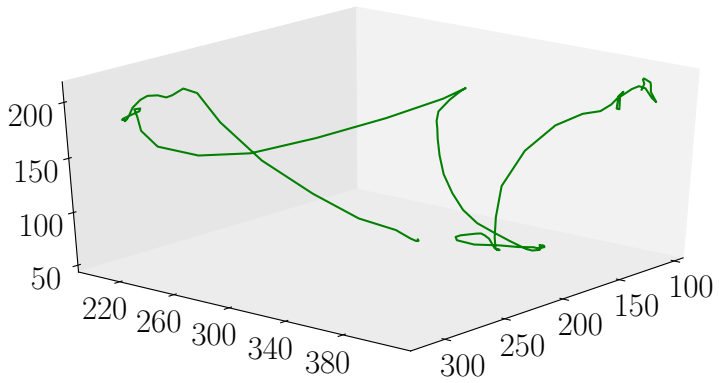
- energy consumption
- electrocorticogram signals (ECoG)

Autoregressive approach

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ x_2 & x_3 & \dots & x_{n+1} \\ \dots & \dots & \dots & \dots \\ x_{T-n+1} & x_{T-n+2} & \dots & x_T \end{pmatrix}$$

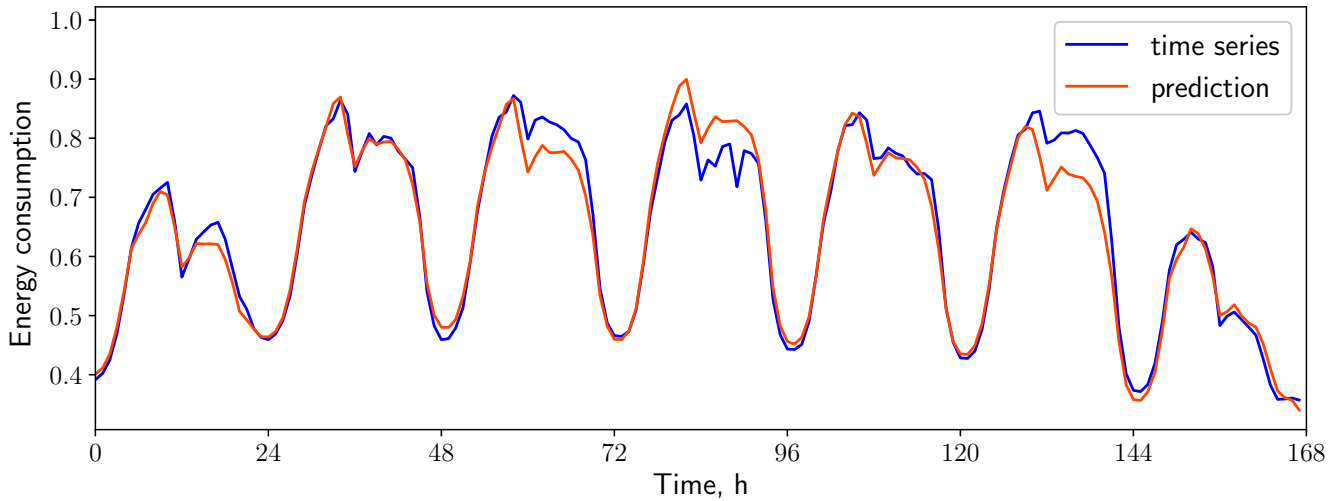
ECoG data





Computational experiment

Energy consumption

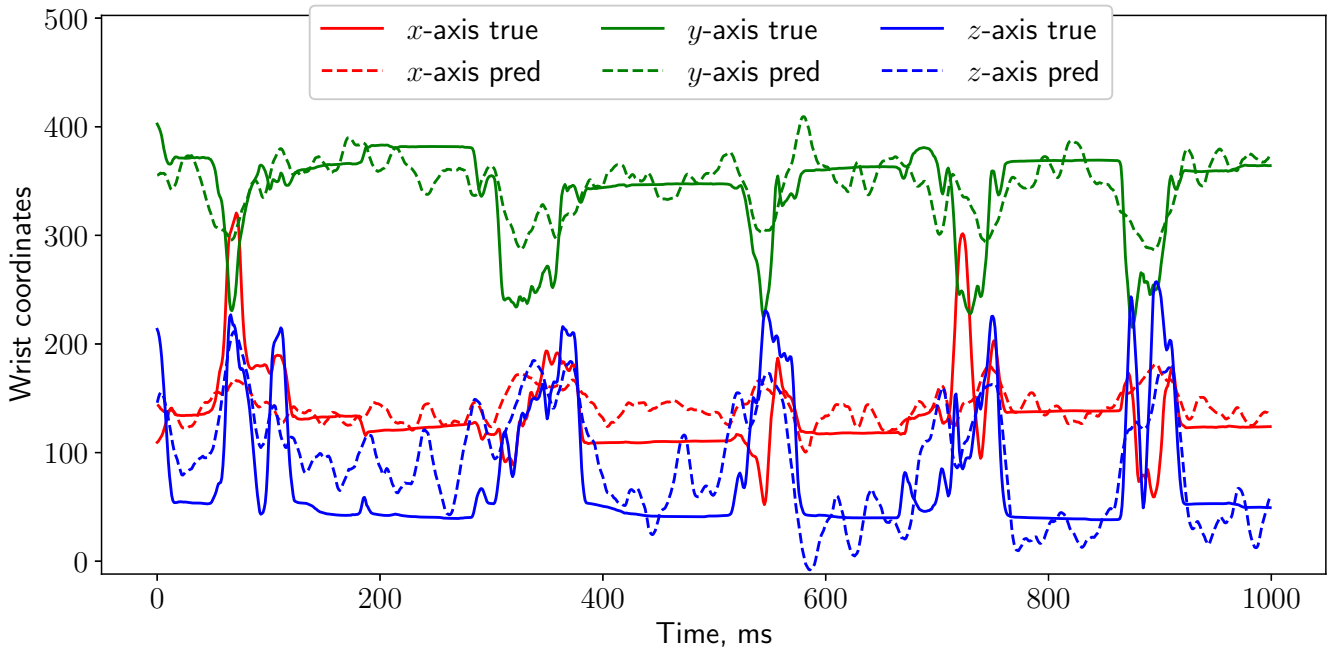


Results

- Space dimensionalities: $\mathbf{X} = 700 \times (24 \cdot 7)$, $\mathbf{Y} = 700 \times 24$.
- Dimensionality of latent space: 14
- NMSE: 0.047

Computational experiment

ECoG



Results

- Space dimensionalities: $\mathbf{X} = 13000 \times (864 \cdot 18)$, $\mathbf{Y} = 13000 \times 3$.
- Dimensionality of latent space: 16
- NMSE: 0.731

Quadratic Programming Model Selection

QPFS

- works for linear problems;
- does not take into account the model;
- ignores the structure of the target space.

Problem

$$\underbrace{(1-\alpha)\mathbf{z}^T\mathbf{Q}\mathbf{z}}_{\text{Sim}} - \underbrace{\alpha\mathbf{b}^T\mathbf{z}}_{\text{Rel}} \to \min_{\substack{\mathbf{z}\in\mathbb{R}^p_+ \\ \|\mathbf{z}\|_1=1}} .$$

- $\mathbf{z} \in \mathbb{R}^p$ — weight importances;
- $\mathbf{Q} \in \mathbb{R}^{p\times p}$ - pairwise weights interactions;
- $\mathbf{b} \in \mathbb{R}^p$ - weight relevances to the target vector.

$$w_j = 0 \Leftrightarrow z_j < \tau.$$

Model: feed-forward neural network

$$f(\mathbf{x}|\mathbf{w}) = \sigma_2\left(\mathbf{W}_2\sigma_1(\mathbf{W}_1\mathbf{x})\right).$$

- weights from different layers do not interact;
- weight interactions $\rightarrow$ similarity between neurons;
- weight relevances $\rightarrow$ model linearization

$$\mathbf{f}(\mathbf{X}|\mathbf{w}+\Delta\mathbf{w}) \approx \mathbf{f}(\mathbf{X}|\mathbf{w}) + \mathbf{J}\cdot\Delta\mathbf{w},$$

where $\mathbf{J} \in \mathbb{R}^{m\times p}$ is a Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f(\mathbf{x}_1|\mathbf{w})}{\partial w_1} & \cdots & \frac{\partial f(\mathbf{x}_1|\mathbf{w})}{\partial w_j} & \cdots & \frac{\partial f(\mathbf{x}_1|\mathbf{w})}{\partial w_p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(\mathbf{x}_m|\mathbf{w})}{\partial w_1} & \cdots & \frac{\partial f(\mathbf{x}_m|\mathbf{w})}{\partial w_j} & \cdots & \frac{\partial f(\mathbf{x}_m|\mathbf{w})}{\partial w_p} \end{pmatrix}$$

Algorithm

1. Find the optimal model parameters:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^p} S(\mathbf{w}|\mathbf{X}, \mathbf{y}, f),$$

   where $S(\mathbf{w}|\mathbf{X}, \mathbf{y}, f)$ — squared error function for regression, cross-entropy for classification.

2. Find active weights subset $\mathcal{A}$

$$\mathbf{z}^* = \arg\min_{\mathbf{z} \in \mathbb{R}^p_+ \|\mathbf{z}\|_1 = 1} Q(\mathbf{z}|\mathbf{X}, \mathbf{y}, f).$$

3. Find the parameters of the reduced model:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^p} S(\mathbf{w}|\mathbf{X}, \mathbf{y}, f), \quad \text{subject to } w_j = 0 \text{ for } j \notin \mathcal{A}.$$

4. (optional) Repeat steps (2) and (3).

Results

Synthetic data

- # features: 100
- # uncorrelated features: 5
- # feature combinations: 95

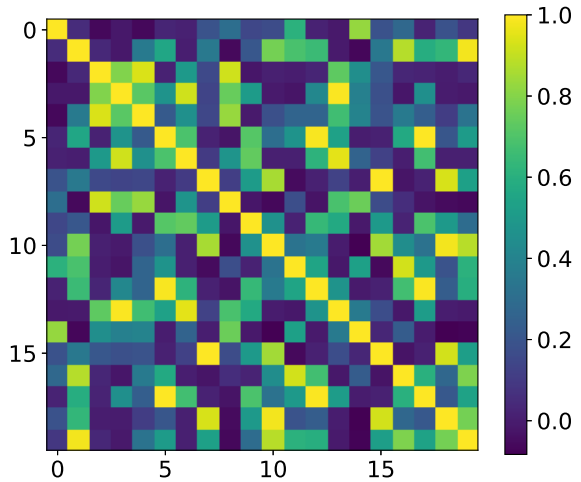


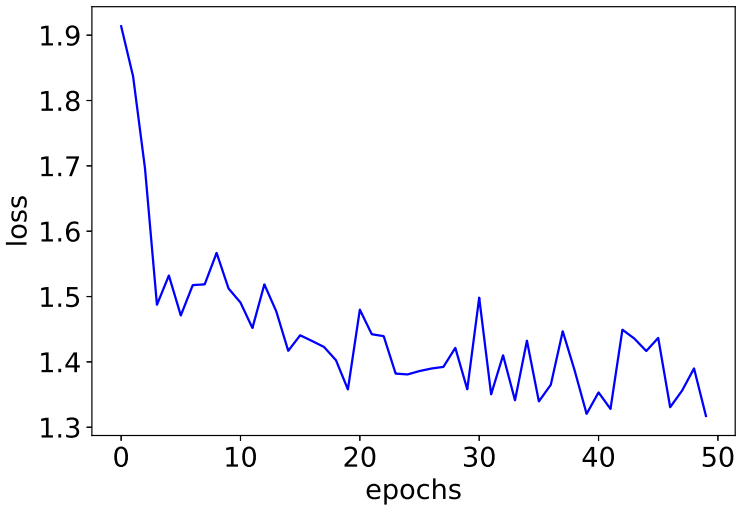
Figure: Correlation matrix

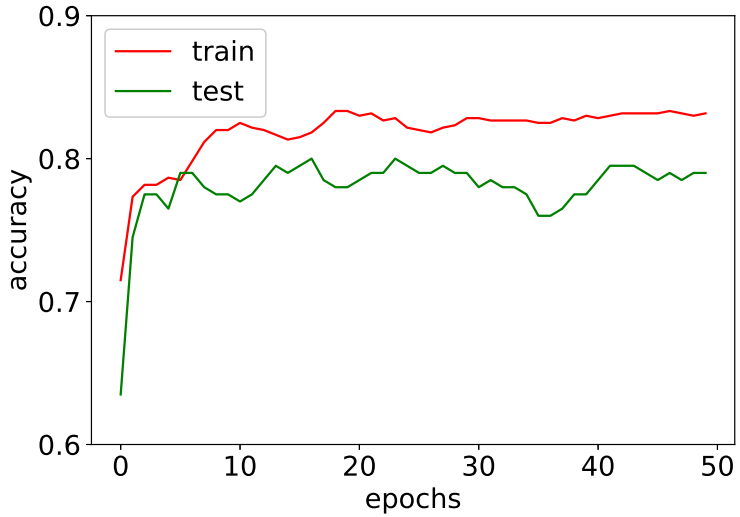
Results

- # parameters in the initial network: $100 \times 40 + 40 \times 2 = 4080$ (100%);
- accuracy of the initial network: train/test=83.7/78.5;
- # parameters in the reduced network: $92 + 47 = 139$ (4%);
- accuracy of the reduced network: train/test=81.6/78.4

Results

Initial network





Reduced network